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*

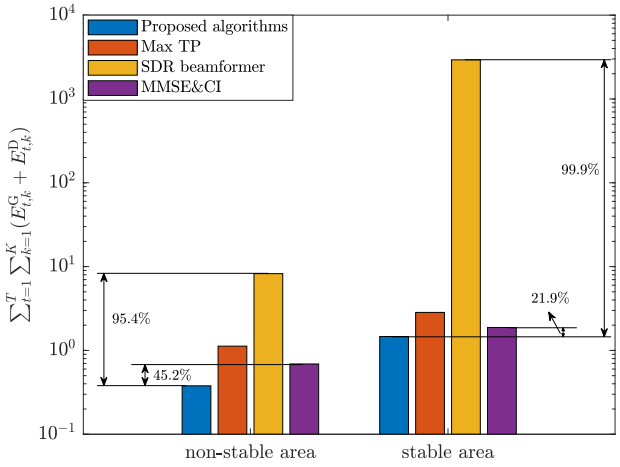
In the document, we provide additional simulation results in term of conserving energy consumption. Moreover, we also present the derivations of Theorem 1, Theorem 2, and Lemma 1 in detail.

Appendix A
Additional Simulation Results

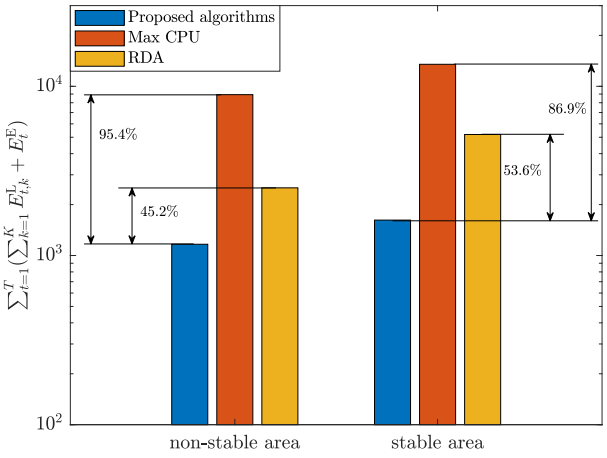
To validate the effectiveness of the proposed algorithms in conserving energy, we compare with the following baselines:

- **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 (41h) and (41i), receive beamformers $\{\mathbf{v}_k\}$ and $\mathbf{b}$ are solved using SDR.
- **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 CL.
- **Random data allocation (RDA)**: The ratios of CL data, $\{\theta_k\}$, are randomly determined.

Fig. A1 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 (32a) 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. A1(a), it is seen that the proposed algorithms achieve the lowest overall uploading energy consumption in both regions. Particularly, our proposed algorithms conserves 95.4% and 45.2% 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 99.9% and 21.9% of uploading energy compared to SDR Beamformer and MMSE&CI schemes, respectively. In Fig. A1(b), it is observed that our proposed algorithms outperform Max CPU and RDA by saving 94.5% and 45.2% of computing energy, respectively,



(a) Overall uploading energy consumption comparison.



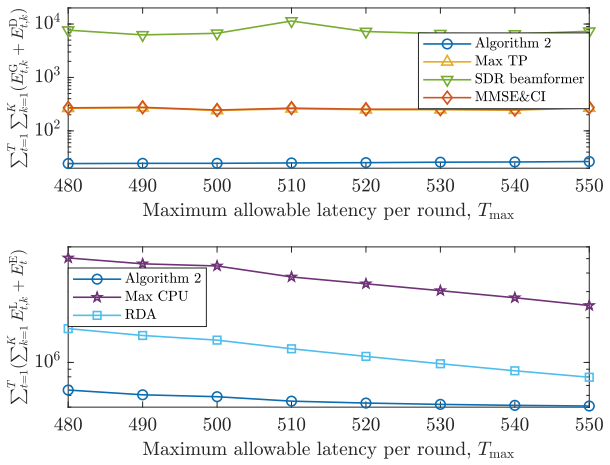
(b) Overall computing energy consumption comparison.

Fig. A1. 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.

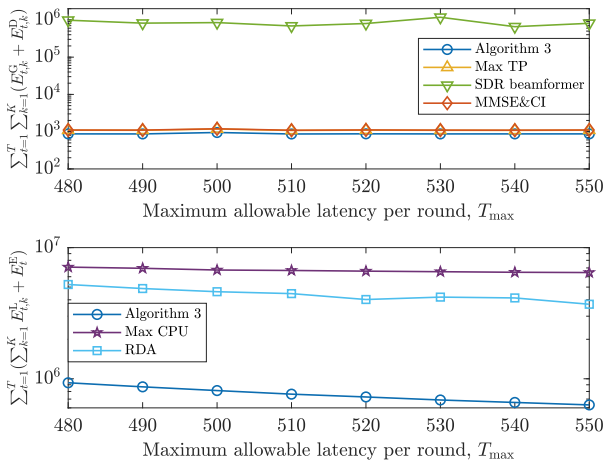
Jingheng Zheng, Hui Tian and Ping Zhang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhengjh@bupt.edu.cn; tianhui@bupt.edu.cn; pzhang@bupt.edu.cn).

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. A2. Overall energy consumption versus $T_{\max}$ with $\epsilon_1 = 1.2$, $\epsilon_2 = 1$, $\epsilon_3 = 0.8$, and $\epsilon_4 = 0.01$.

in the non-stable region. Furthermore, our proposed algorithms save 86.9% and 53.6% 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 2.

Fig. A2 shows the overall uploading and computing energy consumption versus the maximum allowable latency per round, $T_{\max}$. In Fig. A2(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. A2(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.

Appendix B
Proof of Theorem 1

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

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

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

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

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}})^2 A^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}})^2 A^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}})^2 A^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}})^2 A^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], \qquad \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

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

Plugging the above equation into (A3), we have

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

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

Appendix C
Proof of Theorem 2

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

$$\begin{aligned}&F(\mathbf{w}_t) - F(\mathbf{w}_{t-1}) \\ \leq& -\eta_{t-1}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}(\rho_{t-1}^{\mathrm{L}}\hat{\mathbf{g}}_{t-1}^{\mathrm{L}} + \rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}}) \\ &+ \frac{L}{2}\eta_{t-1}^2\|\rho_{t-1}^{\mathrm{L}}\hat{\mathbf{g}}_{t-1}^{\mathrm{L}} + \rho_{t-1}^{\mathrm{L}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}} + \rho_{t-1}^{\mathrm{E}}\mathbf{g}_{t-1}^{\mathrm{E}}\|^2 \\ &- \eta_{t-1}\rho_{t-1}^{\mathrm{L}}\nabla F(\mathbf{w}_{t-1})^{\mathrm{T}}\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}\end{aligned} \tag{A6}$$

Taking the expectation on both sides of (A6), we derive that

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \\ \leq& -\eta_{t-1}\|\nabla F(\mathbf{w}_{t-1})\|^2 + \frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{g}}_{t-1}^{\mathrm{L}}\|^2] \\ &+ \frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{E}})^2\mathbb{E}[\|\mathbf{g}_{t-1}^{\mathrm{E}}\|^2] + \frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\mathbb{E}[\|\hat{\mathbf{n}}_{t-1}^{\mathrm{G}}\|^2] \\ &+ L\eta_{t-1}^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}}\mathbb{E}[(\hat{\mathbf{g}}_{t-1}^{\mathrm{L}})^{\mathrm{T}}\mathbf{g}_{t-1}^{\mathrm{E}}] \\ \leq& (L\eta_{t-1}^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}} - \eta_{t-1})\|\nabla F(\mathbf{w}_{t-1})\|^2 \\ &+ \frac{L}{2}\eta_{t-1}^2A^2[(\rho_{t-1}^{\mathrm{L}})^2 + (\rho_{t-1}^{\mathrm{E}})^2] + \frac{L}{2}\eta_{t-1}^2(\rho_{t-1}^{\mathrm{L}})^2\frac{\sigma^2 Q}{2\nu_t}.\end{aligned} \tag{A7}$$

Then, we employ the PL-inequality [5], i.e.,

$$\|\nabla F(\mathbf{w}_{t-1})\|^2 \geq 2\mu[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)]. \tag{A8}$$

Correspondingly, we have

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_{t-1})] \\ \leq& 2\mu(L\eta_{t-1}^2\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}} - \eta_{t-1})[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] \\ &+ \frac{L}{2}A^2\eta_{t-1}^2 + \frac{L\sigma^2 Q}{4\nu_t}(\rho_{t-1}^{\mathrm{L}})^2\eta_{t-1}^2.\end{aligned} \tag{A9}$$

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

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \\ \leq& [1 - 2(1 - \frac{L}{\mu}\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}})]\mathbb{E}[F(\mathbf{w}_{t-1}) - F(\mathbf{w}^*)] \\ &+ \frac{L}{2\mu^2}[A^2 + \frac{\sigma^2 Q}{2\nu_t}(\rho_{t-1}^{\mathrm{L}})^2].\end{aligned} \tag{A10}$$

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

$$\begin{aligned}&1 - 2(1 - \frac{L}{\mu}\rho_{t-1}^{\mathrm{L}}\rho_{t-1}^{\mathrm{E}}) \\ =& -2\frac{L}{\mu}(\rho_{t-1}^{\mathrm{L}})^2 + 2\frac{L}{\mu}\rho_{t-1}^{\mathrm{L}} - 1 \leq \frac{L}{2\mu} - 1 \triangleq \xi.\end{aligned} \tag{A11}$$

By using (A11) and setting ν_t to $\nu_t = \nu, \forall t \in \mathcal{T}$, we can further derive that

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

Recursively applying inequality (A12), we have

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \\ \leq& \xi^{t-1}\mathbb{E}[F(\mathbf{w}_1) - F(\mathbf{w}^*)] + \frac{1-\xi^{t-1}}{1-\xi}\frac{L}{2\mu^2}(A^2 + \frac{\sigma^2 Q}{2\nu}).\end{aligned} \tag{A13}$$

Lastly, (30) can be obtained by letting the both sides of (A13) approach infinity. The proof is complete.

Appendix D
Proof of Lemma 1

The Lagrange function of problem (47a) is given by

$$\begin{aligned}&\mathcal{L}(\{\hat{f}_k\}, \tilde{f}, \tau_2, \lambda_1, \{\lambda_{2,k}\}, \lambda_3, \{\lambda_{4,k}\}, \{\lambda_{5,k}\}, \lambda_6, \lambda_7) \\ =& \sum\nolimits_{k=1}^{K} C_{11,k}\hat{f}_k^2 + C_{12}\tilde{f}^2 \\ &+ \lambda_1(\tau_2 - T_{\max}) + \sum\nolimits_{k=1}^{K}\lambda_{2,k}(\frac{C_{13,k}}{\hat{f}_k} - \tau_2 + T^{\mathrm{G}}) \\ &+ \lambda_3(\frac{C_{14}}{\tilde{f}} - \tau_2 + \max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\}) + \sum\nolimits_{k=1}^{K}\lambda_{4,k}(-\hat{f}_k) \\ &+ \sum\nolimits_{k=1}^{K}\lambda_{5,k}(\hat{f}_k - \hat{f}_{\max}) + \lambda_6(-\tilde{f}) + \lambda_7(\tilde{f} - \tilde{f}_{\max}),\end{aligned} \tag{A14}$$

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 (47a) are given by

$$\begin{cases}\frac{\partial\mathcal{L}}{\partial\hat{f}_k} = 0, \forall k \in \mathcal{K}, & \text{(A15a)}\\ \frac{\partial\mathcal{L}}{\partial\tilde{f}} = 0, & \text{(A15b)}\\ \frac{\partial\mathcal{L}}{\partial\tau_2} = 0, & \text{(A15c)}\\ \lambda_1(\tau_2 - T_{\max}) = 0, & \text{(A15d)}\\ \lambda_{2,k}(\frac{C_{13,k}}{\hat{f}_k} - \tau_2 + T^{\mathrm{G}}) = 0, \forall k \in \mathcal{K}, & \text{(A15e)}\\ \lambda_3(\frac{C_{14}}{\tilde{f}} - \tau_2 + \max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\}) = 0, & \text{(A15f)}\\ \lambda_{4,k}(-\hat{f}_k) = 0, \forall k \in \mathcal{K}, & \text{(A15g)}\\ \lambda_{5,k}(\hat{f}_k - \hat{f}_{\max}) = 0, \forall k \in \mathcal{K}, & \text{(A15h)}\\ \lambda_6(-\tilde{f}) = 0, & \text{(A15i)}\\ \lambda_7(\tilde{f} - \tilde{f}_{\max}) = 0, & \text{(A15j)}\\ \lambda_1 \geq 0,\ \lambda_3 \geq 0,\ \lambda_6 \geq 0,\ \lambda_7 \geq 0, \\ \lambda_{2,k} \geq 0,\ \lambda_{4,k} \geq 0,\ \lambda_{5,k} \geq 0, \forall k \in \mathcal{K}. & \text{(A15k)}\end{cases}$$

It is noticed that constraint (47c) can be rewritten as

$$\hat{f}_k \geq \frac{C_{13,k}}{\tau_2 - T^{\mathrm{G}}}, \forall k \in \mathcal{K}. \tag{A16}$$

To minimize the objective (47a), $\hat{f}_k$ should be minimized within the feasible region. Moreover, the right-hand side of inequality (A16) monotonously decreases as τ_2 increases. Since constraint (47a), i.e., $\tau_2 \leq T_{\max}$, should be satisfied, the right-hand side of inequality (A16) obtains its minimum when $\tau_2 = T_{\max}$. Hence, the optimal τ_2 and $\hat{f}_k$ can given by (A17) and (A18), respectively, i.e., given by

$$\tau_2^* = T_{\max}, \tag{A17}$$

$$\hat{f}_k^* = \frac{C_{13,k}}{T_{\max} - T^{\mathrm{G}}}, \forall k \in \mathcal{K}. \tag{A18}$$

similarly, constraint (47d) can be written as

$$\tilde{f} \geq \frac{C_{14}}{\tau_2 - \max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\}}. \tag{A19}$$

Considering that the objective (47a) increases as $\tilde{f}$ increases, $\tilde{f}$ should also be minimized to minimize the objective (47a). In addition, the right-hand side of (A19) monotonously decreases as τ_2 increases. Intuitively, $\tau_2 = T_{\max}$ should be imposed to minimize the right-hand side of (A19), such that objective (47a) can be minimized as well. Therefore, the optimal $\tilde{f}$ can be given by

$$\tilde{f}^* = \frac{C_{14}}{T_{\max} - \max_{k\in\mathcal{K}}\{T_k^{\mathrm{D}}\}}. \tag{A20}$$

To meet constraints (47c) and (47d), 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 (A17), (A18), and (A20) to (A15a), (A15b), and (A15c), we have

$$\begin{cases} \lambda_1 = \sum_{k=1}^{K} \dfrac{2C_{11,k}(\hat{f}_k^*)^3}{C_{13,k}} + \dfrac{2C_{12}(\tilde{f}^*)^3}{C_{14}}, & \text{(A21a)} \\ \lambda_{2,k} = \dfrac{2C_{11,k}(\hat{f}_k^*)^3}{C_{13,k}}, \forall k \in \mathcal{K}, & \text{(A21b)} \\ \lambda_3 = \dfrac{2C_{12}(\tilde{f}^*)^3}{C_{14}}, & \text{(A21c)} \\ \lambda_{5,k} = 0, \forall k \in \mathcal{K}, & \text{(A21d)} \\ \lambda_7 = 0. & \text{(A21e)} \end{cases}$$

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, 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.