Semi-Federated Learning: An Integrated Framework for Pervasive Intelligence in 6G Networks

Jingheng Zheng, Wanli Ni, Hui Tian, Deniz Gündüz, and Tony Q. S. Quek

***Abstract*—In cellular-based federated learning (FL), the base station (BS) is only used to aggregate parameters, which incurs a waste of computing resources at the BS. In this paper, a novel semi-federated learning (SemiFL) framework is proposed to break this bottleneck, where local devices simultaneously send their gradient updates and training samples to the BS for global model computation. To capture the performance of SemiFL over wireless networks, a closed-form convergence upper bound of SemiFL is derived. Then, a non-convex problem is formulated to improve the convergence behavior of SemiFL, subject to the transmit power, communication latency, and computation distortion. To solve this intractable problem, a two-stage algorithm is proposed by controlling the transmit power and receive beamformers. Numerical experiments validate that the proposed SemiFL framework can effectively improve accuracy and accelerate convergence as compared to conventional FL.**

***Index Terms*—Semi-federated learning, convergence analysis, model aggregation, transceiver design.**

I. Introduction

In future sixth generation (6G) wireless networks, edge artificial intelligence (AI) provides an indispensable potential to enable the concept of connected intelligence [1]. Federated learning (FL), as a representative distributed learning paradigm of edge AI, allows edge devices to collaboratively train a shared model [2]. Although FL can effectively break the bottleneck of transmission cost and address privacy concerns to some extend, it still faces certain challenges. For instance, the idleness of the base station (BS) during the local training of FL inevitably leads to insufficient utilization of its computation resources [3]. This prevents the potential possibility of exploiting the substantial computation capability at the BS for better learning performance of FL. In order to unleash this potential, it is important to orchestrate FL over edge devices in parallel to centralized learning (CL) at the BS by developing a more general and efficient learning framework.

So far, only a few works have taken the effort to investigate a fusion of FL and CL. The authors of [4] consider a semi-supervised learning task, where the BS conducts supervised CL while local devices perform unsupervised FL using unlabeled data, and then the obtained models from CL and FL are combined at each round. However, the impact of wireless channels during the model transmission is ignored in [4]. Considering the computation capabilities of local devices, the authors of [3] propose a hybrid FL and CL framework, in which the BS selects the devices with sufficient computing resources to perform FL, while the rest transmit data to the BS for CL. In [5], the devices upload data to the BS for CL during the local training of FL, and a balance between loss function and resource consumption is achieved by jointly designing the communication and computing strategies. Nevertheless, the orthogonal transmission scheme used in [5] is likely to reduce spectrum utilization, and hence, can not support the massive connectivity requirement in future 6G networks.

In this paper, we propose a hybrid FL and CL framework, called semi-federated learning (SemiFL), in which all the devices simultaneously upload local gradients and training samples to the BS to achieve a better convergence behavior. To improve spectrum efficiency, at the user side, training samples and local gradients are transmitted over the same time-frequency resources. At the BS side, training samples are decoded for CL while the local gradients obtained through FL are aggregated over the air [6]–[8]. The global model is then updated by the gradients computed both locally and in a distributed fashion. For scenarios with less privacy concerns, our propose SemiFL can be well suitable. As far as we know, the theoretical analysis and performance optimization of the proposed hybrid learning performance have not been studied in the literature. Our contributions are summarized as follows:

- We propose the SemiFL framework to combine FL over devices and CL at the BS. To tackle the spectrum scarcity problem, we propose a non-orthogonal transmission scheme to send both the gradient information and the raw data from the devices to the BS over a multiple access channel. Accordingly, we design signal decoding and gradient aggregation methods for SemiFL.
- We derive a convergence upper bound to characterize the impact of wireless factors on the performance of SemiFL. Then, we formulate a non-convex transceiver design problem to minimize the bound under the constraints of transmit power, communication latency, and computation distortion. Finally, we propose a two-stage algorithm to solve this challenging problem by employing variable substitution and successive convex approximation (SCA).

Numerical experiments results validate that: i) the designed algorithm can converge within a limited number of iterations; ii) the proposed SemiFL framework outperforms conventional FL at the same cost of communication overhead.

J. Zheng, W. Ni, and H. Tian are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China (e-mail: zhengjh@bupt.edu.cn; charleswall@bupt.edu.cn; tianhui@bupt.edu.cn). D. Gündüz is with the Department of Electrical and Electronic Engineering, Imperial College London, UK (e-mail: d.gunduz@imperial.ac.uk). T. Q. S. Quek is with Information System Technology and Design Pillar, Singapore University of Technology and Design, Singapore (e-mail: tonyquek@sutd.edu.sg).

This work was supported by the National Key R&D Program of China under Grant 2020YFB1807801.

II. System Model

As depicted in Fig. 1, we consider a system with one N_r-antenna BS and K single-antenna devices. The devices are indexed by $\mathcal{K} = \{1, 2, \ldots, K\}$. We consider T communication rounds, where each round lasts a duration of T_c units of time. Let $\mathcal{D}_k$ denote the dataset of the k-th device containing $N_k = |\mathcal{D}_k|$ training samples. All devices collaboratively train a shared global model by minimizing the global loss function $F(\mathbf{w})$ over the total dataset $\mathcal{D} = \cup_k \mathcal{D}_k$, given by

$$F(\mathbf{w}) = \frac{1}{N} \sum\nolimits_{k=1}^{K} \sum\nolimits_{n \in \mathcal{D}_k} f(\mathbf{w}; \mathbf{x}_{k,n}, \mathbf{y}_{k,n}), \quad (1)$$

where $\mathbf{w} \in \mathbb{R}^Q$ denotes the shared global model, $\mathbf{x}_{k,n}$ and $\mathbf{y}_{k,n}$ are the feature and label vector of a training sample, respectively, $f(\mathbf{w}; \mathbf{x}_{k,n}, \mathbf{y}_{k,n})$ is the sample-wise loss function, and $N = \sum_{k=1}^{K} N_k$ is the total number of training samples.

During the t-th round, the k-th device calculates the local gradient $\mathbf{g}_{t,k}^f \in \mathbb{R}^Q$ with training samples $\mathcal{D}_{f,k}$, given by

$$\mathbf{g}_{t,k}^f = \frac{1}{N_{f,k}} \sum\nolimits_{n \in \mathcal{D}_{f,k}} \mathbf{g}_{t,k,n}, \ \forall k \in \mathcal{K}, \quad (2)$$

where $\mathbf{g}_{t,k,n}$ is the gradient corresponding to one training sample, and $N_{f,k} = |\mathcal{D}_{f,k}|$. Considering scenarios with less privacy concerns, the k-th device also transmits $N_{c,k}$ untrained samples, denoted by $\mathcal{D}_{c,k}$, to the BS for CL. Note that $N_{f,k} + N_{c,k} \ll N_k$ due to the limited computation and communication resources. To improve the spectrum efficiency, the local gradients and training samples are transmitted in the same time-frequency resources, and the pre-processing of the transmit signal is two-fold.

On the one hand, the local gradient $\mathbf{g}_{t,k}^f$ needs to be normalized. The normalization procedure is summarized as follows. First, the k-th device transmits $\frac{1}{Q}\sum_{q=1}^{Q} g_{t,k,q}^f$ and $\frac{1}{Q}\sum_{q=1}^{Q} (g_{t,k,q}^f)^2$ to the BS, where $g_{t,k,q}^f$ is the q-th entry of local gradient $\mathbf{g}_{t,k}^f$. Then, the BS calculates the global mean $\bar{g}_t = \frac{1}{K}\sum_{k=1}^{K} (\frac{1}{Q}\sum_{q=1}^{Q} g_{t,k,q}^f)$ and variance $\sigma_t^2 = \frac{1}{K}\sum_{k=1}^{K}[\frac{1}{Q}\sum_{q=1}^{Q} (g_{t,k,q}^f)^2] - \bar{g}_t^2$. Next, the BS broadcasts $\bar{g}_t$ and σ_t^2 to the devices. The local gradients are normalized by

$$\tilde{g}_{t,k,q}^f = \frac{g_{t,k,q}^f - \bar{g}_t}{\sigma_t}, \ q = 1, 2, \ldots, Q, \ \forall k \in \mathcal{K}, \quad (3)$$

where $\tilde{g}_{t,k,q}^f$ is the q-th entry of the normalized gradient $\tilde{\mathbf{g}}_{t,k}^f$, whose entries have zero mean and unit variance. The k-th device constructs the signal vector for the local gradient as $\mathbf{s}_{t,k} = \frac{N_{f,k}}{N_f} \tilde{\mathbf{g}}_{t,k}^f$, where $N_f = \sum_{k=1}^{K} N_{f,k}$.

On the other hand, each training sample in $\mathcal{D}_{c,k}$ has m bits. The k-th device modulate and normalize the $N_{c,k}$ training samples to a signal vector $\mathbf{d}_{t,k} \in \mathbb{R}^Q$ by appropriate modulation schemes, such as the adaptive quadrature amplitude modulation [9]. An arbitrary entry of $\mathbf{d}_{t,k}$ has zero mean and unit variance. We further assume that the entries of $\mathbf{s}_{t,k}$ and $\mathbf{d}_{t,k}$ are independent of each other, i.e., $\mathbb{E}[s_{t,k,q} d_{t,k,q}] = 0, q = 1, 2, \ldots, Q, \forall k \in \mathcal{K}$. Each communication round is equally split into Q slots. In the q-th slot, the devices transmit $s_{t,k,q}$ and $d_{t,k,q}$ to the BS using the same time-frequency resources.

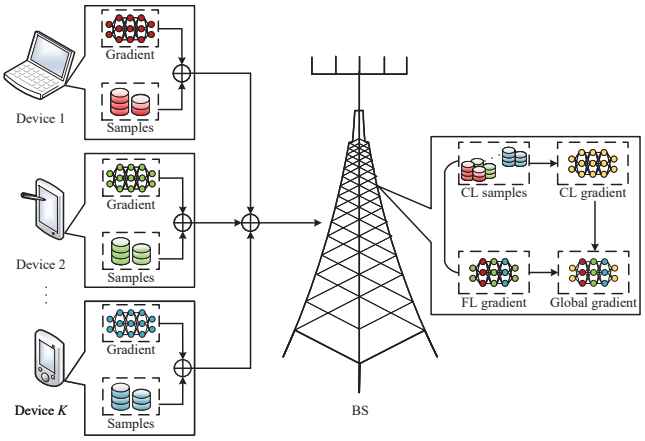

Fig. 1. A demonstration of the proposed SemiFL system.

At the BS side, the received superposition signal in the q-th slot of the t-th round is

$$\mathbf{y}_{t,q} = \underbrace{\sum\nolimits_{k=1}^{K} p_{t,f,k} \mathbf{h}_{t,k} s_{t,k,q}}_{\text{local gradients}} + \underbrace{\sum\nolimits_{k=1}^{K} p_{t,c,k} \mathbf{h}_{t,k} d_{t,k,q}}_{\text{training samples}} + \underbrace{\mathbf{n}_{t,q}}_{\text{noise}}, \quad (4)$$

where $p_{t,f,k} \in \mathbb{C}$ and $p_{t,c,k} \in \mathbb{C}$ denote the transmit power allocation of the local gradients and the training samples, respectively, while $\mathbf{n}_{t,q} \in \mathbb{C}^{N_r}$ is the additive white Gaussian noise with distribution $\mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_r})$, and $\mathbf{h}_{t,k} \in \mathbb{C}^{N_r}$ denotes the Rician channel gain from the k-th device to the BS. We consider a block-fading channel, where $\mathbf{h}_{t,k}$ remains unchanged in a round, but varies independently over rounds. We assume the devices have perfect channel state information.

The post-processing of the received superposition signal is divided into two steps. First, the BS decodes the uploaded training samples from the superposition signal, and utilizes them to calculate the CL gradient. The BS dedicates each device a beamformer, $\{\mathbf{f}_{t,k}\}$, to decode the training samples in parallel. As a result, for the k-th device, the signal for decoding in the q-th slot of the t-th round is given by

$$\hat{d}_{t,k,q} = p_{t,c,k} \mathbf{f}_{t,k}^{\mathrm{H}} \mathbf{h}_{t,k} d_{t,k,q} + \underbrace{\sum\nolimits_{k'=1, k' \neq k}^{K} p_{t,c,k'} \mathbf{f}_{t,k}^{\mathrm{H}} \mathbf{h}_{t,k'} d_{t,k',q}}_{\text{interference of other training samples}}$$
$$+ \underbrace{\sum\nolimits_{k'=1}^{K} p_{t,f,k'} \mathbf{f}_{t,k}^{\mathrm{H}} \mathbf{h}_{t,k'} s_{t,k',q}}_{\text{interference of local gradients}} + \underbrace{\mathbf{f}_{t,k}^{\mathrm{H}} \mathbf{n}_{t,q}}_{\text{noise}}, \ \forall k \in \mathcal{K}. \quad (5)$$

The signal-to-interference-plus-noise ratio (SINR) of the k-th device, $\gamma_{t,k}$, is given by (6) at the top of the next page, where $\|\cdot\|$ denotes the vector 2-norm. The decoded training samples are accumulated over Q slots to reconstruct datasets $\{\mathcal{D}_{c,k}\}$, and utilized to calculate the gradient of CL, $\mathbf{g}_t^c \in \mathbb{R}^Q$. The BS calculates a full-batch gradient based on the entire datasets $\{\mathcal{D}_{c,k}\}$, which is given by

$$\mathbf{g}_t^c = \frac{1}{N_c} \sum\nolimits_{k=1}^{K} \sum\nolimits_{n \in \mathcal{D}_{c,k}} \mathbf{g}_{t,k,n}, \quad (7)$$

where $N_c = \sum_{k=1}^{K} N_{c,k}$ is the number of CL training samples. Then, by removing the signal of the training samples, the BS

$$\gamma_{t,k}=\frac{|p_{t,c,k}\mathbf{f}_{t,k}^{\mathrm{H}}\mathbf{h}_{t,k}|^2}{\sum_{k'=1,k'\neq k}^{K}|p_{t,c,k'}\mathbf{f}_{t,k}^{\mathrm{H}}\mathbf{h}_{t,k'}|^2+\sum_{k'=1}^{K}|p_{t,f,k'}\mathbf{f}_{t,k}^{\mathrm{H}}\mathbf{h}_{t,k'}|^2+\sigma^2\|\mathbf{f}_{t,k}\|^2},\ \forall k\in\mathcal{K}. \tag{6}$$

employs another beamformer $\mathbf{b}_t$ to aggregate local gradients over the air, which is given by

$$\hat{s}_{t,q}=\underbrace{\sum_{k=1}^{K}p_{t,f,k}\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k}s_{t,k,q}}_{\text{aggregated gradient}}+\underbrace{\mathbf{b}_t^{\mathrm{H}}\mathbf{n}_{t,q}}_{\text{noise}}. \tag{8}$$

We measure the signal distortion by the mean square error (MSE) between $s_{t,q}=\sum_{k=1}^{K}s_{t,k,q}$ and $\hat{s}_{t,q}$, i.e.,

$$\mathrm{MSE}_t=\sum_{k=1}^{K}\left|p_{t,f,k}\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k}-1\right|^2+\|\mathbf{b}_t\|^2\sigma^2, \tag{9}$$

After aggregation, the BS de-normalizes $\hat{s}_{t,q}$ to estimate the q-th entry of the aggregated gradient as

$$\hat{g}_{t,q}^{f}=\sigma_t\hat{s}_{t,q}+\bar{g}_t,\ q=1,\ldots,Q. \tag{10}$$

The aggregated gradient at the t-th round is reconstructed by $\hat{\mathbf{g}}_t^f=[\hat{g}_{t,1}^f,\cdots,\hat{g}_{t,Q}^f]^{\mathrm{T}}$.

Finally, the global model for the next round, $\mathbf{w}_{t+1}$, is updated as follows:

$$\mathbf{w}_{t+1}=\mathbf{w}_t-\eta\hat{\mathbf{g}}_t, \tag{11}$$

where η is the learning rate, and the global gradient $\hat{\mathbf{g}}_t$ is calculated by

$$\hat{\mathbf{g}}_t=\frac{N_f}{N_f+N_c}\hat{\mathbf{g}}_t^f+\frac{N_c}{N_f+N_c}\mathbf{g}_t^c. \tag{12}$$

III. Convergence and Problem formulation

A. Convergence Analysis

To investigate the convergence of SemiFL, we introduce some assumptions [10]–[12].

Assumption 1 (μ-strongly convex)**.** *The global loss function $F(\mathbf{w})$ is μ-strongly convex. For any $\mathbf{w}$, $\mathbf{w}'\in\mathbb{R}^Q$ and $\mu>0$, we have*

$$F(\mathbf{w})\geq F(\mathbf{w}')+(\mathbf{w}-\mathbf{w}')^{\mathrm{T}}\nabla F(\mathbf{w}')+\frac{\mu}{2}\|\mathbf{w}-\mathbf{w}'\|^2, \tag{13}$$

where $\nabla F(\mathbf{w})$ is the gradient of $F(\mathbf{w})$ regrading $\mathbf{w}$.

Assumption 2 (L-smooth)**.** *The global loss function $F(\mathbf{w})$ is L-smooth. For any $\mathbf{w}$, $\mathbf{w}'\in\mathbb{R}^Q$ and $L>0$, we have*

$$F(\mathbf{w})\leq F(\mathbf{w}')+(\mathbf{w}-\mathbf{w}')^{\mathrm{T}}\nabla F(\mathbf{w}')+\frac{L}{2}\|\mathbf{w}-\mathbf{w}'\|^2. \tag{14}$$

Assumption 3 (Bounded gradients)**.** *The expected squared 2-norm of any local gradient and the the gradient of any training sample are bounded at $\{\mathbf{w}_t\}$. For constants $\xi_1\geq 0$, $\xi_2>0$ and $G^2\geq 0$, we have*

$$\mathbb{E}\left[\|\mathbf{g}_{t,k}^f\|^2\right]\leq G^2,\ \forall k\in\mathcal{K}, \tag{15}$$

$$\|\mathbf{g}_{t,k,n}\|^2\leq\xi_1+\xi_2\|\nabla F(\mathbf{w}_t)\|^2,\ \forall k\in\mathcal{K},\ \forall n\in\mathcal{D}. \tag{16}$$

Based on the above assumptions, we derive the convergence upper bound of SemiFL in the following theorem.

Theorem 1 (Convergence upper bound of SemiFL)**.** *Given Assumptions 1-4 and learning rate $\eta=\frac{1}{L}$, while letting $\mathbf{w}^*$ denote the optimal model, the convergence upper bound of SemiFL after T rounds is given by:*

$$\mathbb{E}[F(\mathbf{w}_{T+1})-F(\mathbf{w}^*)]\leq\rho_1^T\mathbb{E}\left[F(\mathbf{w}_1)-F(\mathbf{w}^*)\right]+\rho_2\frac{1-\rho_1^T}{1-\rho_1}+\sum_{t=1}^{T}\rho_1^{T-t}\varphi_t\left(\{p_{f,k}\},\mathbf{b}\right), \tag{17}$$

where

$$\varphi_t(\{p_{f,k}\},\mathbf{b})=\frac{G^2(4K\sum_{k=1}^{K}N_{f,k}^2\left|1-p_{t,f,k}\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k}\right|^2+N_f^2\sigma^2\|\mathbf{b}_t\|^2)}{L(N_f+N_c)^2}, \tag{18}$$

$$\rho_1=1-\frac{\mu}{L}+4\mu\xi_2\frac{N_f(N-N_f)+N_c(N-N_c)}{L(N_f+N_c)^2}, \tag{19}$$

$$\rho_2=2\xi_1\frac{N_f(N-N_f)+N_c(N-N_c)}{L(N_f+N_c)^2}. \tag{20}$$

Proof: Please refer to Appendix A. □

Based on Theorem 1, one can find that the convergence upper bound of SemiFL is affected by some wireless-related factors, such as the transmit power $\{p_{t,f,k}\}$ and receive beamformer $\mathbf{b}_t$. In order to accelerate the convergence rate of SemiFL, we aim to minimize the convergence upper bound by conducting meticulous transceiver design, while guaranteeing the communication latency requirements of training samples.

B. Problem Formulation

To minimize the optimality gap between the actual loss and the optimal loss, we formulate a transceiver design problem to improve the convergence behavior of SemiFL. Though the convergence upper bound of SemiFL is affected by the wireless-related factors over the T rounds, the factors are independent of each other between the rounds. Therefore, it is equivalent to solving T independent simple-round problems. For any arbitrary round, the optimization problem is formulated as follows, where subscript t is omitted for brevity.

$$\min_{\substack{\{p_{f,k}\},\{p_{c,k}\},\\ \mathbf{b},\{\mathbf{f}_k\}}}\quad\varphi\left(\{p_{f,k}\},\mathbf{b}\right) \tag{21a}$$

$$\text{s.t.}\quad|p_{f,k}|^2+|p_{c,k}|^2\leq P_{\max},\ \forall k\in\mathcal{K}, \tag{21b}$$

$$\frac{mN_{c,k}}{Wb_1\log_2(1+\frac{\gamma_k}{b_2})}\leq T_c,\ \forall k\in\mathcal{K}, \tag{21c}$$

$$\mathrm{MSE}\leq\epsilon, \tag{21d}$$

where $P_{\max}$ denotes the maximum transmit power at the devices, W denotes the transmission bandwidth, $0<b_1<1$ and $b_2>1$ characterize channel capacity loss, and ϵ denotes the maximum tolerable MSE. Constraint (21c) guarantees the communication latency of training samples. Problem (21) is non-convex due to the non-convexity of the objective function and constraints (21c) and (21d). In the following, we propose an efficient two-stage algorithm to solve problem (21).

IV. Proposed Algorithm

A. Receive Beamformer for Aggregation

Given $\{p_{f,k}\}$, $\{p_{c,k}\}$ and $\{\mathbf{f}_k\}$, the subproblem regarding $\mathbf{b}$ is reduced to

$$\min_{\mathbf{b}} \quad \mathbf{b}^{\mathrm{H}}\mathbf{A}_0\mathbf{b}-2\mathrm{Re}\left\{\mathbf{b}^{\mathrm{H}}\sum\nolimits_{k=1}^{K}\frac{4KN_{f,k}^2p_{f,k}}{(N_f+N_c)^2}\mathbf{h}_k\right\} \tag{22a}$$

$$\text{s.t.} \quad \mathbf{b}^{\mathrm{H}}\mathbf{A}_1\mathbf{b}-2\mathrm{Re}\left\{\mathbf{b}^{\mathrm{H}}\sum\nolimits_{k=1}^{K}p_{f,k}\mathbf{h}_k\right\}+K-\epsilon\leq 0, \tag{22b}$$

where $\mathbf{A}_0$ and $\mathbf{A}_1$ are given by

$$\mathbf{A}_0=\sum\nolimits_{k=1}^{K}\frac{4KN_{f,k}^2|p_{f,k}|^2}{(N_f+N_c)^2}\mathbf{h}_k\mathbf{h}_k^{\mathrm{H}}+\frac{N_f^2\sigma^2}{(N_f+N_c)^2}\mathbf{I}_{N_r}, \tag{23}$$

$$\mathbf{A}_1=\sum\nolimits_{k=1}^{K}|p_{f,k}|^2\mathbf{h}_k\mathbf{h}_k^{\mathrm{H}}+\sigma^2\mathbf{I}_{N_r}. \tag{24}$$

Since problem (22) is convex with respect to (w.r.t.) $\mathbf{b}$, it can be solved by standard toolbox like CVX [13].

B. Transmit Power Allocation

Given $\mathbf{b}$ and $\{\mathbf{f}_k\}$, the subproblem w.r.t. $\{p_{f,k}\}$ and $\{p_{c,k}\}$ is rewritten as

$$\min_{\substack{\{p_{f,k}\},\\ \{p_{c,k}\}}} \quad 4K\sum\nolimits_{k=1}^{K}\frac{N_{f,k}^2}{(N_f+N_c)^2}\left|1-p_{f,k}\mathbf{b}^{\mathrm{H}}\mathbf{h}_k\right|^2 \tag{25}$$

$$\text{s.t.} \quad \text{(21b)}-\text{(21d)},$$

which is non-convex because of the indefinite Hessian matrices in constraint (21c).

Since problem (25) is independent of the angles of $\{p_{c,k}\}$, and $p_{f,k}\mathbf{b}^{\mathrm{H}}\mathbf{h}_k \leq |p_{f,k}||\mathbf{b}^{\mathrm{H}}\mathbf{h}_k|$, we determine the angles of $\{p_{f,k}\}$ and $\{p_{c,k}\}$ by $\angle p_{f,k}=-\angle(\mathbf{b}^{\mathrm{H}}\mathbf{h}_k)$ and $\angle p_{c,k}=0, \forall k\in\mathcal{K}$. Based on the angles, we perform variable substitution by letting $\alpha_k=|p_{f,k}|$ and $\beta_k=|p_{c,k}|^2, \forall k\in\mathcal{K}$. As a result, problem (25) is convexified as

$$\min_{\substack{\{\alpha_k\geq 0\},\\ \{\beta_k\geq 0\}}} \quad 4K\sum\nolimits_{k=1}^{K}\frac{N_{f,k}^2}{(N_f+N_c)^2}\left(1-\alpha_k|\mathbf{b}^{\mathrm{H}}\mathbf{h}_k|\right)^2 \tag{26a}$$

$$\text{s.t.} \quad \alpha_k^2+\beta_k-P_{\max}\leq 0,\ \forall k\in\mathcal{K}, \tag{26b}$$

$$-\beta_k\left|\mathbf{f}_k^{\mathrm{H}}\mathbf{h}_k\right|^2+\gamma_{\min,k}\Big(\sum\nolimits_{k'=1,k'\neq k}^{K}\beta_{k'}|\mathbf{f}_k^{\mathrm{H}}\mathbf{h}_{k'}|^2 +\sigma^2\|\mathbf{f}_k\|^2+\sum\nolimits_{k'=1}^{K}\alpha_{k'}^2|\mathbf{f}_k^{\mathrm{H}}\mathbf{h}_{k'}|^2\Big)\leq 0,\ \forall k\in\mathcal{K}, \tag{26c}$$

$$\sum\nolimits_{k=1}^{K}(1-\alpha_k|\mathbf{b}^{\mathrm{H}}\mathbf{h}_k|)^2+\|\mathbf{b}\|^2\sigma^2-\epsilon\leq 0, \tag{26d}$$

where $\gamma_{\min,k}=b_2(2^{\frac{mN_{c,k}}{b_1WT_c}}-1), \forall k\in\mathcal{K}$. Due to the convexity, problem (26) can be solved by CVX.

C. Receive Beamformers for Decoding

Given $\mathbf{b}$, $\{p_{f,k}\}$ and $\{p_{c,k}\}$, problem (21) is reduced to a feasibility subproblem regarding $\{\mathbf{f}_k\}$. Considering the independence between the devices, the subproblem is decomposed into K feasibility problems. For the k-th device, the subproblem is rewritten as

$$\text{find} \quad \mathbf{f}_k \tag{27a}$$

$$\text{s.t.} \quad \mathbf{f}_k^{\mathrm{H}}\mathbf{A}_{2,k}\mathbf{f}_k\leq 0, \tag{27b}$$

Algorithm 1: The Proposed Two-Stage Algorithm

1: **Initialize** $\{p_{f,k}\}$,$\{p_{c,k}\}$,$\mathbf{b}$,$\{\mathbf{f}_k\}$, maximum iterations N, accuracy ε, $n=0$, and $n'=0$.
2: **repeat**
3: Update $n\leftarrow n+1$.
4: Obtain $\mathbf{b}$ by solving (22).
5: Calculate $\angle p_{f,k}=-\angle(\mathbf{b}^{\mathrm{H}}\mathbf{h}_k),\ \forall k\in\mathcal{K}$ and $\angle p_{c,k}=0,\ \forall k\in\mathcal{K}$.
6: Obtain $\{\alpha_k\}$ and $\{\beta_k\}$ by solving (26).
7: **until** $n\geq N$ or $\frac{|obj^{(n)}-obj^{(n-1)}|}{|obj^{(n)}|}\leq\varepsilon$.
8: **repeat**
9: Update $n'\leftarrow n'+1$.
10: Obtain $\mathbf{f}_k^{(n')}$ and $\nu_k^{(n')}$ by solving (31).
11: **until** $n'\geq N$ or $\frac{|{\nu_k}^{(n')}-{\nu_k}^{(n'-1)}|}{|{\nu_k}^{(n')}|}\leq\varepsilon,\ \forall k\in\mathcal{K}$.
12: **Output** the solution $\{p_{f,k}\}$, $\{p_{c,k}\}$, $\mathbf{b}$, $\{\mathbf{f}_k\}$.

where $\mathbf{A}_{2,k}$ is defined as

$$\mathbf{A}_{2,k}=-|p_{c,k}|^2\mathbf{h}_k\mathbf{h}_k^{\mathrm{H}}+\gamma_{\min,k}\Big(\sum\nolimits_{k'=1,k'\neq k}^{K}|p_{c,k'}|^2\mathbf{h}_{k'}\mathbf{h}_{k'}^{\mathrm{H}} +\sum\nolimits_{k'=1}^{K}|p_{f,k'}|^2\mathbf{h}_{k'}\mathbf{h}_{k'}^{\mathrm{H}}+\sigma^2\mathbf{I}_{N_r}\Big). \tag{28}$$

Similarly to [14], we introduce an auxiliary variable $\nu_k\leq 0$ to transform problem (27) to the following form, which aims to increase the data rate of the k-th device.

$$\min_{\mathbf{f}_k,\nu_k\leq 0} \quad \nu_k \tag{29a}$$

$$\text{s.t.} \quad \mathbf{f}_k^{\mathrm{H}}\mathbf{A}_{2,k}\mathbf{f}_k-\nu_k\leq 0. \tag{29b}$$

However, problem (29) is non-convex due to the indefinite matrix $\mathbf{A}_{2,k}$.

To tackle the non-convexity, we employ SCA to solve it, where the second-order Taylor surrogate function for $\mathbf{f}_k^{\mathrm{H}}\mathbf{A}_{2,k}\mathbf{f}_k$ is constructed as [15]

$$g(\mathbf{f}_k|\mathbf{f}_k^{(n)})=\mathbf{f}_k^{\mathrm{H}}\mathbf{M}_k\mathbf{f}_k+2\mathrm{Re}\{\mathbf{f}_k^{\mathrm{H}}(\mathbf{A}_{2,k}-\mathbf{M}_k)\mathbf{f}_k^{(n)}\} +(\mathbf{f}_k^{(n)})^{\mathrm{H}}(\mathbf{M}_k-\mathbf{A}_{2,k})\mathbf{f}_k^{(n)}. \tag{30}$$

Here, $\mathbf{f}_k^{(n)}$ is the obtained value at the n-th SCA iteration. By placing the first term on the left hand side of (29b) with (30), we can use CVX to solve the following convex problem:

$$\min_{\mathbf{f}_k,\nu_k\leq 0} \quad \nu_k \tag{31a}$$

$$\text{s.t.} \quad g(\mathbf{f}_k|\mathbf{f}_k^{(n)})-\nu_k\leq 0. \tag{31b}$$

The proposed two-stage algorithm for minimizing the convergence upper bound of SemiFL is summarized in Algorithm 1. The fist stage is formed by lines 2-7, and the second stage is formed by lines 8-11. We use CVX, which invokes the standard interior-point method to solve the subproblems. The worst-case complexity for solving problems (22) and (26) are $\mathcal{O}(N_1N_r^3)$ and $\mathcal{O}(8N_2K^3)$, respectively. The worst-case complexity for executing SCA is $\mathcal{O}(N_3N_r^3)$ [16]. Here, N_1, N_2 and N_3 are the maximum iterations of the interior-point method. Therefore, the worst-case complexity of Algorithm 1 is $\mathcal{O}(NN_1N_r^3+8NN_2K^3+KNN_3N_r^3)$.

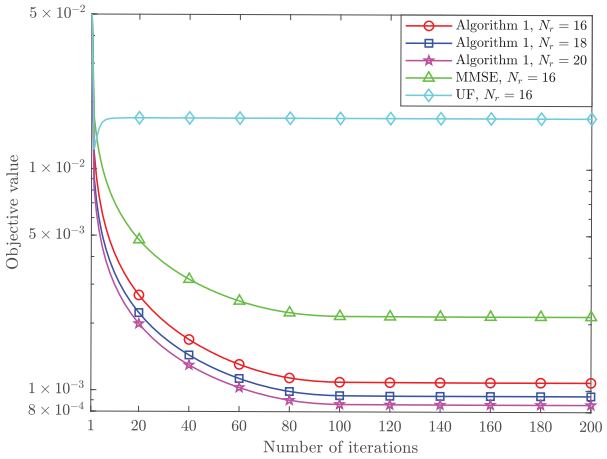

Fig. 2. Convergence behavior of Algorithm 1.

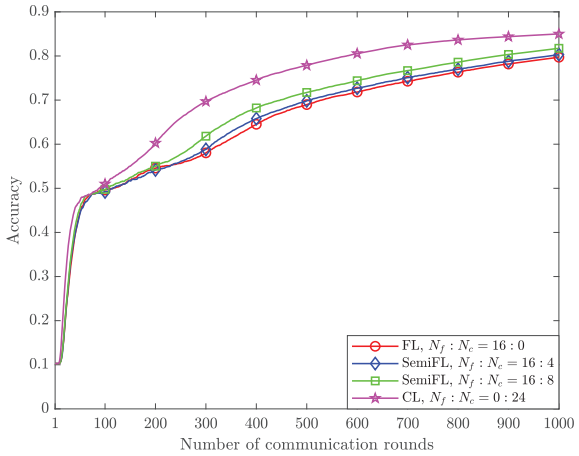

Fig. 3. Accuracy versus number of communication rounds.

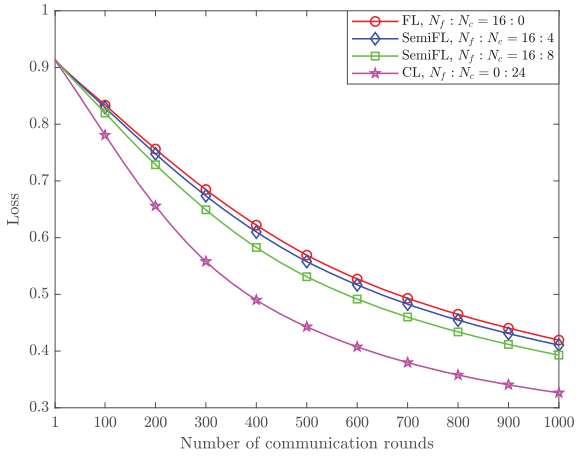

Fig. 4. Loss versus number of communication rounds.

V. Numerical Results

We evaluate our proposed SemiFL framework by letting $K = 10$ devices train a shared multilayer perceptron (MLP) based on the MNIST dataset. The MLP has one hidden layer consisting of 50 neurons, and the loss function is the MSE function. The batch size for FL is $N_{f,k} = 16, \forall k \in \mathcal{K}$, while the number of CL training samples are $N_{c,k} = 8, \forall k \in \mathcal{K}$. We set each entry of a training sample has 16 bits. Since each training sample comprises a 28×28 grey-scale image and a 10-dimension label vector, we set $m = 16 \times (28 \times 28 + 10) = 12704$ bits. The path loss factor is $\tau = 2.2$ with reference path loss $C_0 = -30$ dB, and the Rician factor is $\kappa = 2$ [17]. The transmission bandwidth is $W = 5$ MHz. The noise power is $\sigma^2 = -80$ dBm. The maximum transmit power is $P_{\max} = 0$ dB. Unless extra specification, other simulation parameters are set as: $\eta = 0.01$, $N_r = 16$, $T_c = 500$ ms, $\epsilon = 0.5$, $\varepsilon = 0.01$, $b_1 = 0.905$ and $b_2 = 1.34$ [18].

Fig. 2 plots the convergence behavior of the proposed Algorithm 1. We consider two benchmarks: i) the receive beamformer $\mathbf{b}$ is configured by the minimum MSE (MMSE) criterion; ii) the transmit power allocation $\{p_{f,k}\}$ adopts the uniform-forcing (UF) method [19]. With the same number of receive antennas, we see that Algorithm 1 converges to the lowest value compared with the benchmarks, which confirms the convergence advantage. Moreover, Algorithm 1 converges to a lower objective value when more antennas are deployed at the BS. This can be attributed to the increasing array gain [12].

Fig. 3 and Fig. 4 plot the accuracy and loss with the increasing number of communications rounds, respectively. Conventional FL and CL are considered as the benchmarks. By comparing the red, blue and green curves, we see that SemiFL outperforms FL in both accuracy and loss. Notably, the reason for the different total number of training samples, i.e., $N_f + N_c$, is that we attempt to emphasize the performance gain is brought by the extra training sample for CL. This reveals the advantage of the proposed SemiFL over the conventional FL. However, by utilizing the same total number of training samples, CL is superior to other schemes, which plays the role of a performance upper bound. Although SemiFL achieves moderate learning performance between FL and CL, it would approach CL as more training samples are transmitted to the BS. The observations imply that SemiFL is a more general framework between FL and CL, and integrating CL into FL is a promising method to improve the learning performance when there are less privacy concerns.

VI. Conclusion

This paper proposed an integrated SemiFL framework for pervasive intelligence in 6G networks, where the devices transmit local gradients and training samples to the BS concurrently. The superposed signal was skillfully processed at the BS considering different transmission goals. Then, we derived a closed-form convergence upper bound of SemiFL to reveal the influence of the wireless-related factors. Next, we proposed a two-stage algorithm to solve the formulated non-convex problem efficiently. Numerical experiments confirmed that the proposed SemiFL framework can outperform FL in terms of both accuracy and loss. Compared to CL, the integrated framework can reduce communication overhead at the cost of moderate performance degradation.

APPENDIX A
PROOF OF THEOREM 1

We rewrite $\hat{\mathbf{g}}_t$ as $\hat{\mathbf{g}}_t = \nabla F(\mathbf{w}_t) - \mathbf{e}$, where $\mathbf{e} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_1\mathbf{e}_3$, $\mathbf{e}_1 = \nabla F(\mathbf{w}_t) - \mathbf{g}_t^f$, $\mathbf{e}_2 = \nabla F(\mathbf{w}_t) - \mathbf{g}_t^c$, $\mathbf{e}_3 = \mathbf{g}_t^f - \hat{\mathbf{g}}_t^f$, $\mathbf{g}_t^f = \frac{1}{N_f}\sum_{k=1}^K N_{f,k}\mathbf{g}_{t,k}^f$, $a_1 = \frac{N_f}{N_f+N_c}$ and $a_2 = \frac{N_c}{N_f+N_c}$. By plugging $\mathbf{w} = \mathbf{w}_{t+1}$, $\mathbf{w}' = \mathbf{w}_t$, $\eta = \frac{1}{L}$ and (11) into (14), while taking the expectation of both sides, we have

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_{t+1})] \leq& \mathbb{E}[F(\mathbf{w}_t)] - \frac{1}{2L}\|\nabla F(\mathbf{w}_t)\|^2 + \frac{1}{2L}\|\mathbf{e}\|^2 \\ \overset{(a)}{\leq}& \mathbb{E}[F(\mathbf{w}_t)] - \frac{1}{2L}\|\nabla F(\mathbf{w}_t)\|^2 + \frac{2a_1^2}{L}\mathbb{E}[\|\mathbf{e}_1\|^2] \\ &+ \frac{2a_2^2}{L}\mathbb{E}[\|\mathbf{e}_2\|^2] + \frac{a_1^2}{L}\mathbb{E}[\|\mathbf{e}_3\|^2], \end{aligned} \tag{32}$$

where (a) stems from the Cauchy-Schwarz inequality and triangle inequality.

For $\mathbb{E}[\|\mathbf{e}_1\|^2]$, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{e}_1\|^2] =& \frac{1}{NN_f^2}\mathbb{E}[\|\sum\nolimits_{n\in(\cup_k\mathcal{D}_{f,k})}(N_f - N)\,\mathbf{g}_{t,k,n} \\ &+ \sum\nolimits_{n\in\mathcal{D}/(\cup_k\mathcal{D}_{f,k})} N_f\mathbf{g}_{t,k,n}\|^2] \\ \overset{(a)}{\leq}& \frac{1}{NN_f^2}\mathbb{E}[\sum\nolimits_{n\in(\cup_k\mathcal{D}_{f,k})}(N_f - N)^2\|\mathbf{g}_{t,k,n}\|^2 \\ &+ \sum\nolimits_{n\in\mathcal{D}/(\cup_k\mathcal{D}_{f,k})} N_f^2\|\mathbf{g}_{t,k,n}\|^2] \\ \overset{(b)}{\leq}& \frac{N - N_f}{N_f}(\xi_1 + \xi_2\|\nabla F(\mathbf{w}_t)\|^2), \end{aligned} \tag{33}$$

where (a) holds because of the Cauchy-Schwarz inequality and triangle inequality, and (b) comes from applying (16).

Similarly, we can obtain

$$\mathbb{E}[\|\mathbf{e}_2\|^2] \leq \frac{N - N_c}{N_c}(\xi_1 + \xi_2\|\nabla F(\mathbf{w}_t)\|^2). \tag{34}$$

Based on (10), we derive $\mathbb{E}[\|\mathbf{e}_3\|^2]$ as

$$\begin{aligned}\mathbb{E}[\|\mathbf{e}_3\|^2] \overset{(a)}{=}& \mathbb{E}[\sum_{q=1}^Q|\sum_{k=1}^K\frac{N_{f,k}}{N_f}(1-p_{t,f,k}\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k})(g_{t,k,q}^f-\bar{g}_t)|^2] \\ &+ Q\sigma_t^2\sigma^2\|\mathbf{b}_t\|^2 \\ \overset{(b)}{\leq}& Q\sigma_t^2\sigma^2\|\mathbf{b}_t\|^2 + 2(\sum_{k=1}^K\mathbb{E}[\sum_{q=1}^Q|g_{t,k,q}^f|^2] + KQ|\bar{g}_t|^2) \\ &(\sum_{k=1}^K|\frac{N_{f,k}}{N_f}(1-p_{t,f,k}\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k})|^2), \end{aligned} \tag{35}$$

where (a) takes the expectation w.r.t. noise $\{\mathbf{n}_{t,q}\}$, and (b) holds because of the Cauchy-Schwarz inequality. Moreover, it is obtained that $|\bar{g}_t|^2 \leq \frac{1}{KQ}\sum_{k=1}^K\|\mathbf{g}_{t,k}^f\|^2$ and $\sigma_t^2 \leq \frac{1}{KQ}\sum_{k=1}^K\|\mathbf{g}_{t,k}^f\|^2$. By plugging $\sum_{q=1}^Q|g_{t,k,q}^f|^2 = \|\mathbf{g}_{t,k}^f\|^2$, the obtained two inequalities, and (15) into (35), we have

$$\mathbb{E}[\|\mathbf{e}_3\|^2] \leq \frac{4KG^2}{N_f^2}\sum_{k=1}^K N_{f,k}^2|1-p_{t,f,k}\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k}|^2 + G^2\sigma^2\|\mathbf{b}_t\|^2. \tag{36}$$

According to [10], we have Polyak-Lojasiewicz inequality as $\|\nabla F(\mathbf{w}_t)\|^2 \geq 2\mu(F(\mathbf{w}_t) - F(\mathbf{w}^*))$. By plugging (33), (34), (36) and Polyak-Lojasiewicz inequality into (32), while subtracting $F(\mathbf{w}^*)$ from both sides, we have

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_{t+1}) - F(\mathbf{w}^*)\right] \leq& \rho_1\mathbb{E}\left[F(\mathbf{w}_t) - F(\mathbf{w}^*)\right] + \rho_2 \\ &+ \varphi_t\left(\{p_{f,k}\}, \mathbf{b}\right). \end{aligned} \tag{37}$$

Recursively applying (37) for T times, we finally reach (17). The proof is complete.

REFERENCES

[1] K. B. Letaief, Y. Shi, J. Lu *et al.*, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2021.

[2] W. Y. B. Lim, N. C. Luong, D. T. Hoang *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surv. Tutorials*, vol. 22, no. 3, pp. 2031–2063, Apr. 2020.

[3] A. M. Elbir and S. Coleri, "A family of hybrid federated and centralized learning architectures in machine learning," May 2021. [Online]. Available: https://arxiv.org/abs/2105.03288v1

[4] E. Diao, J. Ding, and V. Tarokh, "SemiFL: Communication efficient semi-supervised federated learning with unlabeled clients," Oct. 2021. [Online]. Available: https://arxiv.org/abs/2106.01432

[5] W. Hong, X. Luo, Z. Zhao *et al.*, "Optimal design of hybrid federated and centralized learning in the mobile edge computing systems," in *Proc. 2021 IEEE ICC Workshops*, Montreal, QC, Canada, Jun. 2021.

[6] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

[7] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, Mar. 2020.

[8] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[9] A. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[10] M. Chen, Z. Yang, W. Saad *et al.*, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[11] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7595–7609, Nov. 2021.

[12] M. M. Amiri, T. M. Duman, D. Gündüz *et al.*, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5129–5143, Aug. 2021.

[13] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014. [Online]. Available: http://cvxr.com/cvx

[14] Y. Liu, J. Zhao, M. Li *et al.*, "Intelligent reflecting surface aided MISO uplink communication network: Feasibility and power minimization for perfect and imperfect CSI," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1975–1989, Mar. 2021.

[15] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[16] W. Ni, X. Liu, Y. Liu *et al.*, "Resource allocation for multi-cell IRS-aided NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4253–4268, Jul. 2021.

[17] W. Ni, Y. Liu, Z. Yang *et al.*, "Federated learning in multi-RIS aided systems," *IEEE Internet of Things J.*, 2021, early access, doi: 10.1109/JIOT.2021.3130444.

[18] H. Lu, X. Jiang, and C. W. Chen, "Distortion-aware cross-layer power allocation for video transmission over multi-user NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1076–1092, Feb. 2021.

[19] L. Chen, X. Qin, and G. Wei, "A uniform-forcing transceiver design for over-the-air function computation," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 942–945, Dec. 2018.