Knowledge-Guided Learning for Transceiver Design in Over-the-Air Federated Learning

Yinan Zou, Zixin Wang, Xu Chen, Haibo Zhou, and Yong Zhou

Abstract

In this paper, we consider communication-efficient over-the-air federated learning (FL), where multiple edge devices with non-independent and identically distributed datasets perform multiple local iterations in each communication round and then concurrently transmit their updated gradients to an edge server over the same radio channel for global model aggregation using over-the-air computation (AirComp). We derive the upper bound of the time-average norm of the gradients to characterize the convergence of AirComp-assisted FL, which reveals the impact of the model aggregation errors accumulated over all communication rounds on convergence. Based on the convergence analysis, we formulate an optimization problem to minimize the upper bound to enhance the learning performance, followed by proposing an alternating optimization algorithm to facilitate the optimal transceiver design for AirCompassisted FL. As the alternating optimization algorithm suffers from high computation complexity, we further develop a knowledge-guided learning algorithm that exploits the structure of the analytic expression of the optimal transmit power to achieve computation-efficient transceiver design. Simulation results demonstrate that the proposed knowledge-guided learning algorithm achieves a comparable performance as the alternating optimization algorithm, but with a much lower computation complexity. Moreover, both proposed algorithms outperform the baseline methods in terms of convergence speed and test accuracy.

Index Terms

Federated learning, over-the-air computation, knowledge-guided learning.

I. INTRODUCTION

With the ever increasing volume of distributed data, computing power of edge devices, and concerns of data privacy, federated learning (FL) [1]–[3] has recently been recognized as a

Y. Zou, Z. Wang, and Y. Zhou are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (E-mail: {zouyn, wangzx2, zhouyong}@shanghaitech.edu.cn). X. Chen is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China (e-mail: chenxu35@mail.sysu.edu.cn). H. Zhou is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: haibozhou@nju.edu.cn).

promising distributed machine learning (ML) paradigm for edge artificial intelligence (AI) [4]– [6]. FL exploits the geographically dispersed data and computing power to distill intelligence at the network edge by employing an edge server to coordinate multiple edge devices to collaboratively train a shared ML model in an iterative manner. By executing the local training based on the local dataset and the up-to-date global model, each edge device only shares its model information instead of raw data with the edge server to alleviate the privacy leakage concerns. FL is expected to support various intelligent applications [7], including smart healthcare, industrial Internet of Things (IoT), and autonomous vehicles.

FL over wireless networks has recently attracted considerable attention, where the communicationexpensive model/gradient exchange between the edge server and edge devices is a critical issue that needs to be addressed. Because of limited radio spectrum resource and finite computing power of edge devices, it is crucial to study the communication and computation co-design. For instance, the authors in [8] proposed to jointly optimize the device selection, power control, and bandwidth allocation to minimize the FL training loss. By jointly optimizing the computation and communication resources, the authors in [9] developed an efficient algorithm to enable energyefficient FL over wireless networks. Most existing studies adopt the orthogonal multiple access (OMA) scheme, e.g., frequency division multiple access (FDMA) and time division multiple access (TDMA), to ensure that the model update of each participating edge device is successfully received by the edge server before performing global model aggregation. Such a "communicatethen-compute" strategy may not be spectrum-efficient as the number of required frequency/time resource blocks is proportional to the number of participating edge devices.

Over-the-air computation (AirComp) [10], as an emerging non-orthogonal multiple access technique, has the potential to enable spectrum-efficient wireless model/gradient aggregation. By exploiting the waveform superposition nature, AirComp enables the edge server to receive a target nomographic function (e.g., arithmetic mean, weighted average) of the signals concurrently transmitted by multiple edge devices over the same radio channel. During the model aggregation process of FL, the edge server is only interested in receiving a weighted average of the local model updates from the edge devices, rather than each individual local model update. Such a model aggregation process matches well with the principle of AirComp, based on which the edge server can directly obtain a noisy version of the aggregated model update by allowing multiple edge devices to concurrently transmit their local model updates. Such a "compute-whencommunicate" strategy requires only one resource block regardless of the number of participating

edge devices. The communication efficiency of AirComp-assisted FL has recently been demonstrated by the existing studies [11]–[13] and further enhanced by leveraging intelligent reflecting surface (IRS) [14].

A. Motivation and Contributions

Most existing studies on AirComp-assisted FL [11], [14]–[17] mainly treat each communication round equally important, and optimize the learning performance according to the instantaneous mean squared error (MSE) of the aggregated global model at a typical communication round, which leads to a sub-optimal learning performance [18], [19]. This is because these studies ignore an inherent property of FL, i.e., the training process of FL involves multiple communication rounds and the model aggregation errors across all communication rounds collectively affect the final training performance. On the other hand, the existing studies [11], [14]–[17], [20], [21] mainly adopt optimization-based methods for the transceiver design of AirComp. However, the optimization-based methods typically suffer from high computation complexity and require the global channel state information (CSI), which hinder their practical applications. These two issues motivate us to develop a both communication and computation efficient framework to design, analyze, and optimize AirComp-assisted FL, taking into account the impact of the aggregation errors over all communication rounds on the FL performance.

In this paper, we consider over-the-air FL over a single-cell wireless network, where the edge devices with non-independent and identically distributed (non-i.i.d.) datasets first perform multiple local iterations and then concurrently transmit their gradients to the edge server over the same radio channel using AirComp in each communication round. Under this setup, we aim to characterize the convergence of the proposed communication-efficient AirComp-assisted FL and further develop a learning-based resource allocation algorithm to enhance the transceiver design, taking into account the model aggregation errors accumulated over all communication rounds. AirComp-assisted FL and learning-based transceiver design, as two critical components of our proposed unified framework, can be regarded as communication for AI and AI for communication, respectively. The main contributions of this paper are summarized as follows.

• We theoretically analyze the convergence of the proposed communication-efficient AirCompassisted FL system, taking into account multiple local stochastic gradient descent (SGD) iterations and the non-i.i.d. data at edge devices. The convergence analysis demonstrates that the time-average MSE is a critical factor that captures the model aggregation errors accumulated over all communication rounds and determines the convergence performance of AirComp-assisted FL. To enhance the learning performance, we formulate an optimization problem to minimize the time-average MSE of the aggregated global model, while taking into account the maximum and average transmit power budgets of each edge device.

- To minimize the time-average MSE of the aggregated global model, we propose an alternating optimization algorithm to optimize the transmit power of each edge device and the receive normalizing factor at the edge server. Due to the non-convexity arsing from the coupling between the transmit power of edge devices and the receive normalizing factor, we decouple the optimization variables and transform the non-convex optimization problem into two convex subproblems. We further derive the optimal receive normalizing factor and the optimal transmit power of edge devices by leveraging KKT conditions.
- As the proposed optimization-based algorithm demands relatively high computation complexity and requires global CSI, we further develop a novel knowledge-guided learning algorithm, which constructs a deep neural network (DNN) with domain knowledge to map the instantaneous CSI to the transmit power of edge devices and the receive normalizing factor. By exploiting the structure of the analytical expression of the optimal transmit power, the proposed knowledge-guided learning algorithm reduces the searching space of the transmit power and in turn achieves a lower computation complexity than the conventional optimization-based algorithms. Moreover, as collecting the optimal solutions to the optimization problem as labels is generally time-consuming, we adopt unsupervised learning to train the DNN specifically developed for effective AirComp transceiver design.
- Simulation results demonstrate that the proposed alternating optimization algorithm and knowledge-guided learning algorithm achieve faster convergence rates and better learning performance than the baseline methods, including full power method, channel inversion method, and knowledge-free learning method. Moreover, the proposed knowledge-guided learning algorithm can achieve a comparable learning performance compared to the proposed alternating optimization algorithm, but with a much lower computation complexity.

B. Related Works

1) OMA-based FL: Various studies have recently been proposed to optimize resource allocation for FL over wireless networks [22]–[27]. In particular, the authors in [22] proposed an FL algorithm for the scenario with non-i.i.d. data and developed an efficient resource allocation algorithm to improve the training performance. The authors in [23] proposed to adapt the frequency of global model aggregation to minimize the training loss. With a fixed training time budget, a joint bandwidth allocation and scheduling policy was proposed in [24]. By considering imperfect CSI, the authors in [25] proposed a joint device scheduling and resource allocation algorithm to improve the training performance. Taking into account the CPU-GPU heterogeneous computing, the authors in [26] designed a joint computation and communication resource allocation scheme to enhance the energy-efficiency of FL. In addition to wirelss resource allocation, learning parameters (e.g., batch-size) can be further adjusted to enhance FL. To accelerate the training process, the authors in [27] proposed a co-design of batch-size selection and communication resource allocation that can adapt to time-varying wireless channels. Note that all the aforementioned studies adopted the OMA scheme, which may not be spectrum-efficient for uplink model aggregation, especially when the number of edge devices is large.

2) AirComp-assisted FL: Leveraging AirComp to support wireless FL has recently been studied from different perspectives. In particular, the authors in [11] proposed a joint design of device selection and receive beamforming to improve the learning performance of AirComp-assisted FL. To mitigate the aggregation error induced by AirComp, the authors in [15] developed an efficient transmit power control strategy. To alleviate the communication bottleneck of AirComp, the authors in [14], [16] leveraged IRS to mitigate the magnitude misalignment during model/gradient aggregation. The authors in [20] proposed to exploit receiver noise as a source of randomness to ensure differential privacy. Moreover, the local learning rates can be further optimized to enhance the learning performance based on the channel conditions in [17]. To reduce the implementation complexity, the authors in [28] utilized momentum-based gradient to update the global model. However, most existing studies on AirComp-assisted FL did not take into account the model aggregation error accumulated over all communication rounds, which determines the final learning performance.

3) Deep learning for resource allocation: Due to the recent advancement of deep learning (DL), DNN can be applied to reduce the computation complexity of optimization-based resource allocation algorithms in wireless networks. The authors in [29]–[31] proposed to train DNNs for interference management and sum rate maximization. The authors in [32] utilized the graph neural network (GNN) for IRS configuration, beamformer design, and power control. However, these data-driven methods generally require a large amount of training samples and lack interpretability and predictability [33]. These issues can be tackled by model-driven methods that

construct neural networks based on domain knowledge. The authors in [34] proposed a modeldriven DNN to replace the conventional orthogonal frequency-division multiplexing (OFDM) receiver. For joint activity detection and channel estimation, the authors in [35], [36] proposed to unfold the numerical iterative methods as the recurrent neural network (RNN). Furthermore, the authors in [37] exploited the structure of the optimal solutions to design DNN for fast beamforming design. Leveraging model-driven DL to achieve computation-efficient transceiver design for AirComp-assisted FL is a critical issue, which, however, has not been explored.

C. Organization and Notations

The rest of this paper is organized as follows. In Section II, we describe the system model of AirComp-assisted FL. We present the convergence analysis and problem formulation in Section III. In Section IV, we propose an alternating optimization algorithm. In Section V, we develop a novel knowledge-guided learning algorithm. In Section VI, the simulation results are provided. Finally, the paper is concluded in Section VII.

We use \mathbb{R}^n to denote the real domain of dimension *n*. Italic, boldface lower-case, and boldface upper-case letters are used to denote scalar, vector, and matrix, respectively. We denote $(\cdot)^T$ as the transpose and $(\cdot)^H$ as Hermitian transpose. $\mathbb{E}[\cdot]$ denotes the statistical expectation operator and $\|\cdot\|$ denotes the Euclidean norm.

II. SYSTEM MODEL

A. Federated Learning Model

Consider FL over a single-cell wireless network, where an edge server co-located with a singleantenna base station coordinates K single-antenna edge devices to collaboratively train a shared ML model, as shown in Fig. 1. We denote the index set of edge devices as $\mathcal{K} = \{1, \ldots, K\}$. Each edge device $k \in \mathcal{K}$ owns a local dataset $\mathcal{D}_k = \{(\boldsymbol{x}_{k,i}, y_{k,i}), 1 \leq i \leq |\mathcal{D}_k|\}$, where $\boldsymbol{x}_{k,i}$ and $y_{k,i}$ denote the *i*-th data sample and its associated label at edge device k, respectively, and $|\mathcal{D}_k|$ denotes the cardinality of set \mathcal{D}_k . The local data at the k-th edge device are generated according to the data distribution $\mathcal{T}_k, \forall k$. In practice, the local data at different edge devices are usually non-i.i.d., i.e., $\mathcal{T}_k \neq \mathcal{T}_j, \forall k \neq j \in \mathcal{K}$. The local loss function at edge device k with respect to local model vector $\boldsymbol{w}_k \in \mathbb{R}^N$ of dimension N is defined by the empirical risk over its local data



Fig. 1. Illustration of AirComp-assisted FL.

$$F_k(\boldsymbol{w}_k; \mathcal{D}_k) = \frac{1}{|\mathcal{D}_k|} \sum_{i=1}^{|\mathcal{D}_k|} f(\boldsymbol{w}_k; (\boldsymbol{x}_{k,i}, y_{k,i})), \quad \forall k \in \mathcal{K},$$
(1)

where $f(\boldsymbol{w}_k; (\boldsymbol{x}_{k,i}, y_{k,i}))$ denotes the sample-wise loss function with respect to $(\boldsymbol{x}_{k,i}, y_{k,i})$. For simplicity, we follow [14] and assume that each edge device has the same amount of data samples, i.e., $|\mathcal{D}_1| = \cdots = |\mathcal{D}_K| = \frac{1}{K} |\mathcal{D}_{tot}|$, where $\mathcal{D}_{tot} = \bigcup_{k=1}^K \mathcal{D}_k$ denotes the global dataset. The empirical global loss function with respect to global model vector $\boldsymbol{w} \in \mathbb{R}^N$ over the global dataset, denoted as $F(\boldsymbol{w}; \mathcal{D}_{tot})$, is

$$F(\boldsymbol{w}; \mathcal{D}_{\text{tot}}) = \frac{1}{|\mathcal{D}_{\text{tot}}|} \sum_{k=1}^{K} |\mathcal{D}_k| F_k(\boldsymbol{w}_k; \mathcal{D}_k) = \frac{1}{K} \sum_{k=1}^{K} F_k(\boldsymbol{w}_k; \mathcal{D}_k).$$
(2)

The objective of the training procedure is to find the optimal weight vector w^* that minimizes the global loss function $F(w; D_{tot})$, i.e.,

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^N} F(\boldsymbol{w}; \mathcal{D}_{\text{tot}}).$$
(3)

B. Over-the-Air Federated Learning

To achieve communication-efficient FL, we adopt the over-the-air FedAvg algorithm, where all edge devices first execute multiple local iterations for local gradient computation and then concurrently transmit their accumulated local gradients to the edge server using AirComp for global gradient aggregation. The computation and communication processes are elaborated as follows. At the beginning of communication round t, the edge server broadcasts global model w(t) to all edge devices in the downlink. As the edge server generally has a much higher transmit power than the edge devices, we assume that each edge device can receive global model w(t)with negligible distortion [20]. After receiving global model vector w(t), each edge device k initializes its local model by setting $w_k(t, 0) = w(t)$, and then updates its local model for $\phi > 1$ iterations using the local stochastic gradient as follows

$$\boldsymbol{w}_k(t,\zeta+1) = \boldsymbol{w}_k(t,\zeta) - \lambda \tilde{\boldsymbol{g}}_k(t,\zeta), \quad \zeta = 0,\dots,\phi-1,$$
(4)

where λ is the learning rate, $\boldsymbol{w}_k(t,\zeta)$ is the local model at device k in round t after ζ local iterations, and

$$\tilde{\boldsymbol{g}}_{k}(t,\zeta) = \nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta); \boldsymbol{\mathcal{B}}_{k}(t,\zeta)) = \frac{1}{B} \sum_{(\boldsymbol{x}_{k},y_{k})\in\boldsymbol{\mathcal{B}}_{k}(t,\zeta)} \nabla f(\boldsymbol{w}_{k}(t,\zeta); (\boldsymbol{x}_{k},y_{k}))$$
(5)

is the stochastic gradient evaluated using the mini-batch $\mathcal{B}_k(t,\zeta)$ that contains B randomly sampled data samples from the local dataset \mathcal{D}_k .

To update the global model, the edge server needs to obtain the aggregation of the accumulated local gradients. Although the OMA scheme can be adopted for gradient uploading in the uplink, e.g., TDMA, the number of required resource blocks scales linearly with the number of edge devices. When there are a large number of edge devices participating in FL training, the incurred communication latency may be very large and becomes the main performance-limiting factor. To this end, we resort to using the AirComp technique to reduce the communication latency and thus enable fast gradient aggregation for FL. Specifically, during the uplink transmission process, all edge devices concurrently transmit their accumulated local gradients to the edge server with appropriate pre-processing over the same radio channel. By exploiting the waveform superposition property of multiple-access channels, the server is capable of directly receiving an aggregation of all accumulated local gradients. By enabling all edge devices to transmit concurrently, the communication latency introduced by AirComp is independent of the number of edge devices, thereby achieving communication-efficient global gradient aggregation.

To facilitate transmit power control for edge devices, we normalize the N-dimensional accumulated local gradients $\boldsymbol{\theta}_k(t) \in \mathbb{R}^N \triangleq (\boldsymbol{w}_k(t,0) - \boldsymbol{w}_k(t,\phi))/\lambda = \sum_{\zeta=0}^{\phi-1} \tilde{\boldsymbol{g}}_k(t,\zeta)$ before the uplink transmission. In particular, after computing the local model, device k computes the mean $\bar{\theta}_k(t)$ and variance $\pi_k^2(t)$ of $\boldsymbol{\theta}_k(t)$ as follows

$$\bar{\theta}_k(t) = \frac{1}{N} \sum_{j=1}^N \theta_{k,j}(t), \ \forall k \in \mathcal{K},$$
(6)

$$\pi_k^2(t) = \frac{1}{N} \sum_{j=1}^N \left(\theta_{k,j}(t) - \bar{\theta}_k(t) \right)^2, \ \forall k \in \mathcal{K},$$
(7)

where $\theta_{k,j}(t)$ denotes the *j*-th element of $\theta_k(t)$. By denoting $\bar{\theta}(t) = \frac{1}{K} \sum_{k \in \mathcal{K}} \bar{\theta}_k(t)$ and $\pi^2(t) = \frac{1}{K} \sum_{k \in \mathcal{K}} \pi_k^2(t)$, device *k* normalizes $\theta_k(t)$ as

$$\mathbf{s}_k(t) = \frac{\boldsymbol{\theta}_k(t) - \boldsymbol{\theta}(t)}{\pi(t)}, \ \forall k \in \mathcal{K},$$
(8)

and transmits $s_k(t)$ to the edge server over wireless fading channels. Note that $s_k(t)$ has zero mean and unit variance, i.e., $\mathbb{E}[s_k(t)s_k(t)^T] = I_N$.

We consider block-fading channels, i.e., the channel coefficients remain invariant within

each communication round but vary independently from one round to another. In the *t*-th communication round, we denote the complex-valued channel coefficient between edge device k and the edge server as $h_k(t)$, which is assumed to be known by edge device k, as in [12]–[14]. Before transmission, signal $s_k(t)$ is multiplied by a pre-processing factor $\psi_k(t)$ to compensate for the phase distortion due to channel fading. In particular, we set $\psi_k(t) = \frac{\sqrt{p_k(t)}h_k^{\rm H}(t)}{|h_k(t)|}$, where $p_k(t) \ge 0$ denotes the transmit power of device k in the *t*-th communication round. We assume that all edge devices are synchronized, which can be achieved by either utilizing a reference-clock [38] or adopting the timing advance technique [39]. The received signal of dimension N at the edge server can be expressed as

$$\boldsymbol{y}(t) = \sum_{k=1}^{K} h_k(t) \psi_k(t) \boldsymbol{s}_k(t) + \boldsymbol{n}(t), \qquad (9)$$

where $\mathbf{n}(t) \sim \mathcal{N}(\sigma^2 \mathbf{I}_N)$ denotes the additive white Gaussian noise (AWGN) vector. Upon receiving the signal, we apply the receive normalizing factor $\eta(t)$ at the edge server for signal amplitude alignment and noise suppression. Hence, we have

$$\hat{\boldsymbol{s}}(t) = \frac{\boldsymbol{y}(t)}{\sqrt{\eta(t)}} = \sum_{k=1}^{K} \frac{\sqrt{p_k(t)}|h_k(t)|}{\sqrt{\eta(t)}} \boldsymbol{s}_k(t) + \frac{\boldsymbol{n}(t)}{\sqrt{\eta(t)}}.$$
(10)

Note that $\hat{s}(t)$ is an estimation of the target variable $s(t) = \sum_{k=1}^{K} s_k(t)$. After de-normalization, we obtain

$$\hat{\boldsymbol{\theta}}(t) = \frac{1}{K} \bigg(\pi(t) \hat{\boldsymbol{s}}(t) + K \bar{\boldsymbol{\theta}}(t) \bigg).$$
(11)

Recall that $\boldsymbol{\theta}_k(t) = \pi(t)\boldsymbol{s}_k(t) + \bar{\theta}(t)$ and $\boldsymbol{s}(t) = \sum_{k=1}^{K} \boldsymbol{s}_k(t)$, (11) can be rewritten as

$$\boldsymbol{\theta}(t) = \frac{1}{K} \pi(t) \left(\hat{\boldsymbol{s}}(t) - \boldsymbol{s}(t) \right) + \boldsymbol{\theta}(t), \tag{12}$$

where $\boldsymbol{\theta}(t) = \frac{1}{K} \sum_{k \in \mathcal{K}} \boldsymbol{\theta}_k(t)$. The edge server can only obtain an estimation of $\boldsymbol{\theta}(t)$, i.e., $\hat{\boldsymbol{\theta}}(t)$, to update the global model parameter $\boldsymbol{w}(t+1)$ as follows

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \lambda \hat{\boldsymbol{\theta}}(t) = \boldsymbol{w}(t) - \lambda (\boldsymbol{\theta}(t) + \boldsymbol{e}(t)),$$
(13)

where $e(t) = \frac{1}{K}\pi(t)(\hat{s}(t) - s(t))$ represents the random aggregation error in each communication round. This error is introduced by channel fading and receiver noise, and determines the convergence performance of FL.

The learning process proceeds by performing (4), (10), and (13) iteratively, until the global model is converged or the maximum number of communication rounds is reached. In addition, we consider that each edge device has the following maximum transmit power constraint and average transmit power constraint

$$p_k(t) \le P_k^{\max}, \quad \forall k \in \mathcal{K}, \ \forall t \in \{0, \dots, T-1\},$$
(14)

$$\frac{1}{T}\sum_{t=0}^{T-1} p_k(t) \le \bar{P}_k, \quad \forall k \in \mathcal{K},$$
(15)

where $P_k^{\text{max}} > 0$ and $\bar{P}_k > 0$ denote the maximum and average transmit power budgets of edge device k, respectively, and T is the maximum number of communication rounds [18], [21], [40]. To make the average transmit power constraint non-trival, we assume $\bar{P}_k < P_k^{\text{max}}$.

III. CONVERGENCE ANALYSIS AND PROBLEM FORMULATION

In this section, we present the convergence analysis for AirComp-assisted FL, taking into account multiple local SGD iterations and the non-i.i.d. data, followed by formulating an optimization problem to minimize the upper bound of the time-average norm of the gradients.

A. Preliminary

1) Non-i.i.d. data: With non-i.i.d. datasets among edge devices, the local optimum of the local loss function may not be consistent with the global optimum of the global loss function. As the heterogeneity level of the local gradients reflects that of the local data, we define the following metric.

Definition 1. For K edge devices with local gradients $\{\nabla F_k(\boldsymbol{w}_k)\}\)$, we define metric χ to characterize the heterogeneity level of the local gradients as follows

$$\frac{\frac{1}{K}\sum_{k=1}^{K} \|\nabla F_k(\boldsymbol{w}_k)\|_2^2}{\|\frac{1}{K}\sum_{k=1}^{K} \nabla F_k(\boldsymbol{w}_k)\|_2^2} = \frac{\frac{1}{K}\sum_{k=1}^{K} \|\nabla F_k(\boldsymbol{w}_k)\|_2^2}{\frac{1}{K^2}\sum_{k=1}^{K} \|\nabla F_k(\boldsymbol{w}_k)\|_2^2 + \frac{1}{K^2}\sum_{i\neq j} \langle \nabla F_i(\boldsymbol{w}_i), \nabla F_j(\boldsymbol{w}_j) \rangle} \le \chi.$$
(16)

Remark 1. The inner product between two local gradients indicates the divergence of the directions of these two local gradients. Note that $\chi \ge 1$ due to Jensen's inequality. When the data across edge devices are i.i.d., the local gradients tend to be the same with tremendous data samples, and thus $\chi = 1$. However, with statistically heterogeneous data, the data distributions among edge devices are different, which implies that the local gradients pointing to different directions. Hence, the inner product between the local gradients is small, leading to a large value of χ . In particular, a higher level of non-i.i.d. data incurs a larger value of χ .

2) *Basic assumptions:* To facilitate the convergence analysis, we make the following assumptions on the loss function and gradients.

Assumption 1 (Bounded loss function). For any parameter w, the global loss function is lower bounded, i.e., $F(w) \ge F(w^*) > -\infty$. Assumption 2 (Lipshchitz continuity and smoothness). The local loss function $F_k(w)$ is smooth with a non-negative constant L and continuously differentiable, i.e.,

$$\|\nabla F_k(\boldsymbol{w}) - \nabla F_k(\boldsymbol{w}')\|_2 \le L \|\boldsymbol{w} - \boldsymbol{w}'\|_2, \quad \forall \, \boldsymbol{w}, \boldsymbol{w}'.$$
(17)

Inequality (17) directly leads to the following inequality

$$F_k(\boldsymbol{w}') \le F_k(\boldsymbol{w}) + \langle \nabla F_k(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w} \rangle + \frac{L}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|_2^2, \quad \forall \, \boldsymbol{w}, \boldsymbol{w}'.$$
(18)

Assumption 3 (Bounded stochastic gradient variance). The local mini-batch stochastic gradients $\{\tilde{g}_k\}$ are assumed to be independent and unbiased estimates of the batch gradient $\{\nabla F_k(w_k)\}$ with bounded variance, i.e.,

$$\mathbb{E}[\tilde{\boldsymbol{g}}_k] = \nabla F_k(\boldsymbol{w}_k), \quad \forall k \in \mathcal{K},$$
(19)

$$\operatorname{Var}(\tilde{\boldsymbol{g}}_k) = \mathbb{E}[\|\tilde{\boldsymbol{g}}_k - \nabla F_k(\boldsymbol{w}_k)\|_2^2] \le \xi^2, \quad \forall k \in \mathcal{K},$$
(20)

where $\xi \ge 0$ is a constant introduced to quantify the sampling noise of stochastic gradients. **Assumption 4** (Bounded variance). The variance of N elements of θ_k is upper bounded by a constant $\Gamma \ge 0$, i.e., $\pi_k^2 \le \Gamma$.

Remark 2. While Assumption 1 is necessary for converging to a stationary point [41], Assumption 2 is standard for convergence analysis [20]. Assumption 3 indicates that the stochastic gradient is an unbiased estimate of the batch gradient. Due to non-i.i.d. data across edge devices, the local stochastic gradient is no longer an unbiased estimate of the batch gradient of the global loss function [42]. For Assumption 4, since the elements of θ_k have finite values, it is reasonable to assume that π_k^2 , as the variance of N elements of θ_k , is upper bounded as in [14].

B. Convergence Analysis

Based on the above assumptions, we present the following theorem for the convergence of AirComp-assisted FL with multiple local iterations and non-i.i.d. data.

Theorem 1. With Assumptions 1-4, if the learning rate λ and the number of local iterations ϕ satisfy $\phi^2 L^2 \lambda^2 \chi + 2\phi \lambda L \leq 1$, then the time-average norm of the gradients after T communication rounds is upper bounded by

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} \leq \underbrace{\frac{2(F(\boldsymbol{w}(0) - F(\boldsymbol{w}^{*})))}{\lambda(\phi - 1)T}}_{\text{Initial optimality gap}} + \underbrace{\frac{2}{\phi - 1} \left(\frac{\phi^{2}\lambda^{2}L^{2}}{2} + \frac{\phi\lambda L}{K}\right)\xi^{2}}_{\text{Variance of stochatic gradient}} + \underbrace{\frac{1 + 2\lambda L}{\phi - 1} \frac{N\Gamma(K+1)}{K^{2}} \frac{1}{T} \sum_{t=0}^{T-1} \underbrace{\frac{MSE(t)}{\text{instantaneous MSE}}}_{\text{Interval} \text{Initial optimality gap}}, \quad (21)$$

where

$$MSE(t) = \sum_{k=1}^{K} \left(\frac{\sqrt{p_k(t)} |h_k(t)|}{\sqrt{\eta(t)}} - 1 \right)^2 + \frac{\sigma^2}{\eta(t)}.$$
 (22)

Proof. Please refer to Appendix A.

Remark 3. For non-i.i.d. data, when the level of the local gradient heterogeneity χ has a larger value, we set a smaller learning rate and perform a smaller number of local iterations to ensure that condition $\phi^2 L^2 \lambda^2 \chi + 2\phi \lambda L \leq 1$ is satisfied.

Remark 4. In Theorem 1, we adopt the average norm of the gradients as the convergence indicator, which is widely used in the convergence analysis for non-convex loss [43]. Note that the FL algorithm achieves an ϵ -approximation solution if

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\boldsymbol{w}(t))\|_2^2 \le \epsilon.$$
(23)

We observe that the upper bound (21) is composed of three terms. The first two terms are the initial optimality gap and the variance of stochastic gradient. The last term is the timeaverage MSE resulting from analog gradient transmission. As $T \to \infty$, the initial optimality gap decreases to zero, and the upper bound approaches to the summation of the variance of the stochastic gradient and the time-average MSE. Besides, when the number of edge devices, the number of local iterations, and the learning rate are given, the variance of the stochastic gradient is a constant. Consequently, in order to improve the convergence performance, it is necessary to minimize the time-average MSE given in (21), which incorporates the model aggregation errors accumulated over T communication rounds.

C. Problem Formulation

By omitting the constant terms, we rewrite the time-average MSE as

$$\overline{\text{MSE}} = \sum_{t=0}^{T-1} \left[\sum_{k=1}^{K} \left(\frac{\sqrt{p_k(t)} |h_k(t)|}{\sqrt{\eta(t)}} - 1 \right)^2 + \frac{\sigma^2}{\eta(t)} \right].$$
(24)

We aim to minimize the time-average MSE by jointly optimizing the transmit power $\{p_k(t)\}$ of edge devices and the receive normalizing factors $\{\eta(t)\}$ of the edge server. Hence, the optimization problem can be formulated as

$$\mathscr{P}: \min_{\substack{\{p_k(t)\},\\\{\eta(t)\}}} \quad \overline{\mathsf{MSE}}$$
(25a)

s.t.
$$0 \le p_k(t) \le P_k^{\max}, \quad \forall k, \forall t,$$
 (25b)

$$0 \le \frac{1}{T} \sum_{t=0}^{T-1} p_k(t) \le \bar{P}_k, \quad \forall k,$$
 (25c)

$$\eta(t) \ge 0, \quad \forall t. \tag{25d}$$

The objective function of problem \mathscr{P} contains the noise-induced error (i.e., $\sum_{t=0}^{T-1} [\sigma^2/\eta(t)]$) and the signal misalignment error (i.e., $\sum_{t=0}^{T-1} [\sum_{k=1}^{K} (\sqrt{p_k(t)} |h_k(t)| / \sqrt{\eta(t)} - 1)^2]$). To minimize the time-average MSE, an intuitive idea is to enlarge the receive normalizing factors to diminish the noise-induced error and adjust the transmit power of edge devices to align the signal amplitudes. However, the finite average and maximum transmit power budgets make the signal amplitude alignment not always possible. Hence, it is tricky to simultaneously reduce the signal misalignment error and the noise-induced error. Moreover, problem (25) is a non-convex optimization problem as the transmit power of edge devices and the receive normalizing factor are highly coupled over different communication rounds. All these issues make problem \mathscr{P} challenging to be solved.

IV. ALTERNATING OPTIMIZATION ALGORITHM

In this section, we propose an alternating optimization algorithm to decouple the optimization variables and tackle the non-convex optimization problem \mathcal{P} .

A. Receive Normalizing Factor Optimization

We first optimize the receive normalizing factors $\{\eta(t)\}$ with given transmit power of edge devices $\{p_k(t)\}$ by solving the following problem

$$\mathscr{P}_{1}: \min_{\{\eta(t)\geq 0\}} \sum_{t=0}^{T-1} \left[\sum_{k=1}^{K} \left(\frac{\sqrt{p_{k}(t)}|h_{k}(t)|}{\sqrt{\eta(t)}} - 1 \right)^{2} + \frac{\sigma^{2}}{\eta(t)} \right].$$
(26)

We decompose problem \mathscr{P}_1 into T subproblems for T communication rounds. Each subproblem can be expressed as

$$\min_{\eta(t)\geq 0} \mathcal{E}(\eta(t)) \triangleq \sum_{k=1}^{K} \left(\frac{\sqrt{p_k(t)} |h_k(t)|}{\sqrt{\eta(t)}} - 1 \right)^2 + \frac{\sigma^2}{\eta(t)},\tag{27}$$

where $\mathcal{E}(\eta(t))$ denotes the objective function of problem (27). By denoting $\Omega(t) = 1/\sqrt{\eta(t)}$, we rewrite the objective function of problem (27) as

$$\mathcal{E}(\Omega(t)) = \sum_{k=1}^{K} \left(\sqrt{p_k(t)} |h_k(t)| \Omega(t) - 1 \right)^2 + (\sigma \Omega(t))^2, \tag{28}$$

which is convex with respect to $\Omega(t)$. By setting the first-order derivative of the objective function $\mathcal{E}(\Omega(t))$ to zero, we obtain the closed-form expression of the optimal $\Omega^*(t)$. As a result, the optimal receive normalizing factor $\eta^*(t)$ to problem (27) can be expressed as

$$\eta^*(t) = \frac{1}{(\Omega^*(t))^2} = \left(\frac{\sigma^2 + \sum_{k=1}^K (\sqrt{p_k(t)} |h_k(t)|)^2}{\sum_{k=1}^K \sqrt{p_k(t)} |h_k(t)|}\right)^2.$$
(29)

B. Transmit Power Optimization

We fix the obtained optimal receive normalizing factor and optimize $\{p_k(t)\}$ by solving the following problem

$$\mathscr{P}_{2}: \min_{\{p_{k}(t)\}} \quad \sum_{t=0}^{T-1} \left[\sum_{k=1}^{K} \left(\frac{\sqrt{p_{k}(t)} |h_{k}(t)|}{\sqrt{\eta(t)}} - 1 \right)^{2} \right]$$
(30a)

We decompose problem \mathscr{P}_2 into K subproblems and optimize the transmit power of the k-th edge device by solving the following problem

$$\min_{\{p_k(t)\}} \sum_{t=0}^{T-1} \left(\frac{\sqrt{p_k(t)} |h_k(t)|}{\sqrt{\eta(t)}} - 1 \right)^2$$
(31a)

s.t.
$$0 \le p_k(t) \le P_k^{\max}, \quad \forall t,$$
 (31b)

$$0 \le \frac{1}{T} \sum_{t=0}^{T-1} p_k(t) \le \bar{P}_k.$$
(31c)

Note that problem (31) is a convex problem and satisfies the Slater's condition. Thus, we can leverage the KKT conditions to obtain the optimal solution given in the following theorem.

Theorem 2. The optimal solution to problem (31) is given as follows:

• If condition

where μ_k^*

$$\sum_{t=0}^{T-1} \min\left\{\frac{\eta(t)}{|h_k(t)|^2}, P_k^{\max}\right\} \le T\bar{P}_k$$
(32)

holds, then the optimal transmit power $p_k^*(t)$ is given by

$$p_k^*(t) = \min\left\{\frac{\eta(t)}{|h_k(t)|^2}, P_k^{\max}\right\}.$$
 (33)

In this case, the transmit power either uses up the maximum power budget or has a form of channel inversion.

• Otherwise, the optimal transmit power $p_k^*(t)$ is given by $r_k^*(t) = \min \left\{ \left(-\sqrt{\eta(t)} |h_k(t)| \right)^2 \right\}_{D^{\max}}$

$$p_k^*(t) = \min\left\{\left(\frac{\sqrt{|h_k(t)|^2 + \mu_k^*\eta(t)}}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right), P_k^{\max}\right\},\tag{34}$$
can be found via the one-dimensional bisection search method to ensure that the

average transmit power constraint $\sum_{t=0}^{T-1} p_k^*(t) = T\bar{P}_k$ holds.

Proof. Please refer to Appendix B.

By now, problem \mathscr{P} can be tackled by solving problems \mathscr{P}_1 and \mathscr{P}_2 alternately. The proposed algorithm is summarized in Algorithm 1.

Algorithm 1: Proposed alternating optimization algorithm for problem \mathscr{P}

1: Input: $\{h_k(t)\}_{t=0}^{T-1}$, stopping condition ϵ_0 .

- 2: Initialize: Transmit power $\{p_k(t)\}^0$ and i = 0.
- 3: repeat
- 4: i = i + 1.

5: Given
$$\{p_k(t)\}^{i-1}$$
, update $\{\eta(t)\}^i$ via $\{\eta(t)\}^i = \left(\frac{\sigma^2 + \sum_{k=1}^K (\sqrt{\{p_k(t)\}^{i-1}} |h_k(t)|)^2}{\sum_{k=1}^K \sqrt{\{p_k(t)\}^{i-1}} |h_k(t)|}\right)^2$.
6: Given $\{\eta(t)\}^i$, update $\{p_k(t)\}^i$ via $\{p_k(t)\}^i = \min\left\{\frac{\{\eta(t)\}^i}{|h_k(t)|^2}, P_k^{\max}\right\}$, if (32) holds or $\{p_k(t)\}^i = \min\left\{\left(\frac{\sqrt{\{\eta(t)\}^i} |h_k(t)|}{|h_k(t)|^2 + \mu_k^*\{\eta(t)\}^i}\right)^2, P_k^{\max}\right\}$, otherwise, where μ_k^* can be obtained via bisection search.
7: until $\frac{\overline{\mathsf{MSE}}^{i-1} - \overline{\mathsf{MSE}}^i}{\overline{\mathsf{MSE}}^i} < \epsilon_0$.
8: Output: $\{\eta(t)\}$ and $\{p_k(t)\}$.

Remark 5. Although Algorithm 1 can optimally solve problem \mathscr{P} , it has the following two limitations. First, the computation complexity of Algorithm 1 is relatively high, as the iterative algorithm requires a large number of iterations to compute the transmit power of edge devices and the receive normalizing factor. Besides, one-dimensional bisection search is required in each iteration and hence introduces additional computation complexity. Second, the alternating optimization algorithm requires the CSI of all communication rounds to solve problem \mathscr{P} . However, it may not be practical to know the CSI of all communication rounds in advance, especially in time-varying wireless networks. To address these limitations, we shall propose an efficient knowledge-guided learning algorithm to solve problem \mathscr{P} in the following section.

V. KNOWLEDGE-GUIDED LEARNING ALGORITHM

In this section, we propose an unsupervised learning algorithm and construct a DNN with domain knowledge to map the instantaneous CSI to the transmit power of edge devices and the receive normalizing factor. Instead of directly learning the mapping function, the proposed learning algorithm leverages the structure of the optimal transmit power derived in Theorem 2 to enable low-complexity transceiver design for AirComp-assisted FL.



Fig. 2. Architecture of the proposed knowledge-guided learning algorithm for the transceiver design of AirComp-assisted FL.

A. Knowledge-Guided Learning for AirComp Transceiver Design

We develop a knowledge-guided learning algorithm, which imitates the proposed alternating optimization algorithm in Section IV. In particular, the proposed learning algorithm learns a mapping between the instantaneous CSI of the current communication round and the transmit power of edge devices and the receive normalizing factor by exploiting the structure information based on (34). The proposed neural network consists of multiple fully-connected layers and a structure mapping layer, as shown in Fig. 2. The fully-connected layers are designed to predict the dual variables (i.e., $\{\mu_k\}$) and the receive normalizing factor (i.e., η), while the structure mapping layer after the fully-connected layers transforms the dual variables to the transmit powers of edge devices by exploiting the structure of the optimal transmit power derived in Section IV. Specifically, in the structure mapping layer, the transmit powers $\{p_k\}$ of edge devices are generated according to the following structure

$$p_k = \min\left\{\left(\frac{\sqrt{\eta}|h_k|}{|h_k|^2 + \mu_k \eta}\right)^2, P_k^{\max}\right\}, \quad \forall k \in \mathcal{K}.$$
(35)

The structure mapping layer, which converts the dual variables and the receive normalizing factor to transmit powers of edge devices, is an important component of our proposed learning algorithm. In particular, the proposed learning algorithm, rather than directing estimating the transmit powers of edge devices, first predicts the dual variables that are key features extracted from the transmit power, and then utilizes the structure mapping layer to recover transmit power

of edge devices from the predicted dual variables.

The usage of the structure of the optimal transmit power enables the proposed neural network to efficiently find the optimal transmit power of edge devices and the optimal receive normalizing factor to minimize the signal misalignment error and the noise-induced error. Compared to the traditional fully-connected neural network that directly estimates the transmit power of edge devices, a salient feature of the proposed neural network is that the structure information is explicitly embedded into network architecture. It is generally difficult for the conventional fullyconnected neural network to learn this structure because of the huge searching space of the transmit power. By adopting the structure of the optimal transmit power, the searching space of the optimal transmit power can be reduced to a finite-dimensional subspace. In addition, the optimal transmit power is represented by dual variables μ_k and receive normalizing factor η . Note that μ_k and η are constrained, i.e., $\mu_k \ge 0$ and $\eta \ge 0$. Hence, the finite-dimensional subspace can be further reduced, which in turn reduces the computation complexity of searching for the optimal transmit power. Although the CSI of multiple communication rounds is required to generate the training samples and to train the neural network, once the training process is completed, the proposed neural network only needs the CSI of the current communication round to compute the transmit powers of edge devices and the receive normalizing factor.

B. Deep Neural Network Design

As shown in Fig. 2, the proposed neural network consists of an input layer, D hidden layers, an output layer, and a structure mapping layer, which are indexed from 0 to D + 2.

1) Input Layer: The input layer has K nodes corresponding to the channel coefficients between K edge devices and the edge server. The input layer, denoted as z_0 , can be expressed as $z_0 = [h_1, \ldots, h_K]^T$.

2) Hidden Layer: The hidden layers between the input layer and the output layer are fully connected. We denote the output of the *d*-th neural layer as z_d and the number of nodes in the *d*-th layer as c_d . We leverage the activation functions to generate the estimated dual variables that are non-negative and continuous. The output of the *d*-th layer is given by

$$\boldsymbol{z}_d = \operatorname{ReLU}(\operatorname{BN}(\boldsymbol{Q}_d \boldsymbol{z}_{d-1} + \boldsymbol{b}_d)), \quad d = 1, \dots, D,$$
 (36)

where Q_d is a weight matrix of dimension $c_d \times c_{d-1}$, b_d is a vector of dimension c_d , ReLU(x) denotes the ReLU function (i.e., $\max(x, 0)$) that introduces nonlinearity, and BN(·) denotes the

batch normalization layer. The batch normalization layer is adopted to mitigate the sensitivity to the weight initialization and to reduce the probability of overfitting.

3) Output Layer: The dimension of the output layer c_{D+1} is set to K+1 and the output of the (D+1)-th layer is

$$\boldsymbol{z}_{D+1} = \operatorname{Sigmoid}(\boldsymbol{Q}_{D+1}\boldsymbol{z}_D + \boldsymbol{b}_{D+1}), \tag{37}$$

where $Sigmoid(x) = \exp(x)/(1 + \exp(x))$ denotes the sigmoid function.

An intuitive approach is to directly train a fully-connected neural network that predicts the transmit power of edge devices and the receive normalizing factor without a structure mapping layer. However, such an intuitive method cannot exploit the specific structure of the optimal solution. It is difficult for a fully-connected network to find the optimal transmit power without using the structure information. In contrast, we design the output of the output layer as the dual variables and the receive normalizing factor, i.e.,

$$\boldsymbol{z}_{D+1} = [\mu_1, \dots, \mu_K, \eta]^{\mathrm{T}}, \tag{38}$$

where the first K entries correspond to K dual variables, and the last entry is the receive normalizing factor. Vector z_{D+1} contains the key features of the transmit power of edge devices, and is passed through the structure mapping layer.

4) Structure Mapping Layer: In the structure mapping layer, the transmit powers $\{p_k\}$ of edge devices are calculated according to (35). Adopting the structure information not only reduces the computation complexity to find the optimal transmit power, but also provides performance guarantee for the proposed learning algorithm.

C. Deep Neural Network Training

The conventional DL models are usually trained using supervised learning, where the solution to the optimization problem is utilized as the ground truth and the squared error between the output of the neural network and the ground truth is the objective to be minimized. An obvious drawback of the supervised learning based algorithm is that a large number of labeled samples are required. However, collecting the optimal solutions to the optimization problem as labels is time-consuming. We in this paper adopt unsupervised learning, where only the CSI rather than the solutions of the optimization problem \mathcal{P} is required as training samples. The parameters of the neural network are optimized by using SGD. Besides the neural network design, the loss function design is also important in our proposed algorithm. We design our loss function as

time-average MSE plus regularizer as follows

$$loss = \frac{1}{B} \sum_{t=1}^{B} MSE(t) + \gamma \cdot Regularizer,$$
(39)

where B is the size of training batch and γ is the penalty parameter. To meet the average transmit power constraint, the regularizer that penalizes the constraint violation is defined as

$$\operatorname{Regularizer} = \sum_{k=1}^{K} \operatorname{ReLU}\left(\frac{1}{B} \sum_{t=1}^{B} p_k(t) - \bar{P}_k\right).$$
(40)

We adopt the ReLU function to ensure that the average transmit power is smaller than P_k rigorously. Note that the regularizer is not designed as $\sum_{k=1}^{K} \|\frac{1}{B} \sum_{t=1}^{B} p_k(t) - \bar{P}_k\|_2^2$ because ℓ_2 -norm can only ensure that the average transmit power to be close to \bar{P}_k .

VI. SIMULATION RESULTS

In this section, we present the simulation results of the proposed alternating optimization algorithm and knowledge-guided learning algorithm for AirComp-assisted FL.

A. Simulation Setup

In the simulations, the wireless channels between the edge devices and the edge server over different communication rounds follow i.i.d. Rayleigh fading. The number of edge devices is set to K = 20 if not specified otherwise. For each edge device, we define the receive signal-tonoise ratio (SNR) of device k as $\text{SNR}_k = \bar{P}_k / \sigma^2$, which is set to 10 dB. Besides, the maximum transmit power is set to $P_k^{\text{max}} = 3\bar{P}_k$. We describe the setting of other parameters as follows.

- Datasets: We adopt non-i.i.d. MNIST and CIFAR-10 datasets for FL training. In particular, we sort the data according to the labels, and divide the dataset into 200 shards with equal size. Each edge device is randomly assigned 2 data shards.
- 2) FL neural network: For the MNIST dataset, we train a convolutional neural network (CNN) with two convolution layers and two fully-connected layers. For the CIFAR-10 dataset, we train a CNN with three convolution layers and two fully-connected layers.
- 3) *Knowledge-based neural network for AirComp transceiver design:* The number of the hidden layers (i.e., *D*) is 2, while the numbers of nodes in two hidden layers are 256 and 64.

We compare the proposed alternating optimization algorithm and knowledge-guided learning algorithm with the following four baseline methods:

• Error-free transmission: The accumulated local gradients are assumed to be transmitted in an error-free manner, i.e., without suffering from channel fading and receiver noise. The

server receives each of the accumulated local gradients from all edge devices without any distortion. This benchmark characterizes the best FL performance.

- Full power: Each edge device transmits with a fixed power that is equal to the average transmit power budget \bar{P}_k . Besides, the receive normalizing factor is set to $\eta =$ $\left(\frac{\sigma^2 + \sum_{k=1}^{K} \bar{P}_k |h_k|^2}{\sum_{k=1}^{K} \sqrt{\bar{P}_k} |h_k|}\right)^2.$ Channel inversion: Each edge device transmits with power defined as follows

$$p_{k} = \begin{cases} \min\left(\bar{P}_{k}, \frac{\eta}{|h_{k}|^{2}}\right), & \bar{P}_{k}|h_{k}|^{2} \ge \epsilon_{c}, \\ 0, & \bar{P}_{k}|h_{k}|^{2} < \epsilon_{c}, \end{cases}$$

$$(41)$$

where $\eta = \min_k \left\{ \frac{\sigma^2 + \bar{P}_k |h_k|^2}{\sqrt{\bar{P}_k} |h_k|} \right\}$ and $\epsilon_c = 0.1$.

• Knowledge-free learning: A fully-connected neural network without structure information is trained to directly predict the transmit power of edge devices and the receive normalizing factor. Except for the structure mapping layer, the neural network structure is same as that of our proposed neural network. The output layer generates the transmit power of edge devices and the receive normalizing factor, i.e., $\boldsymbol{z}_{D+1} = [p_1, \dots, p_K, \eta]^{\mathrm{T}}$. Finally, transmit power p_k is multiplied by \bar{P}_k .

B. Performance Comparison

Fig. 3 compares the test accuracy and the training loss of all schemes under consideration when the number of local iterations (i.e., ϕ) is set to 3 and 50 for MNIST and CIFAR-10 datasets, respectively. We observe that the error-free transmission achieves the best performance because of the ideal model aggregation. The proposed alternating optimization algorithm obtains the optimal solutions of two subproblems and thus achieves the second-best convergence performance. Besides, the performance gap between the alternating optimization algorithm and the knowledge-guided learning algorithm is quite small. By exploiting the structure information in terms of the analytical expression of the optimal transmit power, the proposed knowledge-guided learning algorithm can effectively optimize the transceiver design, while significantly reduces the computation complexity. Compared to the full power method, the channel inversion method, and the knowledge-free learning method, the proposed knowledge-guided learning algorithm and the alternating optimization algorithm achieve faster convergence rates and better learning performance. This demonstrates the importance of the optimization of the transmit power and the receive normalizing factor as well as the exploitation of the structure information. Similar



(c) CIFAR-10 dataset (d) CIFAR-10 dataset Fig. 3. Performance comparison of the proposed optimization and knowledge-guided algorithms with the baseline methods.

performance trends in terms of training loss and test accuracy can be observed for all schemes under the MNIST and CIFAR-10 datasets.

Fig. 4 shows the learning performance of different algorithms with varying number of local iterations. The number of communication rounds (i.e., T) is set to 125 and 150 for MNIST and CIFAR-10 datasets, respectively. As shown in Fig. 4, the test accuracy increases with the number of local iterations. When the number of local iterations is large enough, the speed of performance increase becomes smaller. This is because excessive number of local iterations makes the local optimum deviate from the global mimimum when the datasets are non-i.i.d. Besides, we observe that the knowledge-guided learning algorithm always achieves a performance close to that of the alternating optimization algorithm on both MNIST and CIFAR-10 datasets. By exploiting the structure of the optimal transmit power, the proposed knowledge-guided learning algorithm outperforms the full power, channel inversion, and knowledge-free learning methods.

Fig. 5 shows the learning performance of different algorithms versus the number of edge devices when the number of communication rounds (i.e., T) and the number of local iterations



(c) CIFAR-10 dataset (d) CIFAR-10 dataset Fig. 4. Learning performance versus number of local iterations in terms of test accuracy and training loss.

(i.e., ϕ) are set to 125 and 2, respectively. We observe that the test accuracy of all schemes increases with the number of edge devices. Specifically, the test accuracy increases rapidly when $K \leq 20$, and increases slowly when $K \geq 25$. This is because data redundancy occurs when too many edge devices are involved in FL training. Fig. 5 also shows that the knowledge-guided learning algorithm is able to achieve comparable performance with the alternating optimization algorithm under different number of edge devices. In addition, the knowledge-guided learning algorithm and the alternating optimization algorithm significantly outperform the full power method, the channel-inversion method, and the knowledge-free learning method, which clearly demonstrates the superiority of our proposed algorithms.

TABLE I

Computation time comparison and feasible probability of proposed knowledge-guided learning algorithm when T=200



TABLE II

Computation time comparison and feasible probability of proposed knowledge-guided learning algorithm when K = 20

Number of communication rounds	125	150	175	200	225	250
Alternating optimization	31.22 s	39.98 s	46.32 s	53.53 s	58.58 s	67.61 s
Knowledge-guided learning	14.42 ms	15.37 ms	14.83 ms	14.84 ms	15.67 ms	17.01 ms
Feasible probability of solutions	99.71%	99.66%	99.96%	99.98%	99.99%	99.99%



Fig. 5. Test accuracy versus number of edge devices for MNIST dataset.

We compare the computation time of the knowledge-guided learning algorithm and the alternating optimization algorithm, and evaluate the feasible probability of the solutions returned by the knowledge-guided learning algorithm. Table I shows the computation time versus the number of edge devices when the number of communication rounds (i.e., T) is set to 200. We observe that the computation time grows as the number of edge devices increases. The computation time of the knowledge-guided learning algorithm is three orders of magnitude smaller than the alternating optimization algorithm. This is because the alternating optimization algorithm relies on an iterative process and each iteration involves a bisection search, which are time-consuming. Results demonstrate that the knowledge-guided learning algorithm is more practical for the transceiver design of AirComp-assisted FL with only marginal decrease in learning performance.

Table II shows that the computation time when the number of edge devices (i.e., K) is set to 20. Compared with the alternating optimization algorithm, the knowledge-guided learning algorithm can speed up the computation by 2000 - 4000 times. Both Tables I and II show that the probability of feasible solution of the knowledge-guided learning algorithm is more than 99.6% for different number of communication rounds and different number of edge devices, which demonstrates the robustness of our proposed learning algorithm.

VII. CONCLUSION

In this paper, we studied over-the-air FL, taking in account multiple local SGD iterations and non-i.i.d. data. We derived the convergence of AirComp-assisted FL in terms of the timeaverage norm of gradients, followed by formulating an optimization problem to minimize the convergence bound. We first proposed an alternating optimization algorithm to obtain the optimal transmit power of edge devices and the receive normalizing factor, which, however, requires the global CSI and suffers from high computation complexity. We further developed a knowledgeguided learning algorithm that exploits the domain knowledge to map the instantaneous CSI to the transmit power of edge devices and the receive normalizing factor. Simulation results demonstrated that the knowledge-guided learning algorithm achieves a comparable performance as the alternating optimization algorithm, but with a much lower computation complexity.

Appendix

A. Proof of Theorem 1

Before proving Theorem 1, we first present the following four useful lemmas, which are proved in Appendix C.

Lemma 1. With Assumption 3, the difference between the global model vector and the individual local model vector is bounded, i.e.,

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{w}(t) - \boldsymbol{w}_{k}(t,\zeta)\|_{2}^{2}\right] \leq \phi\lambda^{2}\xi^{2} + \phi\lambda^{2}\sum_{q=0}^{\phi-1}\chi \left\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k}(\boldsymbol{w}_{k}(t,q))\right\|_{2}^{2}.$$
 (42)

Lemma 2. With Assumptions 2 and 3, the following equality holds

$$\mathbb{E}\left[\langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{\theta}(t) \rangle\right] \geq \frac{1}{2} \phi \|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} + \frac{1}{2} \sum_{\zeta=0}^{\phi-1} \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta)) \right\|_{2}^{2} - L^{2} \frac{1}{2} \sum_{\zeta=0}^{\phi-1} \frac{1}{K} \sum_{k=1}^{K} \|(\boldsymbol{w}(t) - \boldsymbol{w}_{k}(t,\zeta))\|_{2}^{2}.$$
(43)

Lemma 3. With Assumption 3, the average accumulated gradient norm is upper bounded as follows

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t)\|_{2}^{2}\right] \leq \frac{\phi}{K}\xi^{2} + \phi\sum_{\zeta=0}^{\phi-1} \left\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2},\tag{44}$$

which relates the average accumulated gradient to the local full gradient.

Lemma 4. With Assumptions 3 and 4, the aggregation error and the instantaneous MSE has the following relationship

$$\mathbb{E}[\|\boldsymbol{e}(t)\|_{2}^{2}] \leq N \frac{\Gamma(K+1)}{K^{2}} \mathbf{MSE}(t).$$
(45)

Proof of Theorem 1. According to Assumption 2, F(w) is *L*-smooth and we have the following inequality

$$F(\boldsymbol{w}(t+1)) - F(\boldsymbol{w}(t)) \leq -\lambda \langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{\theta}(t) + \boldsymbol{e}(t) \rangle + \frac{\lambda^2 L}{2} \|\boldsymbol{\theta}(t) + \boldsymbol{e}(t)\|_2^2$$

$$= -\lambda \langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{\theta}(t) \rangle - \lambda \langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{e}(t) \rangle + \frac{\lambda^2 L}{2} \|\boldsymbol{\theta}(t)\|_2^2 + \frac{\lambda^2 L}{2} \|\boldsymbol{e}(t)\|_2^2 + \lambda^2 L \langle \boldsymbol{\theta}(t), \boldsymbol{e}(t) \rangle$$

$$\stackrel{(a)}{\leq} -\lambda \langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{\theta}(t) \rangle + \frac{\lambda}{2} \|\nabla F(\boldsymbol{w}(t))\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{e}(t)\|_2^2 + \frac{\lambda^2 L}{2} \|\boldsymbol{\theta}(t)\|_2^2 + \frac{\lambda^2 L}{2} \|\boldsymbol{e}(t)\|_2^2 + \lambda^2 L \langle \boldsymbol{\theta}(t), \boldsymbol{e}(t) \rangle$$

$$\stackrel{(b)}{\leq} -\lambda \left\langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{\theta}(t) \right\rangle + \frac{\lambda}{2} \|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} + (\frac{\lambda}{2} + \lambda^{2}L) \|\boldsymbol{e}(t)\|_{2}^{2} + \lambda^{2}L \|\boldsymbol{\theta}(t)\|_{2}^{2},$$

$$(46)$$

where (a) follows from $-a^T b \le \frac{\|a\|_2^2}{2} + \frac{\|b\|_2^2}{2}$ and (b) follows by utilizing $a^T b \le \frac{\|a\|_2^2}{2} + \frac{\|b\|_2^2}{2}$. By taking expectations over stochastic sampling and receiver noise at both sides of (46), we obtain

$$\mathbb{E}[F(\boldsymbol{w}(t+1)) - F(\boldsymbol{w}(t))] \leq -\lambda \mathbb{E}\left[\left[\langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{\theta}(t) \rangle\right]\right] \\ + \frac{\lambda}{2} \|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} + \left(\frac{\lambda}{2} + \lambda^{2}L\right) \mathbb{E}\left[\|\boldsymbol{e}(t)\|_{2}^{2}\right] + \lambda^{2}L \mathbb{E}\left[\|\boldsymbol{\theta}(t)\|_{2}^{2}\right].$$
(47)

Using Lemmas 1, 2, 3 and 4, we have

$$\begin{split} \mathbb{E}[F(\boldsymbol{w}(t+1)) - F(\boldsymbol{w}(t))] &= \frac{\lambda(1-\phi)}{2} \|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} + \left(\frac{\phi^{2}\lambda^{3}L^{2}}{2} + \frac{\phi\lambda^{2}L}{K}\right)\xi^{2} \\ &+ (-1+\phi^{2}L^{2}\lambda^{2}\chi + 2\phi\lambda L)\frac{\lambda}{2}\sum_{\zeta=0}^{\phi-1} \left\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2} + (\frac{\lambda}{2}+\lambda^{2}L)\mathbb{E}[\|\boldsymbol{e}(t)\|_{2}^{2}] \\ &\stackrel{(a)}{\leq}\frac{\lambda(1-\phi)}{2} \|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} + \left(\frac{\phi^{2}\lambda^{3}L^{2}}{2} + \frac{\phi\lambda^{2}L}{K}\right)\xi^{2} + (\frac{\lambda}{2}+\lambda^{2}L)N\frac{\Gamma(K+1)}{K^{2}}\mathrm{MSE}(t), \quad (48) \\ &\text{here (a) holds because } \phi^{2}L^{2}\lambda^{2}\chi + 2\phi\lambda L < 1. \text{ By summing up (48) for all } T \text{ communication} \end{split}$$

where (a) holds because $\phi^2 L^2 \lambda^2 \chi + 2\phi \lambda L \leq 1$. By summing up (48) for all T communication rounds and rearranging the terms, we have

$$\mathbb{E}[F(\boldsymbol{w}(T)) - F(\boldsymbol{w}(0))] \leq \frac{\lambda}{2}(1-\phi)\sum_{t=0}^{T-1} \|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} + \left(\frac{\phi^{2}\lambda^{3}L^{2}}{2} + \frac{\phi\lambda^{2}L}{K}\right)\xi^{2} + \left(\frac{\lambda}{2} + \lambda^{2}L\right)\sum_{t=0}^{T-1} N\frac{\Gamma(K+1)}{K^{2}} \mathsf{MSE}(t).$$
(49)

With Assumption 1, we have $F(\boldsymbol{w}(T)) - F(\boldsymbol{w}(0)) \ge F(\boldsymbol{w}^*) - F(\boldsymbol{w}(0))$, which yields (21).

B. Proof of Theorem 2

The Lagrangian function of problem (31) is given by

$$\mathcal{L}(\{p_k(t)\},\{\alpha(t)\},\mu_k) = \sum_{t=0}^{T-1} \left(\frac{\sqrt{p_k(t)}|h_k(t)|}{\sqrt{\eta(t)}} - 1\right)^2 + \sum_{t=0}^{T-1} \alpha(t) \left(p_k(t) - P_k^{\max}\right) + \mu_k \left(\sum_{t=0}^{T-1} p_k(t) - T\bar{P}_k\right),$$

where $\{\alpha(t) \ge 0\}$ denote the dual variables associated with the constraints in (31b) and $\mu_k \ge 0$

denotes the dual variable associated with constraint (31c). By setting the first derivative of $\mathcal{L}(\{p_k(t)\}, \{\alpha(t)\}, \mu_k)$ with respect to $p_k(t)$ to zero as follows

$$\frac{\partial \mathcal{L}}{\partial p_k(t)} = \frac{|h_k(t)|^2}{\eta(t)} - \frac{|h_k(t)|}{\sqrt{\eta(t)p_k(t)}} + \alpha(t) + \mu_k = 0, \tag{50}$$

we obtain

$$p_k(t) = \left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + (\alpha(t) + \mu_k)\eta(t)}\right)^2.$$
(51)

We denote $\{p_k^*(t)\}$ as the optimal transmit power, and $\{\alpha^*(t)\}$ and μ_k^* as the optimal dual

$$p_k^*(t) = \left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + (\alpha^*(t) + \mu_k^*)\eta(t)}\right)^2, \forall t,$$
(52)

$$0 \le p_k^*(t) \le P_k^{\max}, \forall t,$$
(53)

$$0 \le \sum_{t=0}^{T-1} p_k^*(t) \le T\bar{P}_k, \tag{54}$$

$$\alpha^*(t) \ge 0, \forall t, \tag{55}$$

$$\mu_k^* \ge 0,\tag{56}$$

$$\alpha^*(t) \left(p_k^*(t) - P_k^{\max} \right) = 0, \forall t,$$
(57)

$$\mu_k^* \left(\sum_{t=0}^{T-1} p_k^*(t) - T\bar{P}_k \right) = 0.$$
(58)

If $\alpha^*(t) > 0$, because of the comlementary slackness condition (57), then we obtain $p_k^*(t) = P_k^{\max}$. If $\alpha^*(t) = 0$, then we obtain $p_k^*(t) = \left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right)^2$. Hence, the analytical expression of $p_k^*(t)$ is given by

$$p_{k}^{*}(t) = \begin{cases} \left(\frac{\sqrt{\eta(t)}|h_{k}(t)|}{|h_{k}(t)|^{2} + \mu_{k}^{*}\eta(t)}\right)^{2}, & \text{if } \alpha^{*}(t) = 0, \\ P_{k}^{\max}, & \text{if } \alpha^{*}(t) > 0. \end{cases}$$
(59)

As the value of $p_k^*(t)$ depends on $\alpha^*(t)$, we discuss two following cases. If $\left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right)^2 > P_k^{\max}$ and $\alpha^*(t) = 0$, then

$$p_k^*(t) \stackrel{(a)}{=} \left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right)^2 > P_k^{\max},\tag{60}$$

where (a) is due to (59). Note that the maximum power constraint (53) is not satisfied. Hence, $\alpha^*(t) > 0$ must hold, resulting in $p_k^*(t) = P_k^{\max}$. If $\left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right)^2 \le P_k^{\max}$ and $\alpha^*(t) > 0$, then $p_k^*(t) \stackrel{(a)}{=} P_k^{\max} \ge \left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right)^2 \ge \left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + (\alpha^*(t) + \mu_k^*)\eta(t)}\right)^2$, (61) where (a) is due to (59) and (b) is due to $\alpha^*(t) > 0$. Note that (52) is not satisfied. Hence,

where (a) is due to (59) and (b) is due to $\alpha^*(t) > 0$. Note that (52) is not satisfied. Hence, $\alpha^*(t) = 0$ must holds, which leads to $p_k^*(t) = \left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right)^2$. To sum up, the value of $p_k^*(t)$ is independent of $\alpha^*(t)$ and is given by $p_k^*(t) = \min\left\{\left(\frac{\sqrt{\eta(t)}|h_k(t)|}{|h_k(t)|^2 + \mu_k^*\eta(t)}\right)^2, P_k^{\max}\right\}$.

If $\mu_k^* > 0$, because of the complementary slackness condition (58), then we obtain $\sum_{t=0}^{T-1} p_k^*(t) = T\bar{P}_k$. Hence, we obtain (34), where μ_k^* can be found to ensure $\sum_{t=0}^{T-1} p_k^*(t) = T\bar{P}_k$ by using the one-dimensional bisection search method. Furthermore, if $\mu_k^* = 0$ and $\sum_{t=0}^{T-1} p_{k,m}^*(t) \leq T\bar{P}_k$,

then we obtain (33). If $\mu^* = 0$ and $\sum_{t=0}^{T-1} p_k^*(t) > T\bar{P}_k$, then (54) does not hold. Hence, $\mu_k^* > 0$ must holds, which leads to (34).

To sum up, the optimal transmit power is given by

$$p_{k}^{*}(t) = \begin{cases} \min\left\{\frac{\eta(t)}{|h_{k}(t)|^{2}}, P_{k}^{\max}\right\}, & \text{if } \sum_{t=0}^{T-1}\min\left\{\frac{\eta(t)}{|h_{k}(t)|^{2}}, P_{k}^{\max}\right\} \le T\bar{P}_{k}, \\ \min\left\{\left(\frac{\sqrt{\eta(t)}|h_{k}(t)|}{|h_{k}(t)|^{2} + \mu_{k}^{*}\eta(t)}\right)^{2}, P_{k}^{\max}\right\}, & \text{if } \sum_{t=0}^{T-1}\min\left\{\frac{\eta(t)}{|h_{k}(t)|^{2}}, P_{k}^{\max}\right\} > T\bar{P}_{k}. \end{cases}$$
(62)

where μ^* can be found to ensure the average power constraint $\sum_{t=0}^{T-1} p_k^*(t) = T\bar{P}_k$ via the one-dimensional bisection search method.

C. Proof of Lemmas

Proof of Lemma 1. According to (4), the local model vector is given by
$$\boldsymbol{w}_{k}(t,\zeta) = \boldsymbol{w}(t) - \lambda \sum_{q=0}^{\zeta-1} \tilde{\boldsymbol{g}}_{k}(t,q)$$
. We bound $\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{w}(t) - \boldsymbol{w}_{k}(t,\zeta)\|_{2}^{2}\right]$ as follows

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{w}(t) - \boldsymbol{w}_{k}(t,\zeta)\|_{2}^{2}\right] = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\lambda \sum_{q=0}^{\zeta-1} \tilde{\boldsymbol{g}}_{k}(t,q)\|_{2}^{2}\right] = \frac{1}{K}\sum_{k=1}^{K} \lambda^{2}\mathbb{E}\left[\left\|\sum_{q=0}^{\zeta-1} \tilde{\boldsymbol{g}}_{k}(t,q)\right\|_{2}^{2}\right]$$

$$\stackrel{(a)}{=} \frac{1}{K}\sum_{k=1}^{K} \lambda^{2} \operatorname{Var}\left(\sum_{q=0}^{\zeta-1} \tilde{\boldsymbol{g}}_{k}(t,q)\right) + \frac{1}{K}\sum_{k=1}^{K} \lambda^{2} \left\|\sum_{q=0}^{\zeta-1} \nabla F_{k}(\boldsymbol{w}_{k}(t,q))\right\|_{2}^{2} \stackrel{(b)}{=} \frac{1}{K}\sum_{k=1}^{K} \lambda^{2} \sum_{q=0}^{\zeta-1} \operatorname{Var}\left(\tilde{\boldsymbol{g}}_{k}(t,q)\right)$$

$$+ \frac{1}{K}\sum_{k=1}^{K} \lambda^{2} \left\|\sum_{q=0}^{\zeta-1} \nabla F_{k}(\boldsymbol{w}_{k}(t,q))\right\|_{2}^{2} \leq \frac{1}{K}\sum_{k=1}^{K} \lambda^{2} \sum_{q=0}^{\zeta-1} \operatorname{Var}\left(\tilde{\boldsymbol{g}}_{k}(t,q)\right) + \frac{1}{K}\sum_{k=1}^{K} \lambda^{2} \sum_{q=0}^{\zeta-1} ||\nabla F_{k}(\boldsymbol{w}_{k}(t,q))||_{2}^{2}$$

$$\leq \frac{1}{K}\sum_{k=1}^{K} \lambda^{2} \sum_{q=0}^{\phi-1} \operatorname{Var}\left(\tilde{\boldsymbol{g}}_{k}(t,q)\right) + \frac{1}{K}\sum_{k=1}^{K} \phi\lambda^{2} \sum_{q=0}^{\phi-1} ||\nabla F_{k}(\boldsymbol{w}_{k}(t,q))||_{2}^{2}$$

$$\leq \lambda^{2} \phi\xi^{2} + \phi\lambda^{2} \sum_{q=0}^{\phi-1} \frac{1}{K}\sum_{k=1}^{K} ||\nabla F_{k}(\boldsymbol{w}_{k}(t,q))||_{2}^{2} \stackrel{(c)}{\leq} \phi\lambda^{2}\xi^{2} + \phi\lambda^{2} \sum_{q=0}^{\phi-1} \chi \left\|\frac{1}{K}\sum_{k=1}^{K} \nabla F_{k}(\boldsymbol{w}_{k}(t,q))\right\|_{2}^{2},$$
(63)

where (a) follows from $\mathbb{E}[\boldsymbol{x}^2] = \operatorname{Var}[\boldsymbol{x}] + [\mathbb{E}[\boldsymbol{x}]]^2$ and Assumption 3, (b) holds because $\operatorname{Var}(\sum_{j=1}^n \boldsymbol{x}_j) = \sum_{j=1}^n \operatorname{Var}(\boldsymbol{x}_j)$ with independent $\{\boldsymbol{x}_j\}$, and (c) follows from Definition 1.

Proof of Lemma 2. Recall the definitions of $\boldsymbol{\theta}(t)$ and $\boldsymbol{\theta}_k(t)$, we have $\boldsymbol{\theta}(t) = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\theta}_k(t) = \frac{1}{K} \sum_{k=1}^{K} \sum_{\zeta=0}^{\phi-1} \tilde{\boldsymbol{g}}_k(t,\zeta)$. Thus, we have

$$\mathbb{E}\left[\left\langle \nabla F(\boldsymbol{w}(t)), \boldsymbol{\theta}(t) \right\rangle\right] \stackrel{(a)}{=} \left\langle \nabla F(\boldsymbol{w}(t)), \frac{1}{K} \sum_{k=1}^{K} \sum_{\zeta=0}^{\phi-1} \nabla F_k(\boldsymbol{w}_k(t,\zeta)) \right\rangle$$
$$= \sum_{\zeta=0}^{\phi-1} \left\langle \nabla F(\boldsymbol{w}(t)), \frac{1}{K} \sum_{k=1}^{K} \nabla F_k(\boldsymbol{w}_k(t,\zeta)) \right\rangle \stackrel{(b)}{=} \frac{1}{2} \sum_{\zeta=0}^{\phi-1} \left[\|\nabla F(\boldsymbol{w}(t))\|_2^2 + \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla F_k(\boldsymbol{w}_k(t,\zeta)) \right\|_2^2 \right]$$

$$-\left\|\frac{1}{K}\sum_{k=1}^{K}(\nabla F_{k}(\boldsymbol{w}(t)) - \nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta)))\right\|_{2}^{2}\right] \geq \frac{1}{2}\sum_{\zeta=0}^{\phi-1}\left[\left\|\nabla F(\boldsymbol{w}(t))\right\|_{2}^{2} + \left\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2}\right] - \frac{1}{K}\sum_{k=1}^{K}\left\|(\nabla F_{k}(\boldsymbol{w}(t)) - \nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta)))\right\|_{2}^{2}\right] \geq \frac{1}{2}\sum_{\zeta=0}^{\phi-1}\left[\left\|\nabla F(\boldsymbol{w}(t))\right\|_{2}^{2} + \left\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2}\right] - L^{2}\frac{1}{K}\sum_{k=1}^{K}\left\|(\boldsymbol{w}(t) - \boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2}\right] = \frac{1}{2}\phi\|\nabla F(\boldsymbol{w}(t))\|_{2}^{2} + \frac{1}{2}\sum_{\zeta=0}^{\phi-1}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2}$$

$$-L^{2}\frac{1}{2}\sum_{\zeta=0}^{\phi-1}\frac{1}{K}\sum_{k=1}^{K}\left\|(\boldsymbol{w}(t) - \boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2},$$
(64)

where (a) is due to Assumption 3, (b) is due to $a^T b = \frac{1}{2} ||a||_2^2 + \frac{1}{2} ||b||_2^2 - \frac{1}{2} ||a - b||_2^2$, and (c) is due to Assumption 2.

Proof of Lemma 3. According to the definition of $\theta(t)$, we have

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t)\|_{2}^{2}\right] = \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^{K}\sum_{\zeta=0}^{\phi-1}\tilde{\boldsymbol{g}}_{k}(t,\zeta)\right\|_{2}^{2}\right] \stackrel{(a)}{=} \operatorname{Var}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{\zeta=0}^{\phi-1}\tilde{\boldsymbol{g}}_{k}(t,\zeta)\right) + \left\|\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\sum_{\zeta=0}^{\phi-1}\tilde{\boldsymbol{g}}_{k}(t,\zeta)\right]\right\|_{2}^{2}\right] \stackrel{(b)}{=} \frac{1}{K^{2}}\sum_{k=1}^{K}\operatorname{Var}\left(\sum_{\zeta=0}^{\phi-1}\tilde{\boldsymbol{g}}_{k}(t,\zeta)\right) + \phi\sum_{\zeta=0}^{\phi-1}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{k}(\boldsymbol{w}_{k}(t,\zeta))\right\|_{2}^{2}, \tag{65}$$
where (a) is due to $\mathbb{E}[\|\boldsymbol{x}\|^{2}] = \operatorname{Var}[\boldsymbol{x}] + \|\mathbb{E}[\boldsymbol{x}]\|^{2}$ and (b) is due to $\operatorname{Var}(\sum_{i=1}^{n}\boldsymbol{x}_{i}) = \sum_{i=1}^{n}\operatorname{Var}(\boldsymbol{x}_{i})$

where (a) is due to $\mathbb{E}[\|\boldsymbol{x}\|^2] = \operatorname{Var}[\boldsymbol{x}] + \|\mathbb{E}[\boldsymbol{x}]\|^2$ and (b) is due to $\operatorname{Var}(\sum_{j=1}^n \boldsymbol{x}_j) = \sum_{j=1}^n \operatorname{Var}(\boldsymbol{x}_j)$ with independent $\{\boldsymbol{x}_j\}$. With Assumption 3, we have $\operatorname{Var}\left(\sum_{\zeta=0}^{\phi-1} \tilde{\boldsymbol{g}}_k(t,\zeta)\right) = {}^{(a)} \sum_{\zeta=0}^{\phi-1} \operatorname{Var}(\tilde{\boldsymbol{g}}_k(t,\zeta)) \leq \phi\xi^2$, which yields (44).

Proof of Lemma 4. From the definition of e(t), we have

$$\mathbb{E}[\|\boldsymbol{e}(t)\|_{2}^{2}] = \mathbb{E}\left[\left\|\frac{1}{K}\pi(t)\left(\hat{\boldsymbol{s}}(t) - \boldsymbol{s}(t)\right)\right\|_{2}^{2}\right] \leq \frac{\Gamma}{K^{2}}\mathbb{E}\left[\left\|\left(\hat{\boldsymbol{s}}(t) - \boldsymbol{s}(t)\right)\right\|_{2}^{2}\right] \\
\leq \frac{\Gamma(K+1)}{K^{2}}\left\{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\left(\frac{\sqrt{p_{k}(t)}|h_{k}(t)|}{\sqrt{\eta(t)}}\boldsymbol{I} - \boldsymbol{I}\right)\boldsymbol{s}_{k}(t)\right\|_{2}^{2}\right] + \mathbb{E}\left[\left\|\frac{\boldsymbol{n}(t)}{\sqrt{\eta(t)}}\right\|_{2}^{2}\right]\right\} \\
= \frac{\Gamma(K+1)}{K^{2}}\left\{\sum_{k=1}^{K}N\left(\frac{\sqrt{p_{k}(t)}|h_{k}(t)|}{\sqrt{\eta(t)}} - 1\right)^{2} + \frac{N\sigma^{2}}{\eta(t)}\right\} = N\frac{\Gamma(K+1)}{K^{2}}\mathsf{MSE}(t). \quad (66)$$

REFERENCES

- B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artificial Intell. Stat. (AISTATS)*, 2017, pp. 1273–1282.
- [2] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 2020.
- [3] W. Y. B. Lim, J. S. Ng, Z. Xiong, D. Niyato, C. Miao, and D. I. Kim, "Dynamic edge association and resource allocation in self-organizing hierarchical federated learning networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3640–3653, Dec. 2021.

- [4] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2167–2191, Jul. 2020.
- [5] Y. Shi, K. Yang, Z. Yang, and Y. Zhou, *Mobile Edge Artificial Intelligence: Opportunities and Challenges*. Elsevier, 2021.
- [6] T. K. Rodrigues, K. Suto, H. Nishiyama, J. Liu, and N. Kato, "Machine learning meets computation and communication control in evolving edge and cloud: Challenges and future perspective," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 38–67, 2019.
- [7] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Netw.*, vol. 34, no. 5, pp. 16–22, Sept. 2020.
- [8] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "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.
- [9] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.
- [10] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, Oct. 2007.
- [11] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.
- [12] 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.
- [13] M. M. 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.
- [14] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Letaief, "Federated learning via intelligent reflecting surface," *IEEE Trans. Wireless Commun.*, 2022.
- [15] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, Aug. 2021.
- [16] 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.*, Nov. 2021.
- [17] C. Xu, S. Liu, Z. Yang, Y. Huang, and K.-K. Wong, "Learning rate optimization for federated learning exploiting overthe-air computation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3742–3756, Dec. 2021.
- [18] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, Feb. 2021.
- [19] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE J. Sel. Areas Commun.*, Jul. 2020.
- [20] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2021.
- [21] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE J. Sel. Areas Commun.*, 2021.
- [22] C. T. Dinh, N. H. Tran, M. N. Nguyen, C. S. Hong, W. Bao, A. Y. Zomaya, and V. Gramoli, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 398–409, Feb. 2021.
- [23] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

- [24] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 453–467, Jan. 2021.
- [25] M. M. Wadu, S. Samarakoon, and M. Bennis, "Joint client scheduling and resource allocation under channel uncertainty in federated learning," *IEEE Trans. Commun.*, Sept. 2021.
- [26] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing," *IEEE Trans. Wireless Commun.*, Dec. 2021.
- [27] J. Ren, G. Yu, and G. Ding, "Accelerating dnn training in wireless federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 219–232, Jan. 2021.
- [28] R. Paul, Y. Friedman, and K. Cohen, "Accelerated gradient descent learning over multiple access fading channels," *IEEE J. Sel. Areas Commun.*, 2021.
- [29] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.
- [30] F. Liang, C. Shen, W. Yu, and F. Wu, "Towards optimal power control via ensembling deep neural networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1760–1776, Mar. 2020.
- [31] Y. Li, S. Han, and C. Yang, "Multicell power control under rate constraints with deep learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7813–7825, Dec. 2021.
- [32] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "Graph neural networks for scalable radio resource management: Architecture design and theoretical analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 101–115, Jan. 2021.
- [33] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 77–83, Oct. 2019.
- [34] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "Comnet: Combination of deep learning and expert knowledge in ofdm receivers," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2627–2630, Dec. 2018.
- [35] Y. Shi, H. Choi, Y. Shi, and Y. Zhou, "Algorithm unrolling for massive access via deep neural network with theoretical guarantee," *IEEE Trans. Wireless Commun.*, 2021.
- [36] Y. Zou, Y. Zhou, Y. Shi, and X. Chen, "Learning proximal operator methods for massive connectivity in IoT networks," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Dec. 2021.
- [37] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, Mar. 2019.
- [38] O. Abari, H. Rahul, D. Katabi, and M. Pant, "Airshare: Distributed coherent transmission made seamless," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2015, pp. 1742–1750.
- [39] A. Mahmood, M. I. Ashraf, M. Gidlund, J. Torsner, and J. Sachs, "Time synchronization in 5G wireless edge: Requirements and solutions for critical-MTC," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 45–51, Dec. 2019.
- [40] M. Fu, Y. Zhou, Y. Shi, W. Chen, and R. Zhang, "UAV aided over-the-air computation," *IEEE Trans. Wireless Commun.*, 2021.
- [41] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 560–569.
- [42] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," in *Proc. Int. Conf. Artificial Intell. Stat. (AISTATS)*, 2021, pp. 2350–2358.
- [43] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," 2018. [Online]. Available: https://arxiv.org/abs/1808.07576