Supplementary Material for the Paper: "Semi-Federated Learning Accelerated by Over-the-Air Distortion"

Jingheng Zheng, Hui Tian, *Senior Member, IEEE,* Wanli Ni, *Member, IEEE,*
Yang Tian, and Ping Zhang, *Fellow, IEEE*

In the document, we provide additional simulation results in term of conserving energy consumption. Moreover, we also present the derivations of Theorem 1, Theorem 2, and Lemma 1 in detail.

Appendix A
Architecture of the Adopted ResNet

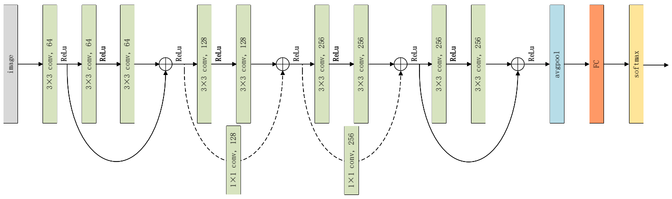
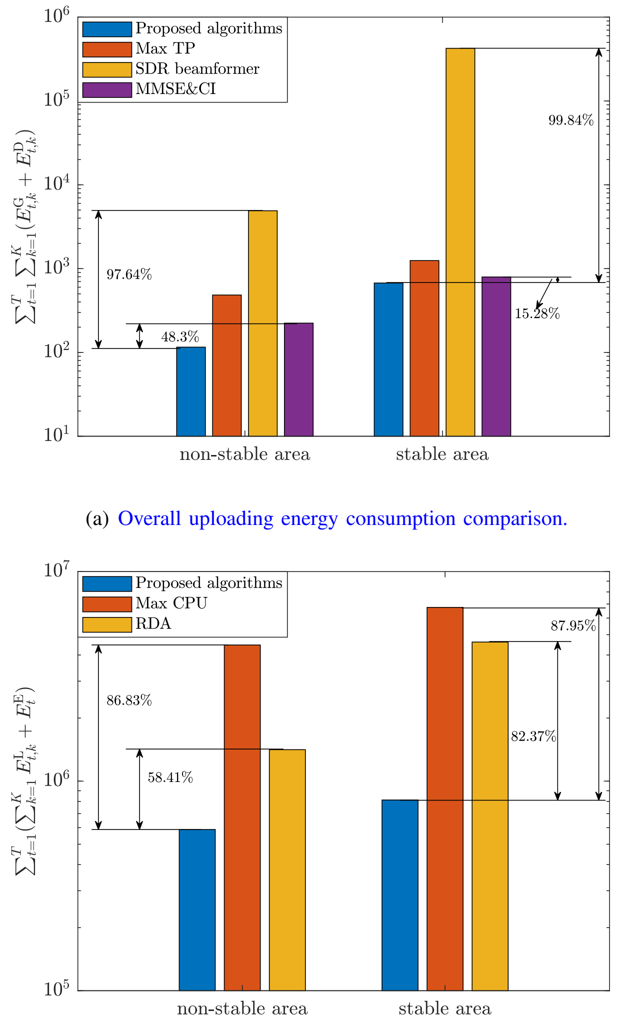


Fig. A1. An architecture demonstration of the adopted ResNet.

The architecture of the adopted ResNet is demonstrated in Fig. A1. The adopted ResNet mainly contains four stacks of layers, where each stack contains two convolutional layers. For the first and fourth stacks, the input is directly added to the output of the two convolutional layers through a skip connection. For the second and third stacks, the input is added to the output of the aforementioned two convolutional layers after applying another convolutional layer to the skip connection. Finally, the four stacks of layers are sequentially followed by an average pooling layer, a fully connected layer, and a softmax layer.

Appendix B
Additional Simulation Results

To validate the effectiveness of the proposed algorithms in conserving energy, we compare with the following baselines:



(a) Overall uploading energy consumption comparison.

(b) Overall computing energy consumption comparison.

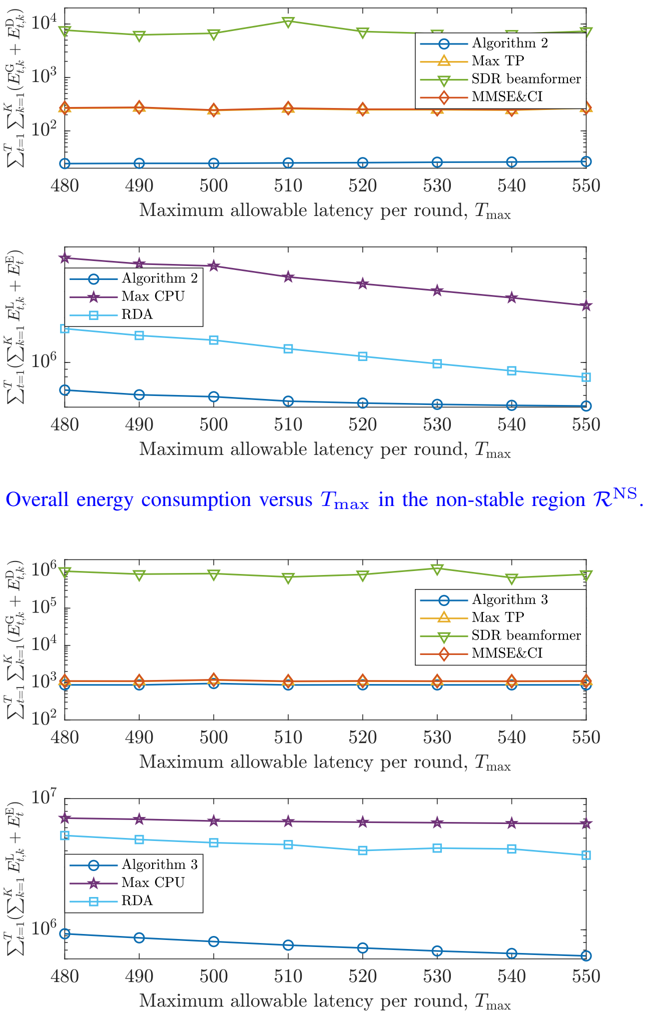
Fig. A2. Overall energy consumption comparison with $\epsilon_1 = 1.2$, $\epsilon_2 = 1$, $\epsilon_3 = 0.8$, and $\epsilon_4 = 0.01$, where $T = 500$ rounds are considered in both the non-stable and stable regions.

- **MMSE&CI** [1]–[3]: By setting $\sqrt{\omega} = \sqrt{\nu}$ in both the non-stable region $\mathcal{R}^{\mathrm{NS}}$ and the stable region $\mathcal{R}^{\mathrm{S}}$, this scheme minimizes MSE to suppress over-the-air distortion throughout the entire SemiFL process.
- **SDR Beamformer** [4]: By dropping the rank-one constraints (44h) and (44i), receive beamformers $\{\mathbf{v}_k\}$ and $\mathbf{b}$ are solved using SDR.

Jingheng Zheng, Hui Tian and Ping Zhang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhengjh@bupt.edu.cn; tianhui@bupt.edu.cn; pzhang@bupt.edu.cn).

Wanli Ni is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: niwanli@tsinghua.edu.cn).

Yang Tian is with the School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 102206, China (e-mail: tianyang9108@163.com).



(a) Overall energy consumption versus $T_{\max}$ in the non-stable region $\mathcal{R}^{\mathrm{NS}}$.

(b) Overall energy consumption versus $T_{\max}$ in the stable region $\mathcal{R}^{\mathrm{S}}$.

Fig. A3. Overall energy consumption versus $T_{\max}$ with $\epsilon_1 = 1.2$, $\epsilon_2 = 1$, $\epsilon_3 = 0.8$, and $\epsilon_4 = 0.01$.

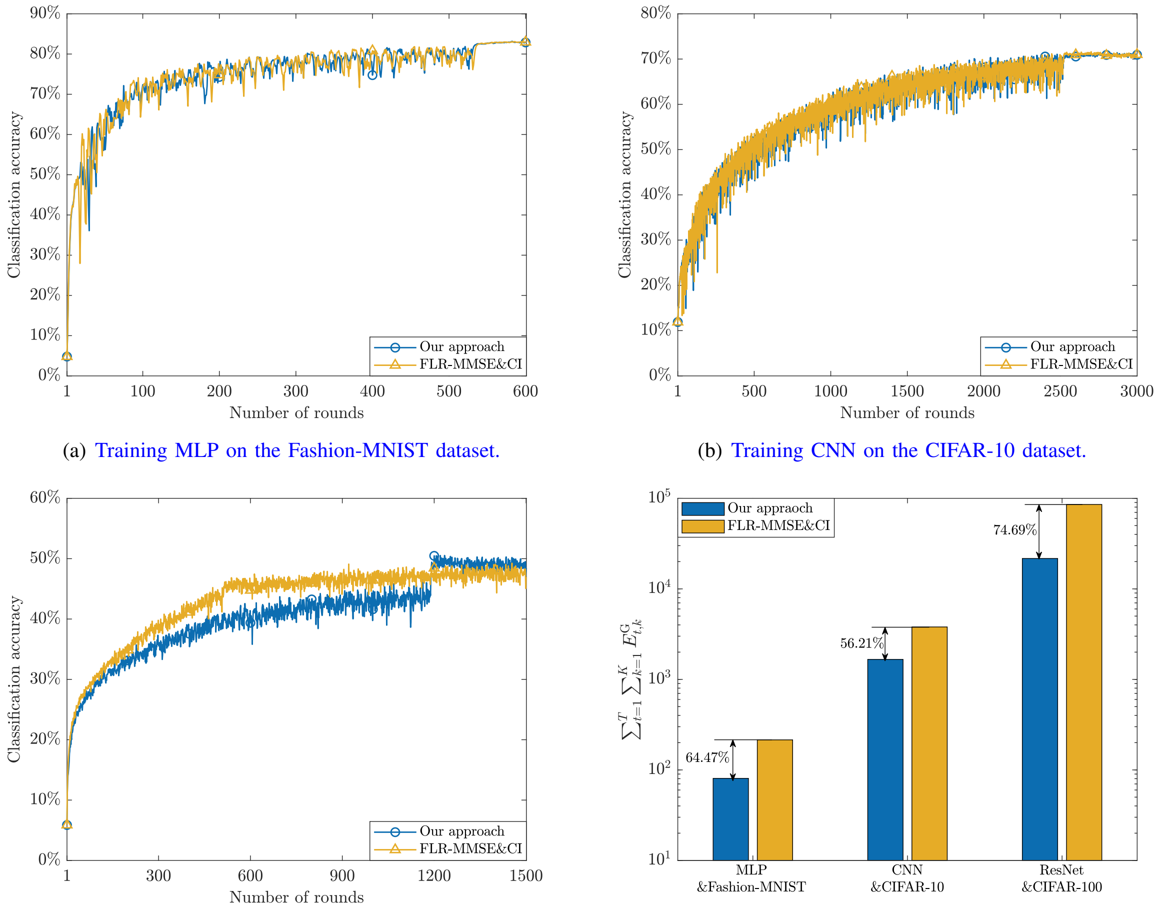
- **Maximum transmit power (Max TP)**: By setting $\zeta_k = p_{\max}|\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k^{\mathrm{DU}}|^2, \forall k \in \mathcal{K}$ and $\omega_k = p_{\max}|\mathbf{b}^{\mathrm{H}}\mathbf{h}_k^{\mathrm{GU}}|^2, \forall k \in \mathcal{K}$, devices use the maximum transmit power to upload gradients and data.
- **Maximum CPU frequencies (Max CPU)**: By setting $\hat{f}_k = \hat{f}_{\max}, \forall k \in \mathcal{K}$ and $\tilde{f} = \tilde{f}_{\max}$, devices and the BS use maximum CPU frequencies to perform FL and SL.
- **Random data allocation (RDA)**: The ratios of SL data, $\{\theta_k\}$, are randomly determined.

Fig. A2 shows the energy consumption of our proposed algorithms in comparison with baselines. Note that $T = 500$ rounds are considered in both regions. To show the performance gains more clearly, the overall energy consumption in objective (35a) is decomposed into two metrics: the overall uploading energy consumption, $\sum_{t=1}^{T}\sum_{k=1}^{K}(E_{t,k}^{\mathrm{G}}+E_{t,k}^{\mathrm{D}})$, and the overall computing energy consumption, $\sum_{t=1}^{T}(\sum_{k=1}^{K}E_{t,k}^{\mathrm{L}}+E_{t}^{\mathrm{E}})$. In Fig. A2(a), it is seen that the proposed algorithms achieve the lowest overall uploading energy consumption in both regions. Particularly, our proposed algorithms conserves 97.64% and 48.3% of uploading energy in the non-stable region, compared to SDR Beamformer and MMSE&CI, respectively. Meanwhile, in the stable region, our proposed algorithms can save 99.84% and 15.28% of uploading energy compared to SDR Beamformer and MMSE&CI schemes, respectively. In Fig. A2(b), it is observed that our proposed algorithms outperform Max CPU and RDA by saving 86.83% and 58.41% of computing energy, respectively, in the non-stable region. Furthermore, our proposed algorithms save 87.95% and 82.37% of computing energy compared to Max CPU and RDA in the stable region, respectively. Additionally, one can find that the proposed algorithms consume more energy in the stable region than the stable region. This is because Algorithm 3 allocates higher transmit power in the stable region to suppress the over-the-air distortion, thereby reducing the optimality gap of SemiFL, as discussed in Remark 3.

Fig. A3 shows the overall uploading and computing energy consumption versus the maximum allowable latency per round, $T_{\max}$. In Fig. A3(a), it is seen that Algorithm 2 achieves lower overall uploading and computing energy consumption than all benchmarks in the non-stable region. Meanwhile, the overall uploading energy consumption is insensitive to changes in $T_{\max}$, whereas the overall computing energy consumption decrease as $T_{\max}$ increases. This is because a larger $T_{\max}$ allows devices and the BS to use lower CPU frequencies, thereby reducing computing energy consumption. In Fig. A3(b), it is seen that Algorithm 3 obtains the lowest energy consumption in the stable region. Since the tendencies of all curves are similar to those in the non-stable region, the same conclusion can be drawn.

Fig. A4 demonstrates the learning performance comparison between our proposed approach and the FLR-MMSE&CI scheme on the considered three datasets. Experiments on the Fashion-MNIST and CIFAR-10 datasets show that FLR-MMSE&CI, which adopts a learning rate equals to $\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}}\eta_t$ while employing the MMSE&CI scheme to suppress over-the-air distortion, achieves nearly identical learning performance to our approach. Note that our approach adopts a fixed learning rate η_t and intentionally introduces the amplitude distortion $\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}}$. In addition, results on the CIFAR-100 dataset show that though FLR-MMSE&CI converges faster than our approach in the non-stable region, our approach which utilizes over-the-air distortion still achieves better final convergence. These findings suggest that the observed learning performance improvements are attributable to our communication-oriented approach, i.e., amplifying over-the-air distortion, particularly amplitude distortion, in the non-stable region while suppressing it in the stable region. Moreover, Fig. A4(d) shows that our approach reduces energy consumption for gradient uploading by 64.47%, 56.21%, and 74.69% on the Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets, respectively. This further underscores our method's superiority in jointly optimizing convergence speed and energy efficiency, compared to FLR-MMSE&CI which treats learning rate adjustment and distortion suppression separately.

To examine the robustness of the proposed approach under



(a) Training MLP on the Fashion-MNIST dataset.

(b) Training CNN on the CIFAR-10 dataset.

(c) Training ResNet on the CIFAR-100 dataset.

(d) Energy consumption for gradient uploading.

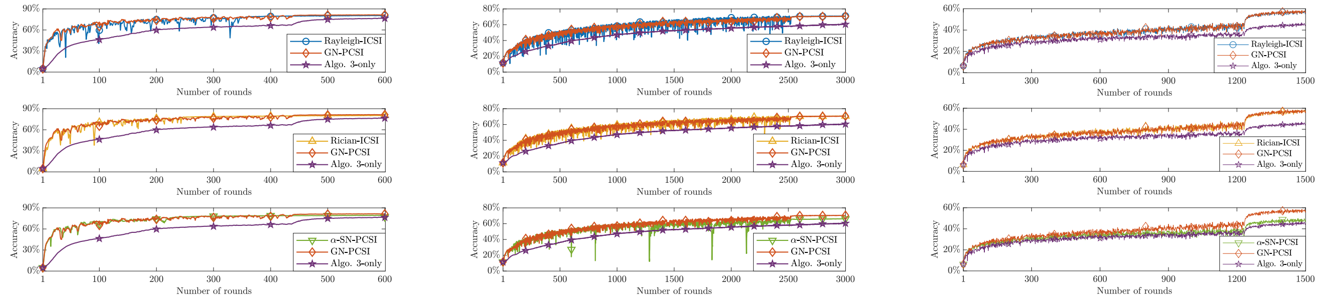
Fig. A4. Learning performance and energy consumption between our approach and the FLR-MMSE&CI scheme.

both imperfect CSI and α-stable noise, we consider the following setting for the channel coefficient and noise as baselines:

- **Rayleigh channel with imperfect CSI (Rayleigh-ICSI)** [5]: The channel coefficient vector is rewritten as $\mathbf{h}_{t,k}^{\mathrm{G}} + \Delta\mathbf{h}_{t,k}^{\mathrm{G}}$, where $\mathbf{h}_{t,k}^{\mathrm{G}}$ denotes the estimated channel which follows a Rayleigh distribution, and $\Delta\mathbf{h}_{t,k}^{\mathrm{G}}$ denotes the estimation error which follows a circularly symmetric complex Gaussian (CSCG) distribution. We set the strength of $\Delta\mathbf{h}_{t,k}^{\mathrm{G}}$ to be 10 times stronger than that of $\mathbf{h}_{t,k}^{\mathrm{G}}$, i.e., $\frac{\|\Delta\mathbf{h}_{t,k}^{\mathrm{G}}\|^2}{\|\mathbf{h}_{t,k}^{\mathrm{G}}\|^2} = 1$. The noise $\mathbf{n}_t^{\mathrm{G}}$ follows a Gaussian distribution
- **Rician channel with imperfect CSI (Rician-ICSI)** [5]: The channel coefficient vector is also rewritten as $\mathbf{h}_{t,k}^{\mathrm{G}} + \Delta\mathbf{h}_{t,k}^{\mathrm{G}}$, whereas $\mathbf{h}_{t,k}^{\mathrm{G}}$ follows a Rician distribution with a Rician factor 10, and $\Delta\mathbf{h}_{t,k}^{\mathrm{G}}$ follows CSCG distribution as well. We also set $\frac{\|\Delta\mathbf{h}_{t,k}^{\mathrm{G}}\|^2}{\|\mathbf{h}_{t,k}^{\mathrm{G}}\|^2} = 1$. The noise $\mathbf{n}_t^{\mathrm{G}}$ follows a Gaussian distribution.
- **α-stable noise with perfect CSI (α-SN-PCSI)** [6]: The channel coefficient vector $\mathbf{h}_{t,k}^{\mathrm{G}}$ follows a Rayleigh distribution. The noise $\mathbf{n}_t^{\mathrm{G}}$ follows a symmetric α-stable distribution. We set the parameter α to $\alpha = 1.4$.
- **Gaussian noise with perfect CSI (GN-PCSI)**:The channel coefficient vector $\mathbf{h}_{t,k}^{\mathrm{G}}$ follows a Rayleigh distribution. The noise $\mathbf{n}_t^{\mathrm{G}}$ follows a Gaussian distribution.

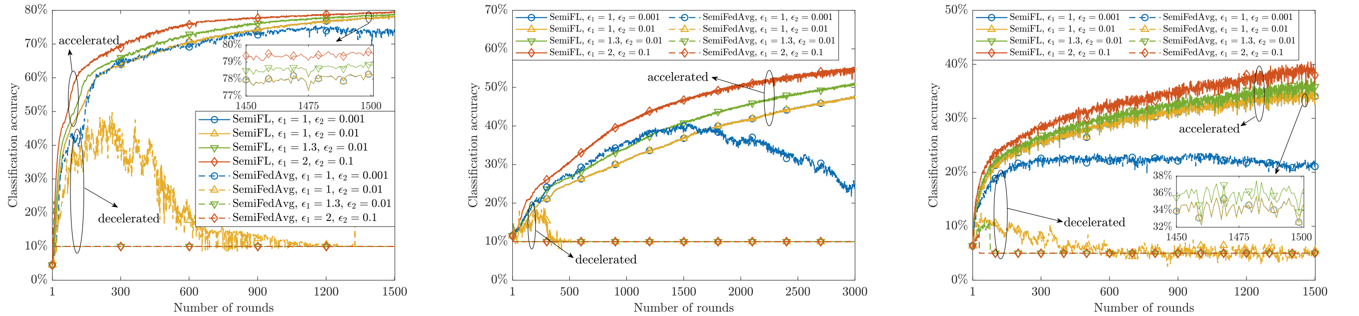
As shown in Fig. A5, SemiFL with the proposed approach, amplifying over-the-air distortion to accelerate convergence, works well across all considered channel and noise conditions. It is seen that SemiFL under the above four network conditions converges faster than Algo. 3-only SemiFL, while gradually approaching the performance of SemiFL in the case of GN-PCSI. This confirms the robustness and effectiveness of the proposed approach under different channel and noise types. Furthermore, it is also noticed that the curves for Rayleigh-ICSI, Rician-ICSI, and α-SN-PCSI schemes exhibit more pronounced fluctuations than GN-PCSI across all experiments. This is because both the imperfect CSI and the α-stable noise introduce stronger interference in gradient aggregation than perfect CSI.

Fig. A6 shows the learning performance comparisons between SemiFL and a benchmark, named SemiFL with federated averaging (SemiFedAvg). The only difference of SemiFedAvg is that devices upload model parameters, i.e., weights and biases, to the BS for aggregation, rather than gradients. In Figs. A6(a), A6(b), and A6(c), it is intriguing to see that as ϵ_1 and ϵ_2 increase, the convergence of SemiFL is gradually accelerated, whereas the convergence of SemiFedAvg is significantly decelerated. This is because over-the-air distortion



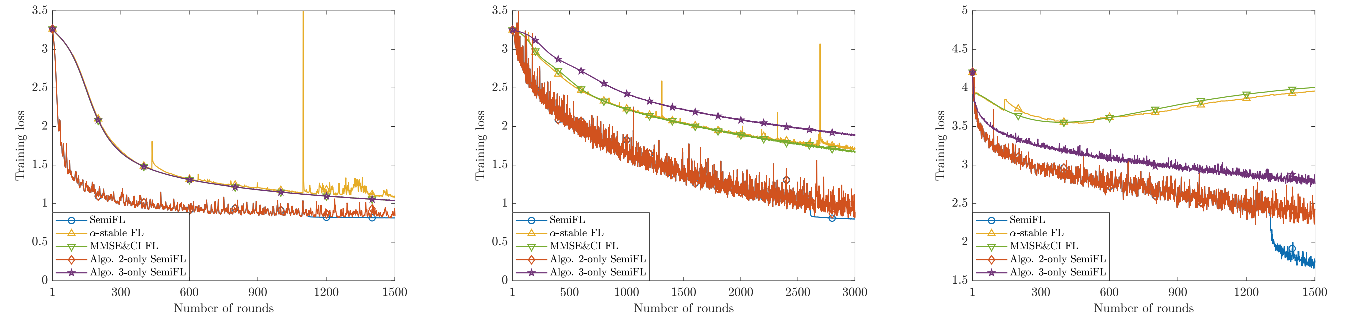
(a) Training MLP on the Fashion-MNIST dataset. (b) Training CNN on the CIFAR-10 dataset. (c) Training ResNet on the CIFAR-100 dataset.

Fig. A5. Learning performance under different network conditions.



(a) Training MLP on the Fashion-MNIST dataset. (b) Training CNN on the CIFAR-10 dataset. (c) Training ResNet on the CIFAR-100 dataset.

Fig. A6. Learning performance comparison between the proposed SemiFL and SemiFedAvg on the Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets with different ϵ_1 and ϵ_2 values, where $\epsilon_3 = 0.8$ and $\epsilon_4 = 0.01$.



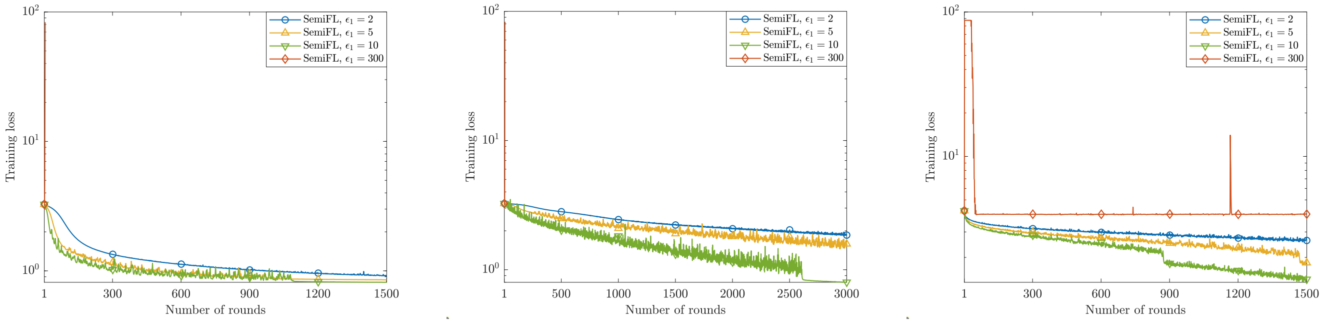
(a) Training MLP on the Fashion-MNIST dataset. (b) Training CNN on the CIFAR-10 dataset. (c) Training ResNet on the CIFAR-100 dataset.

Fig. A7. Training loss comparison between SemiFL and benchmarks on the Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets, where $\epsilon_1 = 10$, $\epsilon_2 = 1$, $\epsilon_3 = 0.8$, and $\epsilon_4 = 0.01$.

is directly imposed on model parameters in SemiFedAvg, fundamentally disrupting the global model. For the proposed SemiFL, over-the-air distortion only affects uploaded gradients, leaving the global model intact. This also highlights a key insight: the convergence acceleration effect of over-the-air distortion is tailored to AirComp-based gradient aggregation. Moreover, Figs. A6(a), A6(b), and A6(c) show that increasing ϵ_2 while keeping ϵ_1 constant cannot trigger the convergence acceleration effect of over-the-air distortion, as evidenced by the overlap of the yellow and blue solid lines. However, it should be emphasized that a large ϵ_2 enables the usage of a large ϵ_1, eventually enhancing the convergence acceleration effect of over-the-air distortion.

As shown in Fig. A7, the proposed SemiFL with our proposed approach, i.e., increasing over-the-air distortion in the non-stable region but suppressing it in the stable region, achieves faster loss function descent than α-stable FL, MMSE&CI FL, and Algo. 3-only SemiFL schemes on all three datasets. This confirms that increasing over-the-air distortion in the non-stable region effectively increases the learning rate, thereby accelerating SemiFL's convergence. Moreover, it is also seen in Fig. A7 that SemiFL with our proposed approach converges to lower training loss than other schemes. This demonstrates that suppressing over-the-air distortion in the stable region helps maintain a small learning rate, thereby facilitating steady and improved final convergence.

Fig. A8 shows the impact of different ϵ_1 values on the training loss of SemiFL. Across all datasets, it is seen



(a) Training MLP on the Fashion-MNIST dataset.

(b) Training CNN on the CIFAR-10 dataset.

(c) Training ResNet on the CIFAR-100 dataset.

Fig. A8. Training loss comparison of the proposed SemiFL on the Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets with different ϵ_1 values, where $\epsilon_3 = 0.8$ and $\epsilon_4 = 0.01$. Note that we set $\epsilon_2 = 1$ when $\epsilon_1 = 2$ or 5, and set $\epsilon_2 = 5$ when $\epsilon_1 = 10$. When $\epsilon_1 = 300$, a sufficiently large ϵ_2 is adopted.

that as ϵ_1 increases, the decent of training loss becomes more pronounced. This is because a larger ϵ_1 value leads to higher over-the-distortion, i.e., $\sqrt{\omega_t}/\sqrt{\nu_t}$, which increases the learning rate, thereby accelerating the convergence of SemiFL in the non-stable region. Meanwhile, it is observed that an excessively large ϵ_1 value causes the training loss curves to vanish for Fashion-MNIST and CIFAR-10 datasets, while making a non-decreasing training loss pattern for the CIFAR-100 dateset. These results indicate that excessive over-the-distortion adversely affects SemiFL, learning to model collapse during training. This necessitates a moderate level of over-the-air distortion that accelerates convergence while maintaining model robustness for SemiFL.

Appendix C
Proof of Theorem 1

Based on (5), (8), and (25) in Assumption 2, we have

$$
\begin{aligned}
&F(\mathbf{w}_t) - F(\mathbf{w}_{t+1})\\
\geq&\eta_t \nabla F(\mathbf{w}_{t+1})^{\mathrm{T}}(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}}\hat{\mathbf{g}}_t^{\mathrm{L}} + \rho_t^{\mathrm{L}}\hat{\mathbf{n}}_t^{\mathrm{G}} + \rho_t^{\mathrm{E}}\mathbf{g}_t^{\mathrm{E}})\\
&+\frac{\mu}{2}\eta_t^2[(\rho_t^{\mathrm{L}})^2\|\mathbf{g}_t^{\mathrm{L}}\|^2 + (\rho_t^{\mathrm{E}})^2\|\mathbf{g}_t^{\mathrm{E}}\|^2 + 2\rho_t^{\mathrm{L}}\rho_t^{\mathrm{E}}(\mathbf{g}_t^{\mathrm{L}})^{\mathrm{T}}\mathbf{g}_t^{\mathrm{E}}]. \quad \text{(A1)}
\end{aligned}
$$

By taking the expectation on both sides, while using Assumption 4, we have

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t+1})]\\
\geq&\eta_t(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})\nabla F(\mathbf{w}_{t+1})^{\mathrm{T}}\nabla F(\mathbf{w}_t)\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2\mathbb{E}[\|\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}}\hat{\mathbf{g}}_t^{\mathrm{L}} + \hat{\mathbf{n}}_t^{\mathrm{G}}\|^2]\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{E}})^2\mathbb{E}[\|\mathbf{g}_t^{\mathrm{E}}\|^2] + \mu\eta_t^2\rho_t^{\mathrm{L}}\rho_t^{\mathrm{E}}\mathbb{E}[(\mathbf{g}_t^{\mathrm{L}})^{\mathrm{T}}\mathbf{g}_t^{\mathrm{E}}]\\
=&\|\frac{\eta_t}{2}(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})\nabla F(\mathbf{w}_{t+1}) + \nabla F(\mathbf{w}_t)\|^2\\
&-\frac{\eta_t^2}{4}(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2\|\nabla F(\mathbf{w}_{t+1})\|^2 - \|\nabla F(\mathbf{w}_t)\|^2\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2(\mathbb{E}[\|\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}}\hat{\mathbf{g}}_t^{\mathrm{L}}\|^2] + \mathbb{E}[\|\hat{\mathbf{n}}_t^{\mathrm{G}}\|^2])\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{E}})^2\mathbb{E}[\|\mathbf{g}_t^{\mathrm{E}}\|^2] + \mu\eta_t^2\rho_t^{\mathrm{L}}\rho_t^{\mathrm{E}}\mathbb{E}[(\mathbf{g}_t^{\mathrm{L}})^{\mathrm{T}}\mathbf{g}_t^{\mathrm{E}}]. \quad \text{(A2)}
\end{aligned}
$$

Then, we incorporate $\|x\| \geq 0, \forall x \in \mathbb{R}$ and Assumption 3 into (A2). As a result, we have

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t+1})]\\
\geq&-\frac{\eta_t^2}{4}(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2A^2 - A^2 + \frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2\frac{\omega_t}{\nu_t}\mathbb{E}[\|\hat{\mathbf{g}}_t^{\mathrm{L}}\|^2]\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{E}})^2\mathbb{E}[\|\mathbf{g}_t^{\mathrm{E}}\|^2] + \frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{n}}_t^{\mathrm{G}}\|^2]\\
&+\mu\eta_t^2\rho_t^{\mathrm{L}}\rho_t^{\mathrm{E}}\mathbb{E}[(\mathbf{g}_t^{\mathrm{L}})^{\mathrm{T}}\mathbf{g}_t^{\mathrm{E}}]\\
\overset{(a)}{\geq}&-\frac{\eta_t^2}{4}(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2A^2 - A^2\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2\frac{\omega_t}{\nu_t}\|\nabla F(\mathbf{w}_t)\|^2 + \frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{E}})^2\|\nabla F(\mathbf{w}_t)\|^2\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{n}}_t^{\mathrm{G}}\|^2] + \mu\eta_t^2\rho_t^{\mathrm{L}}\rho_t^{\mathrm{E}}\mathbb{E}[(\mathbf{g}_t^{\mathrm{L}})^{\mathrm{T}}\mathbf{g}_t^{\mathrm{E}}]\\
\overset{(b)}{=}&-\frac{\eta_t^2}{4}(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2A^2 - A^2\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2\frac{\omega_t}{\nu_t}\|\nabla F(\mathbf{w}_t)\|^2 + \frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{E}})^2\|\nabla F(\mathbf{w}_t)\|^2\\
&+\frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{n}}_t^{\mathrm{G}}\|^2] + \mu\eta_t^2\rho_t^{\mathrm{L}}\rho_t^{\mathrm{E}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}}\|\nabla F(\mathbf{w}_t)\|^2\\
\overset{(c)}{\geq}&-\frac{\eta_t^2}{4}(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2A^2 - A^2 + \frac{\mu}{2}\eta_t^2(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2\varepsilon^2\\
&+\frac{\mu}{2}\eta_t^2\rho_t^{\mathrm{L}}\mathbb{E}[\|\hat{\mathbf{n}}_t^{\mathrm{G}}\|^2], \quad \text{(A3)}
\end{aligned}
$$

where (a) is because $\mathbb{E}[\|\hat{\mathbf{g}}_t^{\mathrm{L}}\|^2] = \sum_{q=1}^{Q}\mathbb{E}[(\hat{g}_{t,q}^{\mathrm{L}})^2] \geq \sum_{q=1}^{Q}(\mathbb{E}[\hat{g}_{t,q}^{\mathrm{L}}])^2 = \|\nabla F(\mathbf{w}_t)\|^2$ and $\mathbb{E}[\|\mathbf{g}_t^{\mathrm{E}}\|^2] = \sum_{q=1}^{Q}\mathbb{E}[(g_{t,q}^{\mathrm{E}})^2] \geq \sum_{q=1}^{Q}(\mathbb{E}[g_{t,q}^{\mathrm{L}}])^2 = \|\nabla F(\mathbf{w}_t)\|^2$. Moreover, (b) is because $\mathbf{g}_t^{\mathrm{L}}$ and $\mathbf{g}_t^{\mathrm{E}}$ are independent, while (c) is because $\|\nabla F(\mathbf{w}_t)\| \geq \varepsilon$ in the non-stable region $\mathcal{R}^{\mathrm{NS}}$.

Recall the definition of $\hat{\mathbf{n}}_t^{\mathrm{G}}$ in (8), one can have

$$
\mathbb{E}[\|\hat{\mathbf{n}}_t^{\mathrm{G}}\|^2] = \sum\nolimits_{q=1}^{Q}\mathbb{E}[(\hat{n}_{t,q}^{\mathrm{G}})^2] = \frac{\sigma^2 Q}{2\nu_t}. \quad \text{(A4)}
$$

Plugging the above equation into (A3), we have

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t+1})]\\
\geq&\frac{\eta_t^2}{4}(2\mu\varepsilon^2 - A^2)(\rho_t^{\mathrm{L}}\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2\\
&-A^2 + \frac{\mu\sigma^2 Q\eta_t^2}{4\nu_t}(\rho_t^{\mathrm{L}})^2. \quad \text{(A5)}
\end{aligned}
$$

By substituting $\rho_t^{\mathrm{E}} = 1 - \rho_t^{\mathrm{L}}$ into (A5), we can obtain (29). The proof is complete.

Appendix D
Proof of Corollary 1

Based on the proof of Theorem 1, one can have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t+1})] \\
\geq&\eta_t \mathbb{E}[\nabla F(\mathbf{w}_{t+1})^{\mathrm{T}}[\rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}}(\hat{\mathbf{g}}_t^{\mathrm{L}} - \mathbf{g}_t^{\mathrm{L}*}) \\
&+ \rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} \mathbf{g}_t^{\mathrm{L}*} + \rho_t^{\mathrm{L}} \hat{\mathbf{n}}_t^{\mathrm{G}} + \rho_t^{\mathrm{E}} \mathbf{g}_t^{\mathrm{E}}]] \\
&+ \frac{\mu}{2}\eta_t^2 (\rho_t^{\mathrm{L}})^2 \frac{\omega_t}{\nu_t} \mathbb{E}[\|\hat{\mathbf{g}}_{t,k}^{\mathrm{L}}\|^2] + \frac{\mu}{2}\eta_t^2 (\rho_t^{\mathrm{E}})^2 \mathbb{E}[\|\mathbf{g}_{t,k}^{\mathrm{E}}\|^2] \\
&+ \frac{\mu}{2}\eta_t^2 (\rho_t^{\mathrm{L}})^2 \mathbb{E}[\|\hat{\mathbf{n}}_t^{\mathrm{G}}\|^2] + \mu\eta^2 \rho_t^{\mathrm{L}} \rho_t^{\mathrm{E}} \mathbb{E}[(\hat{\mathbf{g}}_t^{\mathrm{L}})^{\mathrm{T}} \mathbf{g}_t^{\mathrm{E}}] \\
\geq&\eta_t \rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} \mathbb{E}[\nabla F(\mathbf{w}_{t+1})^{\mathrm{T}}[\frac{1}{K}\sum_{k=1}^{K}(\hat{\mathbf{g}}_{t,k}^{\mathrm{L}} - \mathbf{g}_{t,k}^{\mathrm{L}*})]] \\
&+ \eta_t(\rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}}) \nabla F(\mathbf{w}_{t+1})^{\mathrm{T}} \nabla F(\mathbf{w}_t) \\
&+ \frac{\mu}{2}\eta_t^2 \left(\rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}}\right)^2 \|\nabla F(\mathbf{w}_t)\|^2 + \frac{\mu\sigma^2 Q}{4\nu_t}\eta_t^2(\rho_t^{\mathrm{L}})^2 \\
\geq&-\frac{\eta_t^2}{4}(\frac{\sqrt{\omega_t}}{\sqrt{\nu_t}})^2 \|\nabla F(\mathbf{w}_{t+1})\|^2 - \mathbb{E}[\|\frac{\rho_t^{\mathrm{L}}}{K}\sum_{k=1}^{K}(\hat{\mathbf{g}}_{t,k}^{\mathrm{L}} - \mathbf{g}_{t,k}^{\mathrm{L}*})\|^2] \\
&- \frac{\eta_t^2}{4}(\rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2 \|\nabla F(\mathbf{w}_{t+1})\|^2 - \|\nabla F(\mathbf{w}_t)\|^2 \\
&+ \frac{\mu}{2}\eta_t^2 \left(\rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}}\right)^2 \varepsilon^2 + \frac{\mu\sigma^2 Q}{4\nu_t}\eta_t^2(\rho_t^{\mathrm{L}})^2 \\
\geq&\frac{\mu}{2}\eta_t^2 \left(\rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}}\right)^2 \varepsilon^2 + \frac{\mu\sigma^2 Q}{4\nu_t}\eta_t^2(\rho_t^{\mathrm{L}})^2 \\
&- [\frac{\eta_t^2 \omega_t}{4\nu_t} + \frac{\eta_t^2}{4}(\rho_t^{\mathrm{L}} \frac{\sqrt{\omega_t}}{\sqrt{\nu_t}} + \rho_t^{\mathrm{E}})^2 + 1]A^2 \\
&- \mathbb{E}[\|\frac{\rho_t^{\mathrm{L}}}{K}\sum_{k=1}^{K}(\hat{\mathbf{g}}_{t,k}^{\mathrm{L}} - \mathbf{g}_{t,k}^{\mathrm{L}*})\|^2]. \qquad \text{(A6)}
\end{aligned}$$

Then, for the last term above, we have

$$\begin{aligned}
&-\mathbb{E}[\|\frac{\rho_t^{\mathrm{L}}}{K}\sum_{k=1}^{K}(\hat{\mathbf{g}}_{t,k}^{\mathrm{L}} - \mathbf{g}_{t,k}^{\mathrm{L}*})\|^2] \\
=&-\frac{(\rho_t^{\mathrm{L}})^2}{K^2}\mathbb{E}[\|\sum_{k=1}^{K}\sum_{c=1}^{C}(p_{t,k,c} - \frac{1}{C})\mathbf{g}_{t,k,c}\|^2] \\
\geq&-\frac{(\rho_t^{\mathrm{L}})^2}{K^2}\mathbb{E}[(\sum_{k=1}^{K}\sum_{c=1}^{C}|p_{t,k,c} - \frac{1}{C}|\|\mathbf{g}_{t,k,c}\|)^2] \\
\geq&-\frac{(\rho_t^{\mathrm{L}})^2}{K^2}\mathbb{E}[(\sum_{k=1}^{K}\sum_{c=1}^{C}(p_{t,k,c} - \frac{1}{C})^2)(\sum_{k=1}^{K}\sum_{c=1}^{C}\|\mathbf{g}_{t,k,c}\|^2)] \\
\geq&-\frac{(\rho_t^{\mathrm{L}} A^2 C)^2}{K^2}\sum_{k=1}^{K}\sum_{c=1}^{C}(p_{t,k,c} - \frac{1}{C})^2 \qquad \text{(A7)}
\end{aligned}$$

By plugging (A7) into (A6), one can have (31). The proof is complete.

Appendix E
Proof of Theorem 2

Based on (24) in Assumption 1, by using $\sqrt{\omega_t}/\sqrt{\nu_t} = 1$, we have

$$\begin{aligned}
&F(\mathbf{w}_t) - F(\mathbf{w}_{t-1}) \\
\leq&-\eta_{t-1}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}(\rho_{t-1}^{\mathrm{L}}\mathbf{g}_{t-1}^{\mathrm{L}} + \rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}}) \\
&+ \frac{L}{2}\eta_{t-1}^2\|\rho_{t-1}^{\mathrm{L}}\hat{\mathbf{g}}_{t-1}^{\mathrm{L}} + \rho_{t-1}^{\mathrm{L}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}} + \rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}}\|^2 \\
&- \eta_{t-1}\rho_{t-1}^{\mathrm{L}}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}. \qquad \text{(A8)}
\end{aligned}$$

Taking the expectation on both sides of (A8), we derive that

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \\
\leq&-\eta_{t-1}\|\nabla F(\mathbf{w}_{t-1})\|^2 + \frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{g}}_{t-1}^{\mathrm{L}}\|^2] \\
&+ \frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{E}})^2\mathbb{E}[\|\mathbf{g}_{t-1}^{\mathrm{E}}\|^2] + \frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}\|^2] \\
&+ L\eta_{t-1}^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}}\mathbb{E}[(\hat{\mathbf{g}}_{t-1}^{\mathrm{L}})^{\mathrm{T}}\mathbf{g}_{t-1}^{\mathrm{E}}] \\
\leq&(L\eta_{t-1}^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}} - \eta_{t-1})\|\nabla F(\mathbf{w}_{t-1})\|^2 \\
&+\frac{L}{2}\eta_{t-1}^2 A^2[(\rho_{t-1}^{\mathrm{L}})^2+(\rho_{t-1}^{\mathrm{E}})^2]+\frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\frac{\sigma^2 Q}{2\nu_t}. \qquad \text{(A9)}
\end{aligned}$$

Then, we employ the PL-inequality from our previous work [7], which results in the following result:

$$\|\nabla F(\mathbf{w}_{t-1})\|^2 \geq 2\mu[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)]. \qquad \text{(A10)}$$

Correspondingly, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \\
\leq&2\mu(L\eta_{t-1}^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}} - \eta_{t-1})[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] \\
&+ \frac{L}{2}A^2\eta_{t-1}^2 + \frac{L\sigma^2 Q}{4\nu_t}(\rho_{t-1}^{\mathrm{L}})^2\eta_{t-1}^2. \qquad \text{(A11)}
\end{aligned}$$

By setting η_{t-1} to $\eta_{t-1} = 1/\mu$ while taking the expectation on both sides, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \\
\leq&[1 - 2(1 - \frac{L}{\mu}\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}})]\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] \\
&+ \frac{L}{2\mu^2}[A^2 + \frac{\sigma^2 Q}{2\nu_t}(\rho_{t-1}^{\mathrm{L}})^2]. \qquad \text{(A12)}
\end{aligned}$$

Based on $\rho_{t-1}^{\mathrm{E}} = 1 - \rho_{t-1}^{\mathrm{L}}$, it is noticed that

$$\begin{aligned}
&1 - 2(1 - \frac{L}{\mu}\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}}) \\
=&-2\frac{L}{\mu}(\rho_{t-1}^{\mathrm{L}})^2 + 2\frac{L}{\mu}\rho_{t-1}^{\mathrm{L}} - 1 \leq \frac{L}{2\mu} - 1 \triangleq \xi. \qquad \text{(A13)}
\end{aligned}$$

By using (A13) and setting ν_t to $\nu_t = \nu, \forall t \in \mathcal{T}$, we can further derive that

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \\
\leq&\xi\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] + \frac{L}{2\mu^2}[A^2 + \frac{\sigma^2 Q}{2\nu}(\rho_{t-1}^{\mathrm{L}})^2] \\
\overset{(d)}{\leq}&\xi\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] + \frac{L}{2\mu^2}(A^2 + \frac{\sigma^2 Q}{2\nu}). \qquad \text{(A14)}
\end{aligned}$$

Recursively applying inequality (A14), we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \\
\leq&\xi^{t-1}\mathbb{E}[F(\mathbf{w}_1)-F(\mathbf{w}^*)]+\frac{1-\xi^{t-1}}{1-\xi}\frac{L}{2\mu^2}(A^2+\frac{\sigma^2 Q}{2\nu}). \qquad \text{(A15)}
\end{aligned}$$

Lastly, (32) can be obtained by letting the both sides of (A15) approach infinity. The proof is complete.

Appendix F
Proof of Corollary 2

Based on the proof of Theorem 2, we have

$$\begin{aligned}
&F(\mathbf{w}_t)-F(\mathbf{w}_{t-1})\\
\leq&-\eta_{t-1}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}(\rho_{t-1}^{\mathrm{L}}\mathbf{g}_{t-1}^{\mathrm{L}}+\rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}})\\
&+\frac{L}{2}\eta_{t-1}^2\|\rho_{t-1}^{\mathrm{L}}(\hat{\mathbf{g}}_{t-1}^{\mathrm{L}}-\mathbf{g}_{t-1}^{\mathrm{L}*})+\rho_{t-1}^{\mathrm{L}}\mathbf{g}_{t-1}^{\mathrm{L}*}\\
&+\rho_{t-1}^{\mathrm{L}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}+\rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}}\|^2\\
&-\eta_{t-1}\rho_{t-1}^{\mathrm{L}}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}\\
\leq&-\eta_{t-1}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}(\rho_{t-1}^{\mathrm{L}}\mathbf{g}_{t-1}^{\mathrm{L}}+\rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}})\\
&+L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\|\hat{\mathbf{g}}_{t-1}^{\mathrm{L}}-\mathbf{g}_{t-1}^{\mathrm{L}*}\|^2\\
&+L\eta_{t-1}^2\|\rho_{t-1}^{\mathrm{L}}\mathbf{g}_{t-1}^{\mathrm{L}*}+\rho_{t-1}^{\mathrm{L}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}+\rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}}\|^2\\
&-\eta_{t-1}\rho_{t-1}^{\mathrm{L}}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}.
\end{aligned}\tag{A16}$$

By taking the expectation on both sides, one can have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}_{t-1})]\\
\leq&-\eta_{t-1}\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1})\|^2]+L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\mathbb{E}[\|\mathbf{g}_{t-1}^{\mathrm{L}*}\|^2]\\
&+L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}\|^2]+L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{E}})^2\mathbb{E}[\|\mathbf{g}_{t-1}^{\mathrm{E}}\|^2]\\
&+2L\eta_{t-1}^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}}\mathbb{E}[(\mathbf{g}_{t-1}^{\mathrm{L}*})^{\mathrm{T}}\mathbf{g}_{t-1}^{\mathrm{E}}]\\
&+L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}(\hat{\mathbf{g}}_{t-1,k}^{\mathrm{L}}-\mathbf{g}_{t-1,k}^{\mathrm{L}*})\|^2]\\
\leq&(2L\eta_t^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}}-\eta_{t-1})\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1})\|^2]\\
&+L\eta_{t-1}^2[(\rho_{t-1}^{\mathrm{L}})^2+(\rho_{t-1}^{\mathrm{E}})^2]A^2+L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\frac{\sigma^2Q}{2\nu_t}\\
&+\frac{C}{K}A^2L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2[\sum_{k=1}^{K}\sum_{c=1}^{C}(p_{t-1,k,c}-\frac{1}{C})^2]\\
\leq&2\mu(2L\eta_t^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}}-\eta_{t-1})\mathbb{E}[F(\mathbf{w}_{t-1})-F(\mathbf{w}^*)]\\
&+L\eta_{t-1}^2A^2+L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\frac{\sigma^2Q}{2\nu_t}\\
&+\frac{C}{K}A^2L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2[\sum_{k=1}^{K}\sum_{c=1}^{C}(p_{t-1,k,c}-\frac{1}{C})^2].
\end{aligned}\tag{A17}$$

By setting $\eta_t=\frac{1}{\mu}$ and subtracting $F(\mathbf{w}^*)$ from both sides, we have the following inequality after taking expectation on both sides:

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}_{t-1})]\\
\leq&[1-2(1-2\frac{L}{\mu}\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}})]\mathbb{E}[F(\mathbf{w}_{t-1})-F(\mathbf{w}^*)]\\
&+\frac{L}{\mu^2}[A^2+\frac{\sigma^2Q}{2\nu_t}(\rho_{t-1}^{\mathrm{L}})^2]\\
&+\frac{C}{K}A^2L\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2[\sum_{k=1}^{K}\sum_{c=1}^{C}(p_{t-1,k,c}-\frac{1}{C})^2].
\end{aligned}\tag{A18}$$

Then, we let $\hat{\xi}=1-2(1-2\frac{L}{\mu}\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}})$. Through using the above results, while setting $\nu_t=\nu,\forall t$, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}^*)]\\
\leq&\hat{\xi}\mathbb{E}[F(\mathbf{w}_{t-1})-F(\mathbf{w}^*)]+\frac{L}{\mu^2}(A^2+\frac{\sigma^2Q}{2\nu})\\
&+\frac{CA^2L}{K\mu^2}(\rho_{t-1}^{\mathrm{L}})^2\Delta d_{t-1}.
\end{aligned}\tag{A19}$$

Recursively applying the above inequality for t times, we have

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}^*)]\leq&\hat{\xi}^{t-1}\mathbb{E}[F(\mathbf{w}_1)-F(\mathbf{w}^*)]\\
&+\frac{1-\hat{\xi}^{t-1}}{1-\hat{\xi}}\frac{L}{\mu^2}(A^2+\frac{\sigma^2Q}{2\nu})\\
&+\frac{CA^2L}{K\mu^2}\sum_{\tau=1}^{t-1}\hat{\xi}^{t-1-\tau}(\rho_{\tau}^{\mathrm{L}})^2\Delta d_{\tau}.
\end{aligned}\tag{A20}$$

By forcing $t\to\infty$, we have (33). The poof is complete.

Appendix G
Proof of Lemma 1

The Lagrange function of problem (50a) is given by

$$\begin{aligned}
&\mathcal{L}(\{\hat{f}_k\},\tilde{f},\tau_2,\lambda_1,\{\lambda_{2,k}\},\lambda_3,\{\lambda_{4,k}\},\{\lambda_{5,k}\},\lambda_6,\lambda_7)\\
=&\sum\nolimits_{k=1}^{K}C_{11,k}\hat{f}_k^2+C_{12}\tilde{f}^2\\
&+\lambda_1(\tau_2-T_{\max})+\sum\nolimits_{k=1}^{K}\lambda_{2,k}(\frac{C_{13,k}}{\hat{f}_k}-\tau_2+T^{\mathrm{G}})\\
&+\lambda_3(\frac{C_{14}}{\tilde{f}}-\tau_2+\max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\})+\sum\nolimits_{k=1}^{K}\lambda_{4,k}(-\hat{f}_k)\\
&+\sum\nolimits_{k=1}^{K}\lambda_{5,k}(\hat{f}_k-\hat{f}_{\max})+\lambda_6(-\tilde{f})+\lambda_7(\tilde{f}-\tilde{f}_{\max}),
\end{aligned}\tag{A21}$$

where λ_1, $\{\lambda_{2,k}\}$, λ_3, $\{\lambda_{4,k}\}$, $\{\lambda_{5,k}\}$, λ_6, and λ_7 are non-negative Lagrange multipliers. Then, the KKT conditions of problem (50a) are given by

$$\begin{cases}
\frac{\partial\mathcal{L}}{\partial\hat{f}_k}=0,\forall k\in\mathcal{K}, & \text{(A22a)}\\
\frac{\partial\mathcal{L}}{\partial\tilde{f}}=0, & \text{(A22b)}\\
\frac{\partial\mathcal{L}}{\partial\tau_2}=0, & \text{(A22c)}\\
\lambda_1(\tau_2-T_{\max})=0, & \text{(A22d)}\\
\lambda_{2,k}(\frac{C_{13,k}}{\hat{f}_k}-\tau_2+T^{\mathrm{G}})=0,\forall k\in\mathcal{K}, & \text{(A22e)}\\
\lambda_3(\frac{C_{14}}{\tilde{f}}-\tau_2+\max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\})=0, & \text{(A22f)}\\
\lambda_{4,k}(-\hat{f}_k)=0,\forall k\in\mathcal{K}, & \text{(A22g)}\\
\lambda_{5,k}(\hat{f}_k-\hat{f}_{\max})=0,\forall k\in\mathcal{K}, & \text{(A22h)}\\
\lambda_6(-\tilde{f})=0, & \text{(A22i)}\\
\lambda_7(\tilde{f}-\tilde{f}_{\max})=0, & \text{(A22j)}\\
\lambda_1\geq0,\ \lambda_3\geq0,\ \lambda_6\geq0,\ \lambda_7\geq0, & \\
\lambda_{2,k}\geq0,\ \lambda_{4,k}\geq0,\ \lambda_{5,k}\geq0,\forall k\in\mathcal{K}. & \text{(A22k)}
\end{cases}$$

It is noticed that constraint (50c) can be rewritten as

$$\hat{f}_k\geq\frac{C_{13,k}}{\tau_2-T^{\mathrm{G}}},\forall k\in\mathcal{K}.\tag{A23}$$

To minimize the objective (50a), $\hat{f}_k$ should be minimized within the feasible region. Moreover, the right-hand side of inequality (A23) monotonously decreases as τ_2 increases. Since constraint (50a), i.e., $\tau_2\leq T_{\max}$, should be satisfied, the right-hand side of inequality (A23) obtains its minimum when $\tau_2=T_{\max}$. Hence, the optimal τ_2 and $\hat{f}_k$ can given by

(A24) and (A25), respectively, i.e., given by

$$\tau_2^* = T_{\max}, \tag{A24}$$

$$\hat{f}_k^* = \frac{C_{13,k}}{T_{\max} - T^{\mathrm{G}}}, \forall k \in \mathcal{K}. \tag{A25}$$

similarly, constraint (50d) can be written as

$$\tilde{f} \geq \frac{C_{14}}{\tau_2 - \max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\}}. \tag{A26}$$

Considering that the objective (50a) increases as $\tilde{f}$ increases, $\tilde{f}$ should also be minimized to minimize the objective (50a). In addition, the right-hand side of (A26) monotonously decreases as τ_2 increases. Intuitively, $\tau_2 = T_{\max}$ should be imposed to minimize the right-hand side of (A26), such that objective (50a) can be minimized as well. Therefore, the optimal $\tilde{f}$ can be given by

$$\tilde{f}^* = \frac{C_{14}}{T_{\max} - \max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\}}. \tag{A27}$$

To meet constraints (50c) and (50d), it is clear that $\hat{f}_k \neq 0, \forall k \in \mathcal{K}$ and $\tilde{f} \neq 0$, so as to prevent excessive FL and CL computing latency. Hence, we have $\lambda_{4,k} = 0, \forall k \in \mathcal{K}$ and $\lambda_6 = 0$. Additionally, by applying (A24), (A25), and (A27) to (A22a), (A22b), and (A22c), we have

$$\begin{cases} \lambda_1 = \sum_{k=1}^{K} \frac{2C_{11,k}(\hat{f}_k^*)^3}{C_{13,k}} + \frac{2C_{12}(\tilde{f}^*)^3}{C_{14}}, & \text{(A28a)} \\ \lambda_{2,k} = \frac{2C_{11,k}(\hat{f}_k^*)^3}{C_{13,k}}, \forall k \in \mathcal{K}, & \text{(A28b)} \\ \lambda_3 = \frac{2C_{12}(\tilde{f}^*)^3}{C_{14}}, & \text{(A28c)} \\ \lambda_{5,k} = 0, \forall k \in \mathcal{K}, & \text{(A28d)} \\ \lambda_7 = 0. & \text{(A28e)} \end{cases}$$

The proof is complete.

References

[1] W. Ni, Y. Liu, Z. Yang, H. Tian, and X. Shen, “Federated learning in multi-RIS-aided systems,” *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9608–9624, Jun. 2022.

[2] W. Liu, X. Zang, Y. Li, and B. Vucetic, “Over-the-air computation systems: Optimization, analysis and scaling laws,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5488–5502, Aug. 2020.

[3] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, “Dynamic scheduling for over-the-air federated edge learning with energy constraints,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 227–242, Jan. 2022.

[4] Z.-q. Luo, W.-k. Ma, A. M.-c. So, Y. Ye, and S. Zhang, “Semidefinite relaxation of quadratic optimization problems,” *IEEE Signal Process Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[5] J. Zheng, H. Tian, W. Ni, W. Ni, and P. Zhang, “Balancing accuracy and integrity for reconfigurable intelligent surface-aided over-the-air federated learning,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10 964–10 980, Jul. 2022.

[6] H. H. Yang, Z. Chen, T. Q. S. Quek, and H. V. Poor, “Revisiting analog over-the-air machine learning: The blessing and curse of interference,” *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 3, pp. 406–419, Apr. 2022.

[7] J. Zheng, W. Ni, H. Tian, D. Gündüz, T. Q. S. Quek, and Z. Han, “Semi-federated learning: Convergence analysis and optimization of a hybrid learning framework,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9438–9456, Dec. 2023.