

APPENDICES

A PROOF OF LEMMA 4.1

Lemma A.1 For any interval k , and $t \in [(k-1)\tau, k\tau)$

$$\|\mathbf{w}_i(t) - \mathbf{v}_k(t)\| \leq g_i(t - (k-1)\tau) \quad (25)$$

$$g_i(x) = \frac{\delta_i}{\beta + \mu_{i,k}} ((\eta(\beta + \mu_{i,k}) + 1)^x - 1) \quad (26)$$

Following the proof method from the reference (Wang et al., 2019) which is conducted using mathematical induction, we obtain Lemma A.1, where an additional term $\mu_{i,k}$ is introduced due to the proximal term added in the local training objective function.

Proof. To prove Lemma 4.1, we first establish the upper bound of the difference between the local training parameter \mathbf{w}_i in federated learning and the parameter \mathbf{v}_k in centralized learning, i.e., Lemma A.1. Next, we substitute the definitions into the equations and simplify to demonstrate the upper bound of the difference between the global aggregation parameter \mathbf{w} in federated learning and the parameter \mathbf{v}_k in centralized learning. Finally, we apply the method of cancellation through addition and subtraction to obtain the upper bound of the upper bound of the difference between the optimal parameter \mathbf{w}^* in federated learning and the parameter \mathbf{v}_k in centralized learning.

However, since the $\mu_{i,k}$ values differ across learners, it is necessary to account for the separate calculations for each learner when performing the geometric series summation. We first derive the upper bound of the difference between two consecutive iterations based on the definition formulas of \mathbf{w} and \mathbf{v}_k . Then, by summing and modifying the form of the inequality on the left-hand side, we substitute the upper bound of the consecutive differences to derive the form of the function $h(x)$.

To simplify the notation in the proof, we define $\nabla F^p(\mathbf{w}(t-1), \bar{\mu}_k) = \nabla F(\mathbf{w}(t-1)) + \bar{\mu}_k(\mathbf{w}(t-1) - \mathbf{w}_{k,global})$.

$$\begin{aligned} & \|\mathbf{w}(t) - \mathbf{v}_k(t)\| \\ \stackrel{(a)}{=} & \left\| \mathbf{w}(t-1) - \frac{\eta}{n} \sum_i \nabla F^p_i(\mathbf{w}_i(t-1), \mu_{i,k}) - \mathbf{v}_k(t-1) + \eta \nabla F^p(\mathbf{v}_k(t-1), \bar{\mu}_k) \right\| \\ \stackrel{(b)}{\leq} & \|\mathbf{w}(t-1) - \mathbf{v}_k(t-1)\| + \frac{\eta}{n} \sum_i \|\nabla F^p_i(\mathbf{w}_i(t-1), \mu_{i,k}) - \nabla F^p(\mathbf{v}_k(t-1), \bar{\mu}_k)\| \\ \stackrel{(c)}{\leq} & \|\mathbf{w}(t-1) - \mathbf{v}_k(t-1)\| + \eta(\beta + \bar{\mu}_k) \frac{\sum_i \|\mathbf{w}_i(t-1) - \mathbf{v}_k(t-1)\|}{n} \\ \stackrel{(d)}{\leq} & \|\mathbf{w}(t-1) - \mathbf{v}_k(t-1)\| + \eta(\beta + \bar{\mu}_k) \frac{\sum_i g(t-1 - (k-1)\tau)}{n} \\ \stackrel{(e)}{=} & \|\mathbf{w}(t-1) - \mathbf{v}_k(t-1)\| + \eta\delta \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}} \left((\eta(\beta + \mu_{i,k}) + 1)^{t-1-(k-1)\tau} - 1 \right)}{n}. \end{aligned} \quad (27)$$

Equation (a) is expanded based on the defined update formula, inequality (b) is simplified using the triangle inequality, inequality (c) follows from Assumption 1, inequality (d) applies the conclusion of Lemma A.1, and equation (e) is expanded accordingly.

By applying addition and subtraction cancellation and substituting the terms, we obtain the following proof formula.

$$\begin{aligned}
& \|\mathbf{w}^*(t) - \mathbf{v}_k(t)\| \\
& \stackrel{(a)}{=} \sum_{x=(k-1)\tau+1}^t [\|\mathbf{w}^*(x) - \mathbf{v}_k(x)\| - \|\mathbf{w}^*(x-1) - \mathbf{v}_k(x-1)\|] \\
& \stackrel{(b)}{\leq} \eta \delta \sum_{x=(k-1)\tau+1}^t \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}} \left((\eta(\beta + \mu_{i,k}) + 1)^{t-1-(k-1)\tau} - 1 \right)}{n} \\
& \stackrel{(c)}{=} \eta \delta \sum_{y=1}^{t-(k-1)\tau} \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}} \left((\eta(\beta + \mu_{i,k}) + 1)^{y-1} - 1 \right)}{n} \\
& \stackrel{(d)}{=} \eta \delta \sum_{y=1}^{t-(k-1)\tau} \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}} (\eta(\beta + \mu_{i,k}) + 1)^{y-1}}{n} - \eta \delta (t - (k-1)\tau) \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}}}{n} \\
& \stackrel{(e)}{=} \eta \delta \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}} \frac{1 - (\eta(\beta + \mu_{i,k}) + 1)^{t-(k-1)\tau}}{-\eta(\beta + \mu_{i,k})}}{n} - \eta \delta (t - (k-1)\tau) \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}}}{n} \\
& \stackrel{(f)}{=} \delta \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}} \frac{(\eta(\beta + \mu_{i,k}) + 1)^{t-(k-1)\tau} - 1}{\beta + \mu_{i,k}}}{n} - \eta \delta (t - (k-1)\tau) \frac{\sum_i \frac{\beta + \bar{\mu}_k}{\beta + \mu_{i,k}}}{n} \\
& = h(t - (k-1)\tau). \tag{28}
\end{aligned}$$

Equation (a) is expanded using the method of addition and subtraction cancellation, inequality (b) substitutes the above formula, and equations (c)–(f) are further simplified and summed to obtain the corresponding form of the function $h(x)$.

Based on Lemma 4.1, we can conclude within a given interval, as the number of training iterations increases and $\mu_{i,k}$ becomes larger, the value of the function $h(x)$ also increases, indicating a larger upper bound on the difference between the local training parameters \mathbf{w}_i and the parameters \mathbf{v}_k obtained using a centralized learning approach. However, when the number of training iterations is small, an appropriately chosen $\mu_{i,k}$ value can keep $h(x)$ from becoming too large. This indirectly demonstrates that the adaptive parameter $\mu_{i,k}$ can facilitate better convergence.

B PROOF OF LEMMA 4.2

Lemma B.1 When $\eta \leq \frac{1}{\beta + \max \mu_{i,k}}$ for any k and $t \in [(k-1)\tau, k\tau)$, $\|\mathbf{v}_k(t) - \mathbf{w}^*\|$ will not increase.

Proof.

$$\begin{aligned}
& \|\mathbf{v}_k(t+1) - \mathbf{w}^*\|^2 \\
& \stackrel{(a)}{=} \|\mathbf{v}_k(t) - \eta \nabla F^p(\mathbf{v}_k(t), \bar{\mu}_k) - \mathbf{w}^*\|^2 \\
& \stackrel{(b)}{=} \|\mathbf{v}_k(t) - \mathbf{w}^*\|^2 - 2\eta [\nabla F(\mathbf{v}_k(t), \bar{\mu}_k)]^T (\mathbf{v}_k(t) - \mathbf{w}^*) + \eta^2 \|\nabla F(\mathbf{v}_k(t), \bar{\mu}_k)\|^2. \tag{29}
\end{aligned}$$

Equation (a) substitutes the defined update formula, and equation (b) expands the square.

According to Reference (Wang et al., 2019), we obtain the following inequality.

$$\begin{aligned}
F^p(\mathbf{v}_k(t), \bar{\mu}_k) - F^p(\mathbf{w}^*, \bar{\mu}_k) & \leq [\nabla F^p(\mathbf{v}_k(t), \bar{\mu}_k)]^T (\mathbf{v}_k(t) - \mathbf{w}^*) \\
& \quad - \frac{1}{2(\beta + \bar{\mu}_k)} \|\nabla F^p(\mathbf{v}_k(t), \bar{\mu}_k) - \nabla F^p(\mathbf{w}^*, \bar{\mu}_k)\|^2. \tag{30}
\end{aligned}$$

Substituting the definition of $\theta_k(t)$, we obtain the following inequality.

$$0 \leq \theta_k(t) \leq [\nabla F^p(\mathbf{v}_k(t), \bar{\mu}_k)]^T (\mathbf{v}_k(t) - \mathbf{w}^*) - \frac{1}{2(\beta + \bar{\mu}_k)} \|\nabla F^p(\mathbf{v}_k(t), \bar{\mu}_k) - \nabla F^p(\mathbf{w}^*, \bar{\mu}_k)\|^2. \tag{31}$$

$$\begin{aligned}
& \|\mathbf{v}_k(t+1) - \mathbf{w}^*\| \\
& \stackrel{(a)}{\leq} \|\mathbf{v}_k(t) - \mathbf{w}^*\| - \frac{\eta}{\beta + \bar{\mu}_k} \|\nabla F^p(\mathbf{v}_k(t), \bar{\mu}_k) - \bar{\mu}_k(\mathbf{w}^* - \mathbf{w}_{k,global})\|^2 + \eta^2 \|\nabla F^p(\mathbf{v}_k(t), \bar{\mu}_k)\|^2 \\
& \stackrel{(b)}{\leq} \|\mathbf{v}_k(t) - \mathbf{w}^*\| - \frac{\eta}{\beta + \bar{\mu}_k} \|\nabla F(\mathbf{v}_k(t)) - 2\bar{\mu}_k\xi\|^2 + \eta^2 \|\nabla F(\mathbf{v}_k(t)) + \bar{\mu}_k\xi\|^2 \\
& \stackrel{(c)}{\leq} \|\mathbf{v}_k(t) - \mathbf{w}^*\| - \left(\frac{\eta}{\beta + \bar{\mu}_k} - \eta^2 \right) \|\nabla F(\mathbf{v}_k(t))\|^2 + \left(\frac{4\bar{\mu}_k\xi\eta}{\beta + \bar{\mu}_k} + 2\bar{\mu}_k\xi\eta^2 \right) \|\nabla F(\mathbf{v}_k(t))\| \\
& \quad + \left(-\frac{4\bar{\mu}_k^2\xi^2\eta}{\beta + \bar{\mu}_k} + \bar{\mu}_k^2\xi^2\eta^2 \right). \tag{32}
\end{aligned}$$

Inequality (a) is obtained by substituting the above formula, inequality (b) is derived by substituting the upper bound of the proximal term, and further expansion and simplification yield inequality (c).

From the conclusion of B.1, it can be observed that when the condition $\eta \leq \frac{1}{\beta + \max \mu_{i,k}}$ is satisfied, the second term on the right-hand side of the inequality (c) is negative. If the absolute value of the second term is sufficiently large, it is possible to ensure that $\|\mathbf{v}_k(t) - \mathbf{w}^*\|$ is non-increasing. Additionally, the value of $\mu_{i,k}$ influences the setting of η . Under this constraint, the larger the value of $\mu_{i,k}$, the smaller η should be set.

Lemma B.2 For any k , $\eta \leq \frac{1}{\beta + \max \mu_{i,k}}$, $t \in [(k-1)\tau, k\tau)$,

$$F^p(\mathbf{v}_k(t+1)) - F^p(\mathbf{v}_k(t)) \leq -\eta \left(1 - \frac{\eta(\beta + \bar{\mu}_k)}{2} \right). \tag{33}$$

Proof. Based on the study from reference (Wang et al., 2019), we derive Lemma B.2, which derives the upper bound of the difference in objective function values between two consecutive training iterations in a centralized learning method with a proximal term added to the objective function. By examining the right-hand side of the inequality, it is evident that the inclusion of the proximal term in the objective function makes $\bar{\mu}_k$ affect the value of this upper bound. As $\bar{\mu}_k$ increases, the upper bound decreases. A larger control parameter $\bar{\mu}_k$ leads to a greater impact on the proximal term's variation during training. A smaller difference in the objective function values between consecutive iterations indicates Smoother training. While this smoothness helps reduce fluctuations during training, it may also slow down the training process, underscoring the importance of properly designing the control parameter μ .

Lemma B.3 For any k , $\eta \leq \frac{1}{\beta + \max \mu_{i,k}}$, $t \in [(k-1)\tau, k\tau)$

$$\frac{1}{\theta_k(t+1)} - \frac{1}{\theta_k(t)} \geq \omega\eta \left(1 - \frac{(\beta + \bar{\mu}_k)\eta}{2} \right), \tag{34}$$

where $\omega = \min_k \frac{1}{\|\mathbf{v}_k((k-1)\tau) - \mathbf{w}^*\|^2}$.

We define $\theta_k(t) = F^p(\mathbf{v}_k(t), \bar{\mu}_k) - F^p(\mathbf{w}^*, \bar{\mu}_k)$ as the difference between the parameter obtained using the centralized learning approach and the optimal parameter in the objective function. The proof approach is similar to that in the reference (Wang et al., 2019), leading to the following lemma.

Proof. We then utilize the previously defined lemmas to prove Lemma 4.2. By applying the conclusion of Lemma B.2 and summing over the entire interval, we obtain the following formula:

$$\frac{1}{\theta_k(k\tau)} - \frac{1}{\theta_k((k-1)\tau)} = \sum_{z=(k-1)\tau}^{k\tau-1} \left(\frac{1}{\theta_k(z+1)} - \frac{1}{\theta_k(z)} \right) \geq \tau\omega\eta \left(1 - \frac{\beta + \bar{\mu}_k}{2}\eta \right). \tag{35}$$

Next, by summing over all intervals k based on the above conclusion, we derive the following formula:

$$\sum_{k=1}^K \left(\frac{1}{\theta_k(k\tau)} - \frac{1}{\theta_k((k-1)\tau)} \right) \geq \sum_{k=1}^K \tau\omega\eta \left(1 - \frac{\beta + \bar{\mu}_k}{2}\eta \right) = K\omega\eta \left(1 - \frac{\beta + \frac{1}{K} \sum_{k=1}^K \bar{\mu}_k}{2}\eta \right). \tag{36}$$

Rearrange the terms on the left-hand side of the inequality, we obtain the following formula:

$$\frac{1}{\theta_k(T)} - \frac{1}{\theta_1(0)} \geq T\omega\eta \left(1 - \frac{\beta + \frac{1}{K} \sum_{k=1}^K \bar{\mu}_k}{2} \eta\right) + \sum_{k=1}^{K-1} \left(\frac{1}{\theta_{k+1}(k\tau)} - \frac{1}{\theta_k(k\tau)} \right). \quad (37)$$

Next, we derive the case for two consecutive intervals.

$$\begin{aligned} & \frac{1}{\theta_{k+1}(k\tau)} - \frac{1}{\theta_k(k\tau)} \\ &= \frac{\theta_k(k\tau) - \theta_{k+1}(k\tau)}{\theta_k(k\tau) \theta_{k+1}(k\tau)} \\ &\stackrel{(a)}{=} \frac{1}{\theta_k(k\tau) \theta_{k+1}(k\tau)} (F^p(\mathbf{v}_k(k\tau), \bar{\mu}_k) - F^p(\mathbf{w}^*, \bar{\mu}_k) - F^p(\mathbf{v}_{k+1}(k\tau), \bar{\mu}_{k+1}) + F^p(\mathbf{w}^*, \bar{\mu}_{k+1})) \\ &\stackrel{(b)}{=} \frac{1}{\theta_k(k\tau) \theta_{k+1}(k\tau)} (F^p(\mathbf{v}_k(k\tau), \bar{\mu}_k) - F^p(\mathbf{w}(k\tau), \bar{\mu}_{k+1}) \\ &\quad - \frac{\bar{\mu}_k}{2} \|\mathbf{w}^* - \mathbf{w}_{k,global}\|^2 + \frac{\bar{\mu}_{k+1}}{2} \|\mathbf{w}^* - \mathbf{w}_{k+1,global}\|^2) \\ &\stackrel{(c)}{\geq} \frac{-(\rho + \mu\xi)h(\tau) - \frac{\mu\xi^2}{2}}{\theta_k(k\tau) \theta_{k+1}(k\tau)}. \end{aligned} \quad (38)$$

Equation (a) is obtained by substituting the definition of θ_k , equation (b) is further simplified, and inequality (c) follows from Assumption 1 and the upper bound of the proximal term.

Similarly, considering the summation over all intervals, we obtain the following formula:

$$\begin{aligned} \sum_{k=1}^{K-1} \left(\frac{1}{\theta_{k+1}(k\tau)} - \frac{1}{\theta_k(k\tau)} \right) &\geq - \sum_{k=1}^{K-1} \frac{\rho h(\tau) + \frac{1}{2} (\bar{\mu}_{k+1} + \bar{\mu}_k) \xi^2}{\varepsilon^2} \\ &= -(K-1) \frac{\rho h(\tau) + \frac{1}{2} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} + \bar{\mu}_k) \xi^2}{\varepsilon^2}. \end{aligned} \quad (39)$$

By employing addition and subtraction cancellation and rearranging, we obtain the following formula:

$$\frac{1}{\theta_k(T)} - \frac{1}{\theta_1(0)} \geq T\omega\eta \left(1 - \frac{\beta + \frac{1}{K} \sum_{k=1}^K \bar{\mu}_k}{2} \eta\right) - (K-1) \frac{\rho h(\tau) + \frac{1}{2} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} + \bar{\mu}_k) \xi^2}{\varepsilon^2}. \quad (40)$$

Then, subtracting $\frac{1}{\theta_k(T)}$ from the reciprocal of the difference in objective functions with the proximal term yields the following equation:

$$\begin{aligned} & \frac{1}{F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K)} - \frac{1}{\theta_K(T)} \\ &= \frac{\theta_K(T) - (F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K))}{(F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K)) \theta_K(T)} \\ &\stackrel{(a)}{=} \frac{F^p(\mathbf{v}_K(T), \bar{\mu}_K) - F^p(\mathbf{w}(T), \bar{\mu}_K)}{(F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K)) \theta_K(T)} \\ &\stackrel{(b)}{\geq} \frac{-(\rho + \bar{\mu}_K \xi)h(\tau)}{\varepsilon^2}. \end{aligned} \quad (41)$$

Equation (a) is obtained by substituting the definition of θ_k , and inequality (b) is derived by applying Lemma 4.1 along with the preceding derivation.

$$\begin{aligned}
& \frac{1}{F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K)} - \frac{1}{\theta_1(0)} \\
& \stackrel{(a)}{\geq} T\omega\eta \left(1 - \frac{\beta + \frac{1}{K} \sum_{k=1}^K \bar{\mu}_k}{2} \eta \right) - (K-1) \frac{\rho h(\tau) + \frac{1}{2} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} + \bar{\mu}_k) \xi^2}{\varepsilon^2} - \frac{(\rho + \bar{\mu}_K \xi) h(\tau)}{\varepsilon^2} \\
& \stackrel{(b)}{=} T \left(\omega\eta \left(1 - \frac{\beta + \frac{1}{K} \sum_{k=1}^K \bar{\mu}_k}{2} \eta \right) - \frac{\rho h(\tau)}{\tau \varepsilon^2} \right) - \frac{\bar{\mu}_K \xi h(\tau)}{\varepsilon^2} - \frac{\xi^2}{2\varepsilon^2} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} + \bar{\mu}_k). \quad (42)
\end{aligned}$$

Inequality (a) is obtained by combining the two preceding proofs, and equation (b) is further simplified and organized.

Since $\frac{1}{\theta_1(0)} > 0$, by combining the results from the above derivation, we arrive at the conclusion of Lemma 4.2.

$$\begin{aligned}
& \frac{1}{F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K)} \\
& \geq \frac{1}{F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K)} - \frac{1}{\theta_1(0)} \\
& \geq T \left(\omega\eta \left(1 - \frac{\beta + \frac{1}{K} \sum_{k=1}^K \bar{\mu}_k}{2} \eta \right) - \frac{\rho h(\tau)}{\tau \varepsilon^2} \right) - \frac{\bar{\mu}_K \xi h(\tau)}{\varepsilon^2} - \frac{\xi^2}{2\varepsilon^2} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} + \bar{\mu}_k). \quad (43)
\end{aligned}$$

Taking the absolute values of the above results and continuing to simplify yields the following formula.

$$\begin{aligned}
& F^p(\mathbf{w}(T), \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K) \\
& \leq \frac{1}{T \left(\omega\eta \left(1 - \frac{\beta + \frac{1}{K} \sum_{k=1}^K \bar{\mu}_k}{2} \eta \right) - \frac{\rho h(\tau)}{\tau \varepsilon^2} \right) - \frac{\bar{\mu}_K \xi h(\tau)}{\varepsilon^2} - \frac{\xi^2}{2\varepsilon^2} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} + \bar{\mu}_k)} \\
& = \frac{1}{T \left(\eta\varphi - \frac{\rho h(\tau)}{\tau \varepsilon^2} \right) - \frac{\bar{\mu}_K \xi h(\tau)}{\varepsilon^2} - \frac{\xi^2}{2\varepsilon^2} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} + \bar{\mu}_k)}. \quad (44)
\end{aligned}$$

C SUPPLEMENTARY PROOF FOR THEOREM 1

Proof. Here, we assume that the control parameter value for the model parameters resulting from the algorithm output $\mathbf{w}_{k,global}$ is $\bar{\mu}_K$. During the training process, the model parameters continuously approach the optimal quality, so the final output of the algorithm reflects the optimal quality concentrated in the latter part of the training. Additionally, $\mu_{i,k}$ is not updated after every training iteration, the specific update frequency depends on the variations in B_k and H_k . Thus, $\bar{\mu}_K$ can be considered the control parameter value corresponding to $\mathbf{w}_{k,global}$.

However, during the derivation of ε_0 , an additional term related to $\mu_{i,k}$ and ξ appears in the upper bound due to the conclusion of 4.2, leading to the result $\varepsilon_0 =$

$$\frac{1}{2\eta\varphi T} + \sqrt{\frac{1}{4\eta^2\varphi^2 T^2} + \frac{\rho h(\tau) + \frac{\tau}{T} \bar{\mu}_T \xi h(\tau) + \frac{\tau}{2T} \sum_{k=1}^{K-1} (\bar{\mu}_{k+1} - \bar{\mu}_k)}{\eta\varphi T}}. \text{ We obtain } F^p(\mathbf{w}_{k,global}, \bar{\mu}_K) - F^p(\mathbf{w}^*, \bar{\mu}_K) \leq +\varepsilon_0 + (\rho + \bar{\mu}_K \xi) \rho h(\tau), \text{ which leads to the conclusion of Theorem 1.}$$

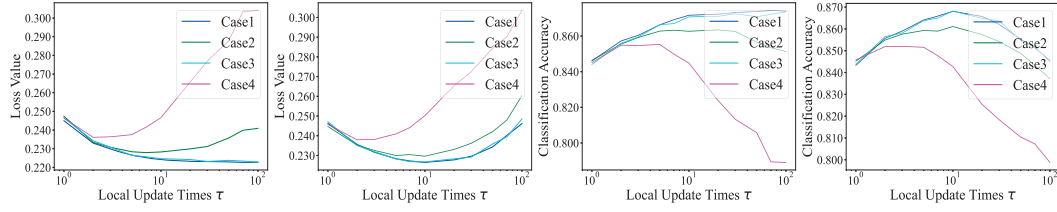


Figure 6: Impact of the optimal local update times τ^* (datasets: MNIST, classifier: SVM): (1) Loss on training data using FedADM, (2) Loss on training data using FedProx, (3) Prediction accuracy on testing data using FedADM, and (4) Prediction accuracy on testing data using FedProx.

D IMPLEMENT DETAILS FOR EXPERIMENTS

D.1 EXPERIMENT ENVIRONMENTS

Table 1: Details of experimental datasets and models

Dataset	Total Images	Training Set	Testing Set	Models
MNIST	70,000	60,000	10,000	SVM/CNN
Fashion-MNIST	70,000	60,000	10,000	SVM
CIFAR-10	60,000	50,000	10,000	CNN

Table 2: Overview of training and control parameters

Parameters	Values
Linear search range γ	10
Maximum local update times τ_{\max}	100
Control parameter φ	SVM: 0.025 CNN: 5×10^{-5}
Gradient descent step size η	0.01
Resource budget R	60 seconds
Constraint value ξ	0.1

D.2 EXPERIMENT RESULTS

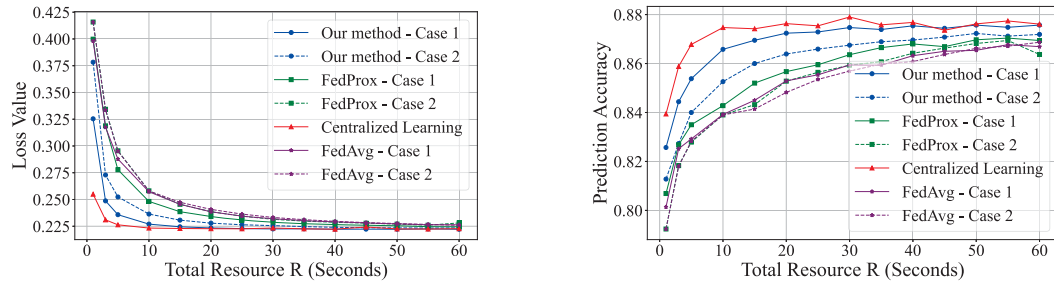


Figure 7: Performance comparison of different methods with different total resource R (datasets: MNIST, classifier: CNN).

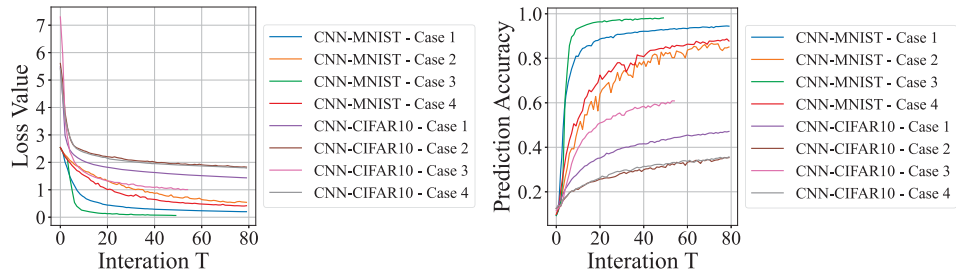


Figure 8: Training performance of our method (dataset: MNIST and CIFAR-10, classifiers: CNN).