Over-the-Air Decentralized Federated Learning

Yandong Shi*†‡, Yong Zhou* , and Yuanming Shi*
*School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
†Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, China
‡University of Chinese Academy of Sciences, Beijing 100049, China
Email: {shiyd, zhouyong, shiym}@shanghaitech.edu.cn

***Abstract*—In this paper, we consider decentralized federated learning (FL) over wireless networks, where over-the-air computation (AirComp) is adopted to facilitate the local model consensus in a device-to-device (D2D) communication manner. However, the AirComp-based consensus phase brings the additive noise in each algorithm iterate and the consensus needs to be robust to wireless network topology changes, which introduce a coupled and novel challenge of establishing the convergence for wireless decentralized FL algorithm. To facilitate consensus phase, we propose an AirComp-based DSGD with gradient tracking and variance reduction (DSGT-VR) algorithm, where both precoding and decoding strategies are developed for D2D communication. Furthermore, we prove that the proposed algorithm converges linearly and establish the optimality gap for strongly convex and smooth loss functions, taking into account the channel fading and noise. The theoretical result shows that the additional error bound in the optimality gap depends on the number of devices. Extensive simulations verify the theoretical results and show that the proposed algorithm outperforms other benchmark decentralized FL algorithms over wireless networks.**

I. Introduction

Federated learning (FL), as a novel type of distributed machine learning [1], has received a growing interest recently, both from academia and industry. With FL, many mobile devices collaboratively train a global model under the orchestration of a parameter server (PS), while keeping the training data unmoved. The PS only receives the local model rather than raw data from mobile devices, thereby significantly reducing the overhead of data collection and preserving additional privacy [2], [3]. FL has enabled a wide range of applications in industrial Internet of Things (IIoT) [4]–[6], e.g., autonomous driving and collaborative robotics, in which high levels of security, privacy and reliability are demanded. In future wireless networks, the driving applications of FL including resource management, channel estimation and signal detection, edge caching, and computation offloading [7], [8], which also poses unique challenges including statistical heterogeneity, communication cost and resource allocation [9]. It thus becomes critical to implement FL for the future 6G wireless networks [10], [11].

Due to limited radio resources, communication is a fundamental bottleneck in FL over wireless networks [12], [13]. The digital [14]–[16] and analog [17]–[19] transmission schemes have been proposed to support FL over wireless networks. In particular, the authors in [16] partitioned a large-scale learning task over multiple resource-constrained edge devices, where joint parameter-and-bandwidth allocation was proposed to reduce the total learning-and-communication latency. In [14], [15], a joint learning, wireless resource allocation, and user selection problem was formulated to minimize FL training loss and FL convergence time. However, the orthogonal channels are required in digital FL systems which suffer from hugely demanding bandwidth especially when number of edge devices is large [18]. As for analog transmission scheme, over-the-air computation (AirComp) was proposed to support fast model aggregation for FL by exploiting the superposition property of multi-access channel (MAC), where multiple edge devices share the same frequency channel [17]–[19]. Specifically, a joint device selection and beamforming design was proposed to improve the statistical learning performance for FL [17]. The authors in [18], [19] focused on designing a stochastic gradient descent (SGD) based algorithm over MAC and investigate the effect of channel fading and noise on the convergence analysis of FL.

The aforementioned studies require a central PS to orchestrate the training process. However, in some application scenarios, e.g., cooperative driving and robotics [6], a central PS may not always be available and reliable when the number of edge devices is large [20]. The centralized FL also faces straggler's dilemma [21] due to the heterogeneity of edge devices, i.e., the FL training speed is limited by the devices with slowest computation and worst channel conditions. To overcome these challenges, device-to-device (D2D) communications based decentralized FL [22]–[24] was proposed, where each device only communicates with its neighbors. In particular, [24] considered digital transmission schemes in a joint learning and network simulation framework, to quantify the effects of model pruning, quantization and physical layer constraints for decentralized FL. Due to limited wireless bandwidth resources, the authors in [22], [23] proposed a decentralized stochastic gradient descent (DSGD) algorithm to improve the convergence performance in decentralized FL, where AirComp based D2D communication was developed to facilitate the consensus phase. However, the AirComp-based consensus phase brings the channel fading and additive noise in each algorithm iterate. On the other hand, the neighborhood weighted average involving the information of network topology performs the consensus phase. The decentralized

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant 62001294.

FL algorithm thus needs to be robust against changes in network topology, otherwise it may not converge [20]. These introduce a coupled and unique challenge of establishing the convergence of decentralized FL algorithm over wireless networks.

In this paper, we shall consider a decentralized FL model over wireless networks, where no central PS exists to orchestrate the training process. To investigate and facilitate consensus phase over wireless network, we propose an AirComp-based DSGT-VR algorithm in decentralized FL, where both precoding and decoding strategies at devices are developed to guarantee algorithm convergence. In addition, the gradient tracking and variance reduction techniques are involved in the proposed algorithm, which can further improve algorithm convergence performance. We analyze the performance of AirComp-based DSGT-VR algorithm theoretically by introducing the auxiliary variable in consensus phase, i.e., the mean of all local parameters. For strongly convex and smooth local loss functions, we prove that the proposed algorithm converges linearly and also establish the optimality gap, taking into account the channel fading and noise. In addition, the convergence result shows that the additive error bound in the optimality gap depends on the number of devices. Numerical experiments are conducted to validate the theoretical analysis and demonstrate the superior performance of the proposed algorithm over wireless networks.

Notations: Throughout this paper, we denote the cardinality of set A by $|A|$, ℓ_2-norm of vector $\boldsymbol{x}$ by $\|\boldsymbol{x}\|_2$ and the second largest singular value of matrix $\boldsymbol{W}$ by $|||\boldsymbol{W}|||$.

II. System Model and Problem Formulation

A. System Model

As shown in Fig. 1, we consider a federated learning system supported by a decentralized wireless communication network in decentralized setting where no central PS exists to coordinate training process of all edge devices [25]. We denote $\mathcal{N} = \{1, ..., N\}$ as the set of devices and $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ as an undirected graph with vertex set $\mathcal{N}$ and edge set $\mathcal{E}$ that represents the set of communication links. The set of connected neighbors of device i is denoted as $\mathcal{N}_i = \{j|(i,j) \in \mathcal{E}\}$. Each device $n \in \mathcal{N}$ has its own data set $\mathcal{D}_n$ and all devices aim to collaboratively learn a common machine learning model by communicating with each other through device-to-device communication link over graph $\mathcal{G}$.

The goal of this paper is to learn a global model by tackling the distributed stochastic optimization problem

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{minimize}} \quad F(\boldsymbol{\theta}) \triangleq \frac{1}{N} \sum_{i=1}^{N} f_i\left(\boldsymbol{\theta}\right), \tag{1}$$

where $F(\boldsymbol{\theta})$ is the global loss function and $f_i(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} f_{i,\xi}(\boldsymbol{\theta})$ is the local loss function at device i. In particular, $f_{i,\xi}(\boldsymbol{\theta})$ is the loss for the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ on a data sample ξ at device i. In the distributed learning process, each device $i \in \mathcal{N}$ has a local parameter vector $\boldsymbol{\theta}_i^t$ that approximates the solution of problem (1) at the t-th iteration.

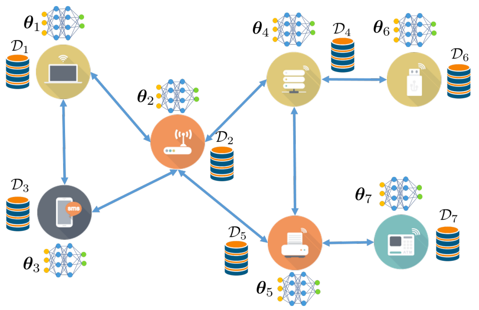

Fig. 1: Illustration of the decentralized FL model over wireless networks consisting of 7 devices.

B. DSGD with Gradient Tracking and Variance Reduction

When the devices communicate over error-free orthogonal links, the commonly adopted method for solving problem (1) is DSGD [26]. We denote $\boldsymbol{W}$ as the mixing matrix encoding the network structure at iteration t.

Definition 1 (Mixing matrix)**.** *Mixing matrix $\boldsymbol{W}$ is a symmetric doubly stochastic matrix whose ij-th entry $w_{ij} > 0$ indicates that device i and j are connected,*

$$\boldsymbol{W} \in \mathcal{W}, \boldsymbol{W} = \boldsymbol{W}^T, \boldsymbol{W}\mathbf{1} = \mathbf{1}, \mathbf{1}^T\boldsymbol{W} = \mathbf{1}^T, \tag{2}$$

where $\mathcal{W} = \{\boldsymbol{W} \in [0,1]^{N \times N} | w_{ij} = 0 \text{ if} (i,j) \notin \mathcal{E} \text{ and } i \neq j\}$.

With the DSGD algorithm, the information exchange can only occur between connected devices. The algorithm for each device $i \in \mathcal{N}$ at iteration t consists of two phases:

1) Computation phase: each device computes a stochastic gradient based on a random data sample ξ_i^t and then performs stochastic gradient updates,

$$\boldsymbol{\theta}_i^{t+\frac{1}{2}} = \boldsymbol{\theta}_i^t - \alpha_t \nabla f_{i,\xi_i^t}(\boldsymbol{\theta}_i^t). \tag{3}$$

2) Consensus phase: devices average their local model based on mixing matrix,

$$\boldsymbol{\theta}_i^{t+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{\theta}_j^{t+\frac{1}{2}}. \tag{4}$$

However, DSGD decays linearly with a fixed step size $\alpha_t \equiv \alpha$, but performs inexact convergence (the steady error of DSGD has an additional bias). Although properly reducing α_t enables exact convergence (each $\boldsymbol{\theta}_n^t$ converges to the same exact solution), DSGD convergences slowly both in practice and theory [27]. With the help of dynamic average-consensus, the DSGD with gradient tracking (DSGT) removes the bias which characterizes the difference between local loss function f_i in DSGD, thereby achieving a much lower steady-state error when the data sample across devices are largely heterogeneous [28]. And by adding variance reduction to DSGT, [27] shows that the DSGT with variance reduction (DSGT-VR) leads to an exact linear convergence with a constant step-size.

In DSGT-VR, each device i maintains a gradient table $\{\nabla f_{i,j}(\hat{\boldsymbol{\theta}}_{i,j})\}_{j=1}^{|\mathcal{D}_i|}$, where $\hat{\boldsymbol{\theta}}_{i,j}$ denotes the most recent parameter when $f_{i,j}$ was computed, to store all local gradients. Then at iteration t, each device i randomly chooses a data sample ξ_i^t and computes the unbiased gradient $\boldsymbol{g}_i^t$ as follows

$$\boldsymbol{g}_i^t = \nabla f_{i,\xi_i^t}(\boldsymbol{\theta}_i^t) - \nabla f_{i,\xi_i^t}(\hat{\boldsymbol{\theta}}_{i,\xi_i^t}) + \frac{1}{|\mathcal{D}_i|}\sum_{j=1}^{|\mathcal{D}_i|} \nabla f_{i,j}(\hat{\boldsymbol{\theta}}_{i,j}). \quad (5)$$

Next, we replace the recent gradient $\nabla f_{i,\xi_i^t}(\boldsymbol{\theta}_i^t)$ with $\nabla f_{i,\xi_i^t}(\hat{\boldsymbol{\theta}}_{i,\xi_i^t})$ in gradient table. Last, the gradient estimator $\boldsymbol{d}_i^{t+1}$ can be computed by using gradient tracking technology,

$$\boldsymbol{d}_i^{t+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{d}_i^t + \boldsymbol{g}_i^{t+1} - \boldsymbol{g}_i^t. \quad (6)$$

C. Communication Model

In this section, we focus on the design of communication model for AirComp-based DSGT-VR algorithm in decentralized setting. The designed communication model consists of the following two parts:

- **Scheduling**: To cope with wireless interference, different devices should be scheduled to communicate in different transmission blocks.
- **Transmission**: The transmission strategies are designed to support consensus phase over wireless networks.

1) Scheduling: To improve the communication performance, we consider the D2D communication [29] in this paper. Since the precoding strategy (8) is designed for target device to guarantee the convergence of proposed algorithm, it brings interference to other devices. So we focus on design interference-free scheduling by scheduling enrolled receiving devices as active "PS" in different transmission block. In this interference-free schedule scheme, there are no two enrolled receiving devices connected to the same device. A naive scheduling policy is to schedule to the enrolled receiving devices one by one, which require N transmission blocks during one consensus phase. The graph coloring algorithm can be adopted to find a suitable scheduling policy that decreases the number of transmission blocks [30]. The authors in [22] has investigated the effect of scheduling policy. In this paper we focus on the convergence versus the consensus iteration t over wireless networks.

2) Transmission: To enhance the spectral efficiency, AirComp has been envisioned to have a wide range of applications in the areas of consensus [31]. In AirComp, the receiver device receives a superposition of the signals that simultaneously transmitted by its neighbor [32], [33]. Block flat-fading channel are considered in this paper and one transmission block contains d time slots. Hence, at the transmission block t, the received signal at enrolled receiving device i can be written as

$$\boldsymbol{y}_i^t = \sum_{j\in\mathcal{N}_i} h_{ij}^t \boldsymbol{x}_j^t + \boldsymbol{z}_i^t, \quad (7)$$

where $h_{ij}^t \in \mathbb{C}$ is the channel coefficient between devices i and j at transmission block t, the transmit signal $\boldsymbol{x}_j^t$ encodes the information of the local model $\boldsymbol{\theta}_j^t$ and $\boldsymbol{z}_i^t \in \mathbb{C}^d$ represents the additive noise vector. In addition, channel inputs are subject to a peak power constraint, i.e., $\mathbb{E}[\|\boldsymbol{x}_i^t\|_2^2] \le P, \forall i \in \mathcal{N}$.

We assume the computation phase at each device to be instantaneous after transmission. Based on the received signal $\boldsymbol{y}_i^t$, device i can average its neighbor's local model and get close to the solution of problem (1). However, due to the influence of channel fading and transmission noise, the received signal is distorted, which will negatively influence the convergence of DSGT-VR algorithms. Hence, we aim to develop a reliable transmission strategy based on DSGT-VR to support FL in decentralized setting.

D. AirComp-based DSGT-VR algortihm

First, we develop precoding and decoding strategies at devices to guarantee the convergence of AirComp-based DSGT-VR. Each consensus iteration k includes $R \le N$ transmission blocks and the devices are scheduled as active "PS" in one transmission block.

In the precoding phase before AirComp, we assume that perfect CSI are available at all devices. Since the topology of the communication network is known, each device contains the mixing matrix weights among its connected neighbors, i.e., $w_{ij}, \forall j \in \mathcal{N}_i$. In transmission block t, the signal $\boldsymbol{x}_j^t, j \in \mathcal{N}_i$ transmitted to device i is precoded as

$$\boldsymbol{x}_j^t = \sqrt{p^t} w_{ij} \frac{(h_{ij}^t)^H}{|h_{ij}^t|^2} \boldsymbol{\theta}_j^t, \quad (8)$$

where $\sqrt{p^t} = \min_i \frac{|h_{ij}^t|\sqrt{P}}{\|\boldsymbol{\theta}_i^t\|_2}, \forall j$ is a uniform scaling factor to satisfy the peak power constraint. However, in decentralized setting, it is observed that the weak channels lead to the small transmit power. In this case, the influence of channel noise will significantly increase. Therefore, we form the connectivity graph $\mathcal{G}$ by connecting the devices i and j if the channel gain $|h_{ij}|$ is above a certain threshold γ [34]–[36]. Then, for device i, we get all its connected neighbors $\mathcal{N}_i$, which is given by

$$\mathcal{N}_i = \{j | (i,j) \in \mathcal{E}, \text{if } |h_{ij}| > \gamma\}. \quad (9)$$

Second, in the decoding phase and based on (8), the received signal at device i after AirComp can be written as

$$\boldsymbol{y}_i^t = \sum_{j\in\mathcal{N}_i\setminus\{i\}} \sqrt{p^t} w_{ij} \boldsymbol{\theta}_j^t + \boldsymbol{z}_i^t. \quad (10)$$

Then the device i can decode the local model signal $\boldsymbol{y}_i^t$, by

$$\boldsymbol{\theta}_i^{t+1} = \frac{1}{\sqrt{p^t}}\boldsymbol{y}_i^t + w_{ii}\boldsymbol{\theta}_i^t = \sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{\theta}_j^t + \widetilde{\boldsymbol{z}}_i^t, \quad (11)$$

where $\widetilde{\boldsymbol{z}}_i^t = \frac{\boldsymbol{z}_i^t}{\sqrt{p^t}} \sim \mathcal{CN}(0, \frac{\sigma^2}{p^t})$ is the channel noise. Since equation (11) is a superposition of the local model parameter, we apply AirComp to perform the consensus phases in DSGT-VR algorithm in wireless communication.

The AirComp techniques operate over star graphs in a pair of consecutive phases of time slots [18], [19]. However, in decentralized setting, since there is no central PS, each device can play a role of central PS through AirComp over star graphs [25]. Specially, in the first phase of time slot, the device i at the center of a star sub-graph receive the signals

Algorithm 1: DSGT-VR with Over-the-Air Consensus

```
Input : Initial: $\boldsymbol{\theta}_i^0 = \boldsymbol{\theta}^0, \boldsymbol{d}_i^0 = \boldsymbol{g}_i^0 = \nabla f_i(\boldsymbol{\theta}_i^0), i \in \mathcal{N}$,
        step sizes $\alpha$, max iterations $T$, $\boldsymbol{W}$ and
        gradient table $\{\nabla f_{i,j}(\hat{\boldsymbol{\theta}}_{i,j})\}_{j=1}^{|\mathcal{D}_i|}, \hat{\boldsymbol{\theta}}_{i,j} = \boldsymbol{\theta}_i^0, \forall j$
for $t = 1, 2, \ldots, T$ do
    for $r = 1, 2, \ldots, M$ transmission block do
        for each $i \in \mathcal{N}$ do in parallel
            Updating $\boldsymbol{\theta}_i^{t+\frac{1}{2}} = \boldsymbol{\theta}_i^t - \alpha \boldsymbol{d}_i^t$.
            Get a data sample $\xi_n^{t+1}$ randomly, compute
             unbiased gradient (5).
            Replace $\nabla f_{i,\xi_i^t}(\hat{\boldsymbol{\theta}}_{i,\xi_i^t})$ by $\nabla f_{i,\xi_i^t}(\boldsymbol{\theta}_i^t)$ in
             gradient table.
            Updating $\boldsymbol{d}_i^{t+\frac{1}{2}} = \boldsymbol{d}_i^t + \boldsymbol{g}_i^t - \boldsymbol{g}_i^{t-1}$.
            if device $i$ is scheduled as active "PS" then
                The device $i$ received signal via (10)
                 and decoding via (11) to get $\boldsymbol{\theta}_i^{t+1}$.
                Broadcast $\boldsymbol{d}_i^{t+\frac{1}{2}}$ back to all its
                 neighbors.
            else
                Precoding $\boldsymbol{\theta}_i^{t+\frac{1}{2}}$ via (8) and send the
                 precoded signal to all its neighbors $\mathcal{N}_i$.
            end
        end
    end
    Each device locally updates gradient eatimator
     $\boldsymbol{d}_i^{t+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{d}_i^{t+\frac{1}{2}}$.
end
```

simultaneously transmitted by all its connected neighbors in $\mathcal{N}_n$ with a superposition form (10). In the second phase of time slot, this central device i broadcasts its update gradient estimator (6) to all its connected neighbors in $\mathcal{N}_n$. In this paper, the devices broadcast the gradient estimator rather than model parameter [22], [23] in downlink communication. The complete implementation of AirComp-based DSGT-VR is summarized in Algorithm 1. Note that gradient estimator of each device is received in different transmission block via error-free link due to scheduling, then we can update gradient estimator based on (6) after M transmission blocks. And we leave the variance reduction techniques over noisy links for future work.

III. Convergence Analysis

In this section, we provide the convergence analysis of the AirComp-based DSGT-VR. The convergence results are established under the following assumptions.

Assumption 1. *The local cost function f_i is both μ-strongly convex and L-smooth for any $i \in \mathcal{N}$.*

Assumption 2. *The network graph $\mathcal{G}$ is undirected and connected, i.e., there exists a path between any two devices.*

Assumption 3. *The local parameter is bounded by a universal constant $B \geq 0$, i.e., $\|\boldsymbol{\theta}_i^t\|^2 \leq B^2, \forall i, t$.*

Assumption 2 implies that $\beta = |||\boldsymbol{W} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top||| < 1$ [37]. Based on above assumptions, we denote $M = \max_i |\mathcal{D}_i|, m = \min_i |\mathcal{D}_i|, \forall i \in \mathcal{N}$ and $\kappa = \frac{L}{\mu}$ be the condition number of the global loss function F. The main convergence result of AirComp-based DSGT-VR is established as follows,

Theorem 1. *Let $\boldsymbol{\theta}^\natural$ denote the solution of the distributed optimization problem (1). If Assumptions 1 and 2 hold and the step-size in AirComp-based DSGT-VR algorithm is such that*

$$\alpha = \min\{\mathcal{O}(\frac{1}{\mu M}), \mathcal{O}(\frac{m(1-\beta)^2}{ML\kappa^2})\},$$

then $\forall t \geq 0, \forall i \in \mathcal{N}$ and for some $0 < \rho < 1$, it holds that

$$\mathbb{E}\left[F(\boldsymbol{\theta}_i^t)\right] - F(\boldsymbol{\theta}^\natural) \leq \frac{cL}{2}\rho^t + \frac{LN}{2(N-1)}\frac{d\sigma^2 B^2}{\gamma^2 P}\sum_{\tau=1}^{t}\rho^{t-\tau}, \tag{12}$$

where $c = \frac{N}{N-1}\|\boldsymbol{\theta}_i^0 - \overline{\boldsymbol{\theta}}^0\|^2 + N\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^\natural\|^2$.

Proof. See Appendix A. □

Theorem 1 establishes the upper bound of estimation error for strongly convex and smooth local loss function in decentralized federated learning system. The error bound is divided into the initial distance due to the error in original algorithm and the additive noise caused by the channel noise. In addition, the initial distance shows that the AirComp-based DSGT-VR achieves the same convergence rate as that of the DSGT-VR algorithm, i.e., linear convergence rate. Theorem 1 demonstrates that impact of the additive noise decreases as the number of devices increases.

IV. Numerical Experiments

In this section, we numerically demonstrate the convergence behavior of the proposed AirComp-based DSGT-VR, and compare with the-state-of-art decentralized FL algorithms. We consider a decentralized federated learning setting where $N = 20$ devices with 1000 data samples locally at each device cooperatively train a regularized logistic regression model for binary classification,

$$F(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\log[1 + e^{-(\boldsymbol{a}_{ij}^T\boldsymbol{\theta})b_{ij}}] + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2, \tag{13}$$

where device i holds m_i training samples $\{\boldsymbol{a}_{ij}, b_{ij}\}_{j=1}^{m_i}$, $\boldsymbol{a}_{ij}$ denotes the feature of j-th training sample at i-th device, $b_{ij} \in \{+1, -1\}$ is the corresponding binary label and λ is a regularization parameter.

For data sample, we use classes "3" and "5" in the MNIST dataset for binary classification. We use 1000 samples for training and 1968 samples for testing, and each device is assigned by 40 training samples. We use the channel $h_{ij}^t \sim \mathcal{CN}(0, 1)$, noise σ^2 to be 0 dBm and the threshold $\gamma = 0.5$. In this paper, we adopted Laplacian matrix [38] as mixing matrix for fast and provable convergence analysis. To stable the training process, all features are normalized to be unit vectors,

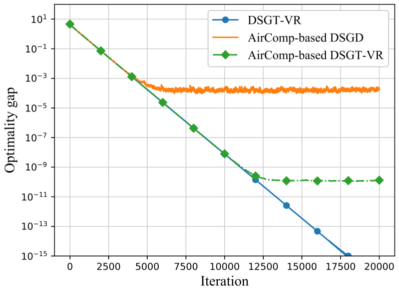

Fig. 2: Optimality gap versus number of consensus iteration t.

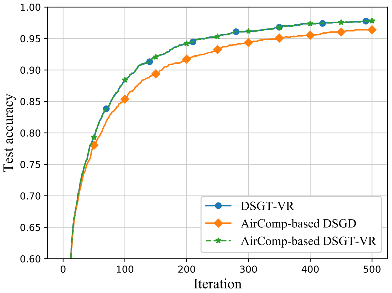

Fig. 3: Test accuracy versus number of consensus iteration t.

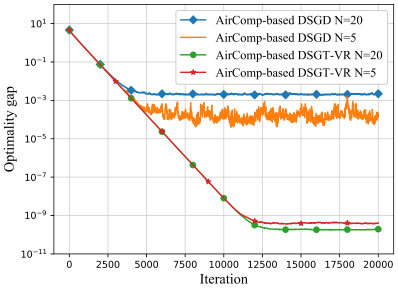

Fig. 4: Optimality gap versus number of consensus iteration t with different N.

i.e., $\|\boldsymbol{a}_{ij}\|_2 = 1, \forall i, j$, and set the regularization parameter $\lambda = \frac{1}{\sum_{i=1}^N m_i}$. And we characterize the performance in terms of the optimality gap, i.e., $\mathbb{E}\left[F(\boldsymbol{\theta}_i^t)\right] - F(\boldsymbol{\theta}^\natural)$. To further verify the effectiveness of the proposed, we compare the proposed algorithm with the AirComp-based DSGD algorithm [22], [23], which is the state-of-the-art algorithm of the wireless decentralized federated edge learning system.

Fig. 2 shows that both AirComp-based DSGD and the proposed algorithm achieve a linear convergence rate, but the proposed algorithm can converge to a more accurate solution. Specially, the error gap of AirComp-based DSGD almost reaches 10^{-4}, while the proposed algorithm only has the optimality gap of 10^{-10}. However, due to the effect by fading channels and additive noise, there is still a gap between the solution obtained by proposed algorithm and optimal solution, which corresponds to our theoretical results. And the test accuracy is evaluated versus the number of consensus iteration is illustrated in Fig. 3. Obviously, the proposed algorithm outperforms the AirComp-based DSGD algorithm while reaches the nearly the same accuracy of DSGT-VR algorithm.

In Fig 4, we illustrate the impact of the different number of devices on optimality gap of the proposed algorithm in the same network topology. Clearly, the proposed algorithm with 20 devices reaches a more accurate solution than that with 5 devices, which verifies our theoretical results. However, the AirComp-based DSGD has the opposite result that the more devices the less accurate solution.

V. Conclusions

In this paper, we proposed a novel AirComp-based DSGT-VR algorithm to facilitate the consensus phase in decentralized FL over wireless networks, where both precoding and decoding strategies at devices are developed for D2D communications. Furthermore, we proved the linear convergence rate of the proposed algorithm and provided the error bound to reveal the impact of fading channel and its noise. Simulation results were conducted to verify the theoretical result and the superior performance of the proposed algorithm.

Appendix A
Proof of Theorem 1

Under assumption 1, the global cost F is also L-smooth and we denote $\boldsymbol{\theta}^\natural$ is the optimal solution of F, then by smoothness

$$\mathbb{E}\left[F(\boldsymbol{\theta}_i^t)\right] - F(\boldsymbol{\theta}^\natural) \le \frac{L}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^\natural\|^2\right]. \tag{14}$$

First, we introduce the auxiliary variable $\{\overline{\boldsymbol{\theta}}^t\}$ which is mean of $\boldsymbol{\theta}_1^t, \cdots, \boldsymbol{\theta}_N^t$ in error-free case, then the error $\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^\natural\|^2\right], \forall i \in \mathcal{N}$ turn out to be

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^\natural\|^2\right] &= \mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t + \overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^\natural\|^2\right]\\ &\overset{(a)}{\le} \underbrace{\frac{N}{N-1}\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|^2\right]}_{T_1} + \underbrace{N\mathbb{E}\left[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^\natural\|^2\right]}_{T_2},\end{aligned} \tag{15}$$

where (a) comes from Young's inequality that $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \le (1+\eta)\|\boldsymbol{a}\|^2 + (1+\frac{1}{\eta})\|\boldsymbol{b}\|^2, \forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d, \forall \eta \ge 0$ and we set $\eta = N-1$. Then the error $\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^\natural\|^2\right]$ is divided into two parts, i.e., T_1 and T_2. Next, we focus on establish the upper bounds for T_1 and T_2. Recall (11), we have $\boldsymbol{\theta}_i^t = \sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{\theta}_j^{t-1} + \tilde{\boldsymbol{z}}_i^{t-1}$, then we can rewrite $\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|^2\right]$ in T_1 as follows

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|^2\right] &\overset{(a)}{=} \mathbb{E}\left[\|\sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{\theta}_j^{t-1} - \overline{\boldsymbol{\theta}}^t\|^2\right] + \mathbb{E}\left[\|\tilde{\boldsymbol{z}}_i^{t-1}\|^2\right],\\ &\overset{(b)}{\le} \rho\mathbb{E}\left[\|\boldsymbol{\theta}_i^{t-1} - \overline{\boldsymbol{\theta}}^{t-1}\|^2\right] + \frac{d\sigma^2 B^2}{\gamma^2 P},\end{aligned}$$

where (a) comes from that noise $\tilde{\boldsymbol{z}}_i^{t-1}$ is zero-mean and independent to model parameters and (b) comes from [39, Theorem 1, Lemma 2] with $0 < \rho < 1$ and the uniform scaling factor $\sqrt{p^t}$, i.e., $\mathbb{E}\left[\|\tilde{\boldsymbol{z}}_i^{t-1}\|^2\right] = \frac{d\sigma^2}{p^t} \le \frac{d\sigma^2 B^2}{\gamma^2 P}$. According to [37, Propositon 1], then if $\alpha = \min\{\mathcal{O}(\frac{1}{\mu M}), \mathcal{O}(\frac{m(1-\beta)^2}{ML\kappa^2})\}$, we have $T_2 \le \rho\mathbb{E}\left[N\|\overline{\boldsymbol{\theta}}^{t-1} - \boldsymbol{\theta}^\natural\|^2\right]$. Hence, we get

$$\mathbb{E}\left[\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^\natural\|^2\right] \le c\rho^t + \frac{N}{N-1}\frac{d\sigma^2 B^2}{\gamma^2 P}\sum_{\tau=1}^{t}\rho^{t-\tau},$$

where $c = \frac{N}{N-1}\|\boldsymbol{\theta}_i^0 - \overline{\boldsymbol{\theta}}^0\|^2 + N\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^\natural\|^2$ and $0 < \rho < 1$. Based on (14), we can obtain

$$\mathbb{E}\left[F(\boldsymbol{\theta}_i^t)\right] - F(\boldsymbol{\theta}^\natural) \le \frac{cL}{2}\rho^t + \frac{LN}{2(N-1)}\frac{d\sigma^2 B^2}{\gamma^2 P}\sum_{\tau=1}^{t}\rho^{t-\tau}, \tag{16}$$

which completes the proof.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. and Statist. (AISTATS)*. PMLR, 2017, pp. 1273–1282.

[2] S. A. Rahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things J.*, Apr. 2021.

[3] 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, Oct. 2020.

[4] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," *IEEE Trans. Industr. Inform.*, vol. 16, no. 6, pp. 4177–4186, Jun. 2020.

[5] Q. V. Pham, K. Dev, P. K. R. Maddikunta, T. R. Gadekallu, T. Huynh-The, et al., "Fusion of federated learning and industrial internet of things: A survey," *arXiv preprint arXiv:2101.00798*, 2021.

[6] S. Savazzi, M. Nicoli, M. Bennis, S. Kianoush, and L. Barbieri, "Opportunities of federated learning in connected, cooperative, and automated industrial systems," *IEEE Commun. Mag.*, Mar. 2021.

[7] Z. Yang, M. Chen, K. Wong, H. Poor, and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," *arXiv preprint arXiv:2101.01338*, 2021.

[8] 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, Apr. 2020.

[9] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, Jul. 2020.

[10] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[11] 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, Oct. 2020.

[12] L. Li, L. Yang, X. Guo, Y. Shi, H. Wang, W. Chen, and K. B. Letaief, "Delay analysis of wireless federated learning based on saddle point approximation and large deviation theory," *arXiv preprint arXiv:2103.16994*, 2021.

[13] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Lataief, "Federated learning via intelligent reflecting surface," *arXiv preprint arXiv:2011.05051*, 2020.

[14] M. Chen, H. Vincent Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, Apr. 2021.

[15] 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.

[16] D. Wen, M. Bennis, and K. Huang, "Joint parameter-and-bandwidth allocation for improving the efficiency of partitioned edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8272–8286, Dec. 2020.

[17] 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.

[18] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, Apr. 2020.

[19] 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.

[20] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5330–5340.

[21] X. Cai, X. Mo, J. Chen, and J. Xu, "D2D-enabled data sharing for distributed machine learning at wireless network edge," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1457–1461, Sept. 2020.

[22] E. Ozfatura, Stefano Rini, and D. Gündüz, "Decentralized SGD with over-the-air computation," in *Proc. IEEE Global Commun. Conf. (Globecom)*, 2020, pp. 1–6.

[23] H. Xing, O. Simeone, and S. Bi, "Decentralized federated learning via SGD over wireless D2D networks," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2020, pp. 1–5.

[24] S. Savazzi, S. Kianoush, V. Rampa, and M. Bennis, "A joint decentralized federated learning and communications framework for industrial networks," in *Proc. IEEE Int. Workshop Comput.-Aided Model., Anal., and Dess. Commun. Links and Netw. (CAMAD)*, 2020, pp. 1–7.

[25] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[26] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. U. Stich, "A unified theory of decentralized sgd with changing topology and local updates," *arXiv preprint arXiv:2003.10422*, 2020.

[27] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 102–113, May 2020.

[28] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Math. Program.*, pp. 1–49, 2020.

[29] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, May 2014.

[30] M. S. Molloy, M. Molloy, and B. Reed, *Graph colouring and the probabilistic method*, vol. 23, Springer Science & Business Media, 2002.

[31] G. Zhu, J. Xu, and K. Huang, "Over-the-air computing for 6G–turning air into a computer," *arXiv preprint arXiv:2009.02181*, 2020.

[32] Z. Wang, Y. Shi, Y. Zhou, H. Zhou, and N. Zhang, "Wireless-powered over-the-air computation in intelligent reflecting surface-aided IoT networks," *IEEE Internet of Things J.*, vol. 8, no. 3, pp. 1585–1598, Feb. 2021.

[33] J. Dong, Y. Shi, and Z. Ding, "Blind over-the-air computation and data fusion via provable wirtinger flow," *IEEE Trans. Signal Process.*, vol. 68, pp. 1136–1151, Jan. 2020.

[34] L. Ruan and V. K. N. Lau, "Dynamic interference mitigation for generalized partially connected quasi-static MIMO interference channel," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3788–3798, Aug. 2011.

[35] L. Ruan, V. K. N. Lau, and X. Rao, "Interference alignment for partially connected MIMO cellular networks," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3692–3701, Jul. 2012.

[36] S. A. Jafar, "Topological interference management through index coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 529–568, Oct. 2014.

[37] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *IEEE Trans. Signal Process.*, vol. 68, pp. 6255–6271, Oct. 2020.

[38] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[39] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control. Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, Sept. 2018.