Federated Learning over Wireless Device-to-Device Networks: Algorithms and Convergence Analysis

Hong Xing, Osvaldo Simeone, and Suzhi Bi

Abstract

The proliferation of Internet-of-Things (IoT) devices and cloud-computing applications over siloed data centers is motivating renewed interest in the collaborative training of a shared model by multiple individual clients via federated learning (FL). To improve the communication efficiency of FL implementations in wireless systems, recent works have proposed compression and dimension reduction mechanisms, along with digital and analog transmission schemes that account for channel noise, fading, and interference. The prior art has mainly focused on star topologies consisting of distributed clients and a central server. In contrast, this paper studies FL over wireless device-to-device (D2D) networks by providing theoretical insights into the performance of digital and analog implementations of decentralized stochastic gradient descent (DSGD). First, we introduce generic digital and analog wireless implementations of communication-efficient DSGD algorithms, leveraging random linear coding (RLC) for compression and over-the-air computation (AirComp) for simultaneous analog transmissions. Next, under the assumptions of convexity and connectivity, we provide convergence bounds for both implementations. The results demonstrate the dependence of the optimality gap on the connectivity and on the signal-to-noise ratio (SNR) levels in the network. The analysis is corroborated by experiments on an image-classification task.

Index Terms

Federated learning, distributed learning, decentralized stochastic gradient descent, over-the-air computation, D2D networks.

Part of this paper has been presented at the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), May 2020 [1].

H. Xing and S. Bi are with the College of Electronic and Information Engineering, Shenzhen University, Shenzhen, 518060, China (e-mails: {hong.xing, bsz}@szu.edu.cn).

O. Simeone is with the King's Communications, Learning, and Information Processing (KCLIP) lab, Department of Engineering, King's College London, London, WC2R 2LS, U.K. (e-mail: osvaldo.simeone@kcl.ac.uk).

I. Introduction

With the proliferation of Internet-of-Things (IoT) devices and cloud-computing applications over siloed data centres, *distributed learning* has become a critical enabler for artificial intelligence (AI) solutions [2], [3]. In distributed learning, multiple agents collaboratively train a machine learning model via the exchange of training data, model parameters and/or gradient vectors over geographically distributed computing resources and data. *Federated learning (FL)* refers to distributed learning protocols that do not directly exchange the training data in an attempt to reduce the communication load and to limit privacy concerns [4]–[7]. In conventional FL, multiple clients train a shared model by exchanging model-related parameters with a central node. This class of protocols hence relies on a parameter-server (PS) architecture, which is typically realized in wireless settings via a base station-centric network topology [8]–[12].

There are important scenarios when there is no central coordinator acting as the parameter server (PS) for FL, and therefore distributed learning must rely on a peer-to-peer communication topology that encompasses device-to-device (D2D) links among individual learning agents over an arbitrary connectivity graph. For example, reference [13] demonstrated the effectiveness of a D2D FL framework across siloed medical centers. There are also other scenarios when PS-based architecture is undesirable due to coverage, privacy, implementation efficiency, or fault-tolerance considerations, and thus a D2D architecture becomes a preferred option thanks to its resilience and increased parallelism. An example is the ring AllReduce architecture for deep learning (DL) described in [14] that enables efficient GPU-and-GPU communications.

A. Related Work

The problem of alleviating the communication load in FL systems has been widely investigated, mostly under the assumption of noiseless, rate-limited links, and *star topologies*. Key elements of these solutions are compression and dimension-reduction operations that map the original model parameters or gradient vectors into representations defined by a limited number of bits and/or sparsity. Important classes of solutions include unbiased compressors (e.g., [8]–[10]) and biased compressors with error-feedback mechanisms (e.g., [11], [12], [15]).

In a *D2D architecture*, devices can only exchange information with their respective neighbors, making consensus mechanisms essential to ensure agreement towards the common learning goal [16]. A well-known protocol integrating stochastic gradient (SGD) and consensus is *Decentralized Stochastic Gradient Descent (DSGD)* [17], [18]. Similar to FL in star topologies, there have

been lots of previous efforts aiming for accelerating consensus and removing communication overhead of the DSGD from algorithmic perspectives by, e.g., variance-reduction for large data heterogeneity among agents [19] and by compression [20]–[23]. The CHOCO-SGD algorithm proposed in [20]–[22], which combines the standard DSGD algorithm with biased compression, was studied for strongly convex and smooth objectives in [20], and for non-convex smooth objectives in [21], combined with event-triggered protocols in [22]. The authors in [23] introduced a compression scheme for model exchange between neighboring nodes that improves convergence while requiring no additional hyperparameters. All these prior works on decentralized FL assume either ideal or rate-limited *noiseless* D2D communications. Wireless implementations of these algorithms are non-trivial in general, because the goal of achieving consensus is compromised to different extents by channel impairments, such as fading, or packet losses.

A large number of recent works have proposed communication strategies and multi-access protocols for *FL in wireless star topologies* [24]–[27]. At the physical layer, *over-the-air computation (AirComp)* was investigated in [15], [24], [25], [28], [29] as a promising solution to support simultaneous transmissions by leveraging the waveform superposition property of the wireless medium. Unlike conventional digital communication over orthogonal transmission blocks, AirComp is based on analog, e.g., uncoded, transmission, which enables the estimate of aggregated statistics directly from the received baseband samples. This reduces the communication burden, relieving the network from the need to decode individual information separately for all participating devices. For example, the authors in [29] proposed an adaptive learning-rate scheduler and investigated the convergence of the resulting protocol.

The literature on *decentralized FL in wireless D2D architecture* is, in contrast, still quite limited. A DSGD based algorithm termed MATCHA was proposed in [30] by accounting for interference among nearby links. By sampling a matching decomposition of an interference graph, MATCHA schedules non-interfering communication links in parallel, among which the connectivity-critical links are activated with a higher probability. However, no attempt was made to take physical layer transmissions into account. A real-time implementation of decentralized FL systems was proposed in [31] over industrial wireless networks, and the joint effects of model pruning, sparisification, and quantization were considered. The wireless transmission constraints under study were only suitable for digital transmissions. References [32] and [33] both considered AirComp-based DSGD for decentralized FL over D2D networks, where precoding/decoding strategies for analog D2D transmissions, possibly combined with interference-free

D2D scheduling policy, were developed. Compared with [32], the work [33] proved that the original performance of DSGD with gradient tracking is compromised by an error floor term due to the channel noise introduced by AirComp, while relying on the strong assumption that the intermediate information broadcast by the center of a "star" sub-graph can be ideally received by its neighbors. Furthermore, reference [34] provided a theoretical analysis for the analog implementation of another fully decentralized optimization algorithm, *decentralized lazy mirror decent (DLMD)*, which accommodates convex non-smooth loss functions under channel noise and rate constraints. However, since the convergence rate is conditioned on an increasing power sequence over training iterations, the performance guarantee may be seriously compromised in power-constrained wireless D2D networks. In addition, how the proposed analog implementation of the DLMD scheme addresses the mismatch between the available channel uses and the typically larger dimension of the model has yet been investigated.

B. Main Contributions

This work investigates the impact of wireless communication constraints induced by blockages, pathloss, channel noise and fading on the convergence of DSGD-based FL algorithms for digital and analog transmission implementations. The analysis applies to general scheduling and power allocation policies, with the CHOCO-SGD chosen as the baseline DSGD algorithm due to its generality and flexibility in the choice of the compression operators. The main contributions are as follows:

1) We study general digital and analog wireless implementations of DSGD algorithms that rely on dimension reduction-based compression via *random linear coding (RLC)* and enable broadcasting for digital transmission, as well as both broadcasting and AirComp for analog transmission.

2) Under the assumptions of convexity and connectivity, we derive *convergence bounds* for the mentioned general class of digital wireless implementations that demonstrate the dependence of the optimality gap on the connectivity of the graph and on the model estimation error due to compression.

3) We also provide *convergence bounds* for the mentioned general class of analog wireless implementations of DSGD that quantify the impact of topology and channel noise. The analysis also reveals the role played by an *adaptive* consensus step size in combating the effect of accumulative channel noise in the analog implementation. To the best of our knowledge, this is the first time that an adaptive consensus step size is shown to be beneficial for convergence.

4) We provide numerical experiments for image classification, confirming the benefits of the proposed adaptive consensus rate, and demonstrating the agreement between analytical and empirical results.

The conference version [1] of this paper provided some preliminary experimental results on digital and analog DSGD, without offering any theoretical results.

The remainder of this paper is organized as follows. The system model is presented in Section II. Digital and analog transmission protocols are introduced in Section III. The convergence analysis for both implementations is presented in Section IV and Section V, respectively. Numerical performance results are described in Section VI, followed by conclusions in Section VII.

C. Notations

We use the upper case boldface letters for matrices and lower case boldface letters for vectors. We also use $\|\cdot\|$ to denote the Euclidean norm of a vector or the spectral norm of a matrix, and $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. We denote by $|\mathcal{V}|$ the cardinality of a set $\mathcal{V}$. The average of vectors $\boldsymbol{x}_i$ over $i \in \mathcal{V}$ is defined as $\bar{\boldsymbol{x}} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \boldsymbol{x}_i$. Notations $\mathsf{Tr}(\cdot)$ and $(\cdot)^T$ denote the trace and the transpose of a matrix, respectively. $\mathbb{E}[\cdot]$ stands for the statistical expectation of a random variable. $\boldsymbol{I}$ represents an identity matrix with appropriate size, and $\triangleq$ indicates a mathematical definition. $\lambda_i(\cdot)$ denotes the ith largest eigenvalue of a matrix.

II. System Model

In this paper, we consider a FL problem in a decentralized setting as shown in Fig. 1, in which a set $\mathcal{V} = \{1, \ldots, K\}$ of K devices can only communicate with their respective neighbors over a wireless D2D network whose connectivity is characterized by an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ denoting the set of nodes and $\mathcal{E} \subseteq \{(i,j) \in \mathcal{V} \times \mathcal{V} \mid i \neq j\}$ the set of edges. The set of neighbors of node i is denoted as $\mathcal{N}_i = \{j \in \mathcal{V} \mid (i,j) \in \mathcal{E}\}$. Following the FL framework, each device has available a local data set, and all devices collaboratively train a machine learning model by exchanging model-related information without directly disclosing data samples to one another.

A. Learning Model

Each device $i \in \mathcal{V}$ has access to its local data set $\mathcal{D}_i$, which may have non-empty intersection with the data set $\mathcal{D}_j$ of any other device $j \in \mathcal{V}$, $i \neq j$. All devices share a common machine

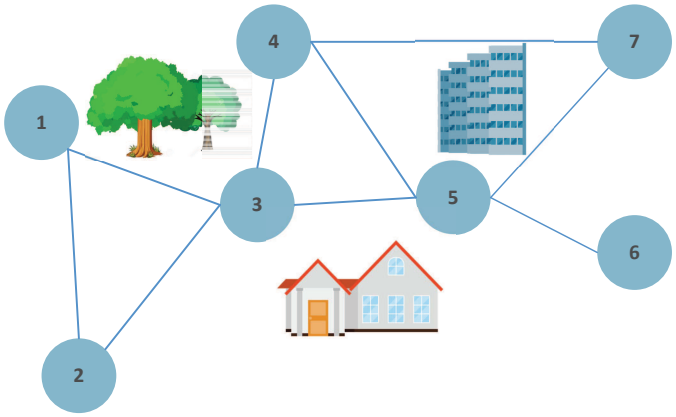


Fig. 1. The connectivity graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ for a wireless D2D network.

learning model class, which is parametrized by a vector $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$. As a typical example, the model class may consist of a neural network with a given architecture. The goal of the network is to solve the empirical risk minimization problem [17], [21]

$$(\text{P0}): \underset{\boldsymbol{\theta}}{\texttt{Minimize}} \quad F(\boldsymbol{\theta}) \triangleq \frac{1}{K}\sum_{i\in\mathcal{V}} f_i(\boldsymbol{\theta}),$$

where $f_i(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_i|}\sum_{\boldsymbol{\xi}\in\mathcal{D}_i} l(\boldsymbol{\theta}, \boldsymbol{\xi})$ is the local empirical risk function for device i with $l(\boldsymbol{\theta}; \boldsymbol{\xi})$ denoting the loss accruing from parameter $\boldsymbol{\theta}$ on data sample $\boldsymbol{\xi} \in \mathcal{D}_i$, which may include the effect of regularization.

Among the communication-efficient variants of the DSGD algorithm reviewed in Section I-A, we develop our analysis in the sequel based on CHOCO-SGD for the following reasons. First, CHOCO-SGD is flexible enough to support the general class of compression schemes satisfying the compression operator condition [20], allowing us to model both analog and digital implementations. Second, it provides state-of-the-art convergence guarantees for in noiseless and rate-limited communications, thus offering a suitable baseline to adapt to various wireless transmission protocols. At the start of each iteration $t+1$, device $i \in \mathcal{V}$ has in its memory its current model iterate $\boldsymbol{\theta}_i^{(t)}$, the corresponding estimated version $\hat{\boldsymbol{\theta}}_i^{(t)}$ and the estimated iterates $\hat{\boldsymbol{\theta}}_j^{(t)}$ for all its neighbors $j \in \mathcal{N}_i$. We note that an equivalent version of the algorithm that requires less memory can be found in [20, Algorithm 6], but we do not consider it here since it does not change the communication requirements. Furthermore, at each iteration t, device $i \in \mathcal{V}$ first executes a local update step by SGD based on its data set $\mathcal{D}_i$ as

$$\boldsymbol{\theta}_i^{(t+1/2)} = \boldsymbol{\theta}_i^{(t)} - \eta^{(t)}\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t)}), \tag{1}$$

where $\eta^{(t)}$ denotes the learning rate, and $\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t)})$ is an estimate of the exact gradient $\nabla f_i(\boldsymbol{\theta}_i^{(t)})$ obtained from a mini-batch $\mathcal{D}_i^{(t)} \subseteq \mathcal{D}_i$ of the data set, i.e., $\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t)}) = \frac{1}{|\mathcal{D}_i^{(t)}|} \sum_{\boldsymbol{\xi} \in \mathcal{D}_i^{(t)}} \nabla l(\boldsymbol{\theta}_i^{(t)}; \boldsymbol{\xi})$. [1]

Then, each device $i \in \mathcal{V}$ compresses the difference $\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}$ between the locally updated model (1) and the previously estimated iterate $\hat{\boldsymbol{\theta}}_i^{(t)}$. The compressed difference $\mathcal{C}^{(t)}(\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)})$ is then exchanged with the neighbors of node i. Assuming that communication is reliable — an assumption that we will revisit in the rest of the paper—each device $i \in \mathcal{V}$ updates the estimated model parameters $\hat{\boldsymbol{\theta}}_j^{(t)}$ for itself and for its neighbors as

$$\hat{\boldsymbol{\theta}}_j^{(t+1)} = \hat{\boldsymbol{\theta}}_j^{(t)} + \mathcal{D}^{(t)}\left(\mathcal{C}^{(t)}(\boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)})\right), \; j \in \{i\} \cup \mathcal{N}_i, \tag{2}$$

where $\mathcal{D}^{(t)}(\cdot)$ is a decoding function. Next, device $i \in \mathcal{V}$ executes a consensus update step by correcting the updated model (1) using the estimated parameters (2) as

$$\boldsymbol{\theta}_i^{(t+1)} = \boldsymbol{\theta}_i^{(t+1/2)} + \zeta^{(t)} \sum_{j \in \mathcal{N}_i \cup \{i\}} w_{ij} \left(\hat{\boldsymbol{\theta}}_j^{(t+1)} - \hat{\boldsymbol{\theta}}_i^{(t+1)}\right), \tag{3}$$

where $\zeta^{(t)}$ is the consensus rate, and the mixing matrix $\boldsymbol{W} = \boldsymbol{W}^T \in \mathbb{R}^{K \times K}$ is selected to be *doubly stochastic*, i.e., $[\boldsymbol{W}]_{ij} = w_{ij} \geq 0$, $\boldsymbol{W}\mathbf{1} = \mathbf{1}$, $\mathbf{1}^T\boldsymbol{W} = \mathbf{1}^T$ and $\|\boldsymbol{W} - \mathbf{1}\mathbf{1}^T/K\|_2 < 1$. We postpone discussion regarding the compression operator $\mathcal{C}^{(t)}(\cdot)$ and the decoding operator $\mathcal{D}^{(t)}(\cdot)$ to Section II-C. The considered decentralized learning protocol is summarized in Algorithm 1.

Finally, we make the following assumptions that are widely adopted in the literature on decentralized stochastic optimization [20].

Assumption 2.1: Each local empirical risk function $f_i(\boldsymbol{\theta})$, $i \in \mathcal{V}$, is L-smooth and μ-strongly convex, that is, for all $\boldsymbol{\theta}_1 \in \mathbb{R}^{d \times 1}$ and $\boldsymbol{\theta}_2 \in \mathbb{R}^{d \times 1}$, it satisfies the inequalities

$$f_i(\boldsymbol{\theta}_1) \leq f_i(\boldsymbol{\theta}_2) + \nabla f_i(\boldsymbol{\theta}_2)^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + \frac{L}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2, \tag{4}$$

and

$$f_i(\boldsymbol{\theta}_1) \geq f_i(\boldsymbol{\theta}_2) + \nabla f_i(\boldsymbol{\theta}_2)^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + \frac{\mu}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2. \tag{5}$$

Assumption 2.2: The variance of the mini-batch gradient $\hat{\nabla} f_i(\boldsymbol{\theta}_i)$ is bounded as

$$\mathbb{E}_{\mathcal{D}_i^{(t)}}[\|\hat{\nabla} f_i(\boldsymbol{\theta}_i) - \nabla f_i(\boldsymbol{\theta}_i)\|^2] \leq \sigma_i^2, \tag{6}$$

[1]The gradient is estimated by averaging the sample-wise gradient $\nabla l(\boldsymbol{\theta}_i^{(t)}; \boldsymbol{\xi})$ over a mini-batch $\mathcal{D}_i^{(t)}$ of the data set with replacement, and therefore the estimate $\hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t)})$ coincides with the true gradient $\nabla f_i(\boldsymbol{\theta}_i^{(t)})$ when $\mathcal{D}_i^{(t)} = \mathcal{D}_i$.

Algorithm 1: Decentralized Learning with Noiseless Communication

Input : Consensus step size $\zeta^{(t)}$, SGD learning step size $\eta^{(t)}$, connectivity graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and mixing matrix $\boldsymbol{W}$

Initialize at each node $i \in \mathcal{V}$: $\boldsymbol{\theta}_i^{(0)}$, $\hat{\boldsymbol{\theta}}_j^{(0)} = \mathbf{0}$, $\forall j \in \mathcal{N}_i \bigcup \{i\}$;

for $t = 0, 1, \ldots, T-1$ **do**

  for each device $i \in \mathcal{V}$ **do** in parallel

    update $\boldsymbol{\theta}_i^{(t+1/2)} = \boldsymbol{\theta}_i^{(t)} - \eta^{(t)} \hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t)})$;

    compress the difference $\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}$ to obtain $\mathcal{C}^{(t)}(\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)})$;

    for each neighboring device $j \in \mathcal{N}_i$ **do** in parallel

      send $\mathcal{C}^{(t)}(\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)})$;

      receive $\mathcal{C}^{(t)}(\boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)})$;

    end

    update $\hat{\boldsymbol{\theta}}_j^{(t+1)} = \hat{\boldsymbol{\theta}}_j^{(t)} + \mathcal{D}^{(t)}\left(\mathcal{C}^{(t)}(\boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)})\right)$, for $j \in \{i\} \cup \mathcal{N}_i$;

    udpate $\boldsymbol{\theta}_i^{(t+1)} = \boldsymbol{\theta}_i^{(t+1/2)} + \zeta^{(t)} \sum_{j \in \mathcal{N}_i \cup \{i\}} w_{ij} \left(\hat{\boldsymbol{\theta}}_j^{(t+1)} - \hat{\boldsymbol{\theta}}_i^{(t+1)}\right)$.

  end

end

Output: $\boldsymbol{\theta}_i^{(T-1)}$, $\forall i \in \mathcal{V}$

and its expected Euclidean norm is bounded as

$$\mathbb{E}_{\mathcal{D}_i^{(t)}}[\|\hat{\nabla} f_i(\boldsymbol{\theta}_i)\|^2] \leq G^2, \tag{7}$$

where the expectation $\mathbb{E}_{\mathcal{D}_i^{(t)}}[\cdot]$ is taken over the selection of a mini-batch $\mathcal{D}_i^{(t)} \subseteq \mathcal{D}_i$.

B. Communication Model

As seen in Fig. 2, at the end of every iteration t, communication takes place within one communication block of a total number N of channel uses spanning over M equal-length slots, denoted by $\mathcal{S} = \{1, \ldots, M\}$. Slow fading remains constant across all iterations, and is binary, determining whether a link is blocked ar not. A link $(i, j) \in \mathcal{E}$ is by definition not *blocked*, while all the other links $(i, j) \notin \mathcal{E}$ are blocked. We assume that the connectivity graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with all the unblocked links as edges satisfies the following assumption.

Assumption 2.3: Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is a connected graph.

For all unblocked links $(i,j) \in \mathcal{E}$, the channel coefficient between device i and j is modelled as

$$h_{ij}^{\prime(t)} \triangleq \sqrt{A_0}\left(\frac{d_{ij}}{d_0}\right)^{-\frac{\gamma}{2}} h_{ij}^{(t)}, \tag{8}$$

where the small-scale fading coefficient $h_{ij}^{(t)} \sim \mathcal{CN}(0,1)$ remains unchanged within one communication block and varies independently across blocks, and the path loss gain $\sqrt{A_0}(d_0/d_{ij})^{\gamma/2}$ is constant across all iterations, where A_0 is the average channel power gain at reference distance d_0; d_{ij} is the distance between device i and j; and γ is the path loss exponent factor.

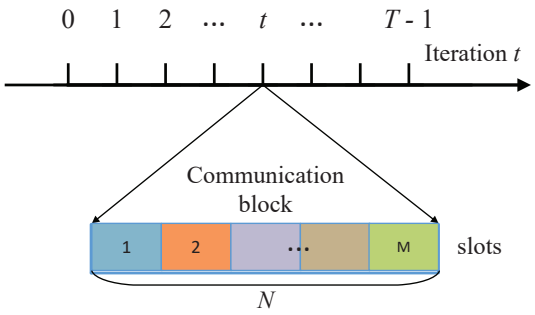


Fig. 2. Timeline of training iterations and communication blocks: A communication block of N channel uses, divided into M slots, is employed to exchange the compressed difference between model parameters among neighboring devices.

Each device is subject to an energy constraint of $NP^{(t)}$ per communication block. If a device is active for $M' \leq M$ slots, the energy per symbol is hence given by $\frac{NP^{(t)}}{M'N/M} = P^{(t)}\frac{M}{M'}$. The mean-square power of the added white Gaussian noise (AWGN) is denoted as N_0.

C. Compression

In this subsection, we describe the assumed compression operator $\mathcal{C}(\cdot)$ and decompression operator $\mathcal{D}(\cdot)$ that are used in the update (3). We specifically adopt *random linear coding (RLC)* compression [10]. Let $\boldsymbol{A}^{(t)} = \frac{1}{\sqrt{m}}\boldsymbol{H}\boldsymbol{R}^{(t)}$ be the linear encoding matrix, where $\boldsymbol{H} \in \{\pm 1\}^{m\times d}$ with $m \leq d$ is a partial Hadamard matrix with mutually orthogonal rows, i.e., $\frac{1}{d}\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{I}$; and $\boldsymbol{R}^{(t)} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with its diagonal entries $[\boldsymbol{R}^{(t)}]_{ii} \triangleq r_i^{(t)}$ drawn from uniform distributions such that $\Pr(r_i^{(t)} = 1) = \Pr(r_i^{(t)} = -1) = 0.5$, for all $i = 1,\ldots,d$. The compression operator is given by the linear projection $\mathcal{C}^{(t)}(\boldsymbol{u}) = \boldsymbol{A}^{(t)}\boldsymbol{u}$, while decoding takes place as $\mathcal{D}^{(t)}(\boldsymbol{v}) = \frac{m}{d}(\boldsymbol{A}^{(t)})^T\boldsymbol{v}$. The complexity of the encoding/decoding operations is linear with order $\mathcal{O}(md)$, and thus easy to implement even on IoT devices. The concatenation of the

compression and decompression operators, namely, $\mathcal{D}^{(t)}(\mathcal{C}^{(t)}(\boldsymbol{u})) = \frac{m}{d}(\boldsymbol{A}^{(t)})^T\boldsymbol{A}^{(t)}\boldsymbol{u}$, satisfies the *compression operator* condition [10], [20]

$$\mathbb{E}\left\|\boldsymbol{u} - \frac{m}{d}(\boldsymbol{A}^{(t)})^T\boldsymbol{A}^{(t)}\boldsymbol{u}\right\|^2 = \left(1 - \frac{m}{d}\right)\|\boldsymbol{u}\|^2, \text{ for all } \boldsymbol{u} \in \mathbb{R}^{d\times 1}. \tag{9}$$

We note that the random matrices $\{\boldsymbol{R}^{(t)}\}$ need to be shared among devices such that the same random sequence $\{\boldsymbol{A}^{(t)}\}$ is agreed upon by all devices. In practice, sharing the random sequence $\{\boldsymbol{A}^{(t)}\}$ can be done offline by tasking a randomly selected device to flood d seeds through the network during a calibration phase. Since this phase takes place only once, its communication overhead can be neglected. Finally, by (9), the quality of signal reconstruction in terms of the mean-square error (MSE) of RLC can be easily quantified, making RLC a natural choice to develop all analysis framework for wireless implementations of FL, as elaborated on in Section IV and Section V.

III. Digital and Analog Transmission Protocols

In this section, we describe digital and analog wireless implementations of the decentralized learning algorithm reviewed in the previous section. The implementations are meant to serve as prototypical templates for the deployment of decentralized learning. In practice, specific scheduling strategies are in need to allocate transmission slots as seen in Fig. 2 to devices in collaboration.

A. Digital Transmission

In digital transmission protocol, devices represent their model updates as digital messages for transmission. At iteration t, each device $i \in \mathcal{V}$ broadcast to all its neighbors using one dedicated time slot of the communication block decided by the scheduling policy in place.

1) Scheduling: The analysis to be developed in Section IV applies to any scheduling protocol that satisfies the following conditions: *(i)* no two connected devices transmit in the same slot due to the half-duplex transmission constraints; and *(ii)* no two devices connected to the same device transmit in the same slot, so as not to cause interference at their common neighbor. To design a scheduling scheme meeting these properties, one can construct an auxiliary graph $\mathcal{G}^d(\mathcal{V}, \mathcal{E}^d)$ with degree $\Delta^d < K$ such that the edge set $\mathcal{E}^d \supseteq \mathcal{E}$ includes not only the original edges in $\mathcal{E}$, but also one edge for each pair of nodes sharing one or more common neighbors. One can then carry out vertex-coloring on the auxiliary graph $\mathcal{G}^d(\mathcal{V}, \mathcal{E}^d)$ using a classical greedy algorithm,

such that any two nodes connected by an edge are assigned distinct colors [35, Algorithm G]. Scheduling proceeds by assigning the nodes with the same color to the same slot (see [32] for an example). The complexity of this centralized scheduling scheme is of the order $\mathcal{O}(K^2)$ [35]. There are also decentralized implementations of vertex-coloring that could be applied to this setting with the complexity of $\mathcal{O}(\Delta^d K)$ [36].

2) Transmission: The number $B_i^{(t)}$ of bits that device $i \in \mathcal{V}$ can successfully broadcast to its neighbors during a slot allocated to it by the scheduling algorithm is limited by the neighboring device with the worst channel power gain. Accordingly, we have

$$B_i^{(t)} = \frac{N}{M} \log_2 \left(1 + \frac{P^{(t)}}{N_0} M \min_{j \in \mathcal{N}_i} |h_{ij}'^{(t)}|^2 \right). \tag{10}$$

We recall that, in (10), the number M of time slots per iteration is decided by the scheduling scheme.

To quantize the encoded vector $\boldsymbol{A}_i^{(t)}(\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}) \in \mathbb{R}^{m \times 1}$ into $B_i^{(t)}$ bits, we employ a simple per-element b-bit quantizer $\mathcal{Q}_b(\cdot)$ with chip-level precision so that $b = 64$ or $b = 32$ is for double-precision or single-precision floating-point, respectively, according to IEEE 754 standard. Communication constraints thus impose the inequality $bm_i^{(t)} \leq B_i^{(t)}$, which is satisfied by setting $m_i^{(t)} = \lfloor B_i^{(t)}/b \rfloor$. Based on the (received) quantized signal, each device $i \in \mathcal{V}$ updates the estimated model parameters of its own as well as of its neighbors in $j \in \{i\} \cup \mathcal{N}_i$ as (cf. (2))

$$\hat{\boldsymbol{\theta}}_j^{(t+1)} = \hat{\boldsymbol{\theta}}_j^{(t)} + \frac{m_j^{(t)}}{d} (\boldsymbol{A}_j^{(t)})^T \mathcal{Q}_b \left(\boldsymbol{A}_j^{(t)} (\boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)}) \right). \tag{11}$$

In order to implement update (11), each node $j \in \mathcal{V}$ and its neighbors in set $\mathcal{N}_j$ can share *a priori* a common sequence of (pseudo-)random matrix $\tilde{\boldsymbol{A}}_j^{(t)} \in \mathbb{R}^{m \times d}$ as described in Section II-C. If node j sends its current value $m_j^{(t)}$ to all neighbors and if $m_j^{(t)} \leq m$, all nodes can thus select the same submatrix $\boldsymbol{A}_j^{(t)}$ from $\tilde{\boldsymbol{A}}_j^{(t)}$ to evaluate (11). The described digital implementation is summarized in Appendix A.

B. Analog Transmission

With analog transmission, devices directly transmit their respective updated parameters by mapping analog signals to channel uses, without the need for digitization. As studied in [24], in addition to broadcast as in digital transmission, it is also useful to schedule all devices that share a common neighbor for simultaneous transmission in order to enable AirComp. Specifically, time slots required to be scheduled in pairs. In the first slot of each pair, one or more center

nodes receive a superposition of the signals simultaneously transmitted by all their respective neighbors. The center nodes are referred to as AirComp receivers. In the second slot, the center nodes serve as broadcast transmitters communicating to all their neighbors. The total number M of time slots is thus given by twice as the number n of pairs of time slots, which is specified by the scheduling policy in use.

1) Scheduling: The considered analog transmission protocols can accommodate any scheduling policy that satisfies the following two principles: *(i)* no two connected nodes are scheduled as AirComp receivers in the same time slot due to half-duplex transmissions; and *(ii)* no two nodes sharing a same neighboring node are scheduled as AirComp receivers in the same time slot. The second condition implies that, when a common neighboring node transmits to enable AirComp, the node's signal transmitted to one of the neighbors will not cause interference to the other. There are potentially many feasible scheduling schemes satisfying the above constraints, and we provide in Appendix B a sequential scheduling policy for the purpose of illustration. The scheduling policy described therein aims at selecting as many non-interfering star-based sub-networks as possible in one pair of time slots.

To elaborate, we will use the following notation. For each device $i \in \mathcal{V}$, we define a set $\mathcal{S}_i^{\rm Tx} \subseteq \mathcal{S}$ of transmission slots, with $\mathcal{S}_i^{\rm Tx} = \mathcal{S}_i^{\rm BT} \cup \mathcal{S}_i^{\rm AT}$ partitioned into disjoint subsets $\mathcal{S}_i^{\rm BT}$ and $\mathcal{S}_i^{\rm AT}$ ($\mathcal{S}_i^{\rm BT} \cap \mathcal{S}_i^{\rm AT} = \emptyset$). Subset $\mathcal{S}_i^{\rm BT}$ denotes the set of transmission slots in which device i broadcasts to its neighbors, and $\mathcal{S}_i^{\rm AT}$ denotes the set of slots in which device i transmits to enable AirComp. Similarly, we define the set $\mathcal{S}_i^{\rm Rx} \subseteq \mathcal{S}$ of receiving slots for device i as $\mathcal{S}_i^{\rm Rx} = \mathcal{S}_i^{\rm BR} \cup \mathcal{S}_i^{\rm AR}$ with $\mathcal{S}_i^{\rm BR} \cap \mathcal{S}_i^{\rm AR} = \emptyset$, where $\mathcal{S}_i^{\rm BR}$ and $\mathcal{S}_i^{\rm AR}$ denote the sets of receiving slots in which device $i \in \mathcal{V}$ receives from a transmitter in broadcast and AirComp modes, respectively. A sequential scheduling policy that satisfies the conditions *(i)* and *(ii)* listed in Section III-B.1) is described as follows.

First, we carry out greedy coloring on the auxiliary graph associated with the original connectivity graph $\mathcal{G}^{(1)} = \mathcal{G}(\mathcal{V}, \mathcal{E})$ as described in Section III-A. Next, defining $d_c^{(1)}$ as the sum of the degrees of all nodes that have been assigned the same color c at the first iteration, we set all nodes assigned color $c^* = \arg\max_c \{d_c^{(1)}\}$ in $\mathcal{G}^{(1)}$ as the center nodes, which compose the set $\mathcal{N}_{c^*}^{(1)}$. In the first slot, the nodes in $\mathcal{N}_{c^*}^{(1)}$ receive combined signals transmitted by their neighbors in $\mathcal{G}^{(1)}$; and in the subsequent second slot, the same set of nodes in $\mathcal{N}_{c^*}(1)$ broadcast their respective signal to their neighbors. As a result, the first slot is concurrently in set $\mathcal{S}_i^{\rm AR}$ and sets $\mathcal{S}_j^{\rm AT}$, for all nodes $j \in \mathcal{N}_i$. Conversely, the second slot is concurrently in set $\mathcal{S}_i^{\rm BT}$ and sets

$\mathcal{S}_j^{\mathrm{BR}}$ for all nodes $j \in \mathcal{N}_i$. The center nodes in $\mathcal{N}_{c^*}^{(1)}$ and their connected edges, along with any nodes disconnected from $\mathcal{G}^{(1)}$, are then removed to produce the residue graph $\mathcal{G}^{(2)}$. The overall procedure is repeated until the residue graph $\mathcal{G}^{(n)}$ ($n \geq 1$) becomes empty, and is summarized in Algorithm 3. Intuitively, the more edges are removed at each iteration, the fewer transmission slots will be needed. Since $d_c^{(n)}$ corresponds to the number of removable edges at iteration n, we schedule all nodes with color $c^* = \arg\max_c\{d_c^{(n)}\}$ as the center nodes to reduce the total number of required slots and improve spectral efficiency. We illustrate the outcome of this scheduling policy on the connectivity graph in Fig. 1 as an example (see Fig. 9 in Appendix B).

Since the number of nodes in the residue graph reduces by at least one at each iteration and the maximum number of iterations is no more than Δ^d, the overall complexity of both centralized and decentralized implementations of the sequential scheduling policy (c.f. Algorithm 3) proves to be $\mathcal{O}(\Delta^d K^2)$ [36].

2) Transmission: We now describe the transmitted and the received signals in each pair of slots of the communication protocol.

Odd slots: All devices $i \in \mathcal{V}$ operating in AirComp mode for a center node j in an odd slot $s \in \mathcal{S}_i^{\mathrm{AT}} \cap \mathcal{S}_j^{\mathrm{AR}}$ concurrently transmit the signals by pre-compensating the channel as

$$\boldsymbol{x}_{ij}^{(t,s)} = \frac{\sqrt{\gamma_j^{(t,s)}}}{h_{ij}'^{(t)}} w_{ji} \boldsymbol{A}^{(t)} (\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}), \tag{12}$$

where $\gamma_j^{(t,s)}$ is a power scaling factor for channel alignment at device j. The channel coefficient $h_{ij}'^{(t)}$ needs to be acquired at each device i operating in AirComp. To this end, at the beginning of each paired slots in the schedule, devices operating in AirComp can estimate channel through a pilot sent by their associated center node assuming channel reciprocity. The receiving center node, device j obtains

$$\boldsymbol{y}_j^{(t,s)} = \sqrt{\gamma_j^{(t,s)}} \sum_{i \in \mathcal{N}_j^{(s)}} w_{ji} \boldsymbol{A}^{(t)} (\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}) + \boldsymbol{n}_j^{(t,s)}, \tag{13}$$

where $\mathcal{N}_j^{(s)}$ is the neighboring set of device j operating in AirComp at slot s, and $\boldsymbol{n}_j^{(t,s)} \sim \mathcal{CN}(\boldsymbol{0}, N_0 \boldsymbol{I})$ is the received AWGN at slot s of iteration t. Device j estimates the *combined* model parameters $\sum_{i \in \mathcal{N}_j^{(s)}} w_{ji} (\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)})$ via the linear estimator

$$\hat{\boldsymbol{y}}_j^{(t,s)} = \frac{m}{d} (\boldsymbol{A}^{(t)})^T \Re \left\{ \boldsymbol{y}_j^{(t,s)} / \sqrt{\gamma_j^{(t,s)}} \right\}. \tag{14}$$

Even slots: Any device $i \in \mathcal{V}$ operating in broadcast mode in an even slot $s \in \mathcal{S}_i^{\mathrm{BT}}$ transmits a signal

$$\boldsymbol{x}_i^{(t,s)} = \sqrt{\alpha_i^{(t,s)}} \boldsymbol{A}^{(t)} (\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}), \tag{15}$$

where $\alpha_i^{(t,s)}$ is device i's transmitting power scaling factor in slot $s \in \mathcal{S}_i^{\mathrm{BT}}$ of iteration t. Each neighboring device $j \in \mathcal{N}_i^{(s)}$, with $s \in \mathcal{S}_j^{\mathrm{BR}}$, receives from device i the signal

$$\boldsymbol{y}_{ij}^{(t,s)} = \sqrt{\alpha_i^{(t,s)}} h_{ij}^{\prime(t)} \boldsymbol{A}^{(t)} (\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}) + \boldsymbol{n}_j^{(t,s)}, \tag{16}$$

where $\boldsymbol{n}_j^{(t,s)} \sim \mathcal{CN}(\boldsymbol{0}, N_0 \boldsymbol{I})$ is the received AWGN. Device j estimates the signal $\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}$ via the linear estimator

$$\hat{\boldsymbol{y}}_{ij}^{(t,s)} = w_{ji} \frac{m}{d} (\boldsymbol{A}^{(t)})^T \Re \left\{ \frac{\boldsymbol{y}_{ij}^{(t,s)}}{\sqrt{\alpha_i^{(t,s)}} h_{ij}^{\prime(t)}} \right\}, \tag{17}$$

where $\Re\{\cdot\}$ denotes the real part of its argument.

Next, device $j \in \mathcal{V}$ updates its estimate of the combined model parameters from all neighboring devices in $\mathcal{N}_j$ by aggregating the estimates obtained at all receiving slots in set $\mathcal{S}_j^{\mathrm{Rx}} = \mathcal{S}_j^{\mathrm{BR}} \cup \mathcal{S}_j^{\mathrm{AR}}$ as

$$\hat{\boldsymbol{y}}_j^{(t+1)} = \hat{\boldsymbol{y}}_j^{(t)} + \sum_{s \in \mathcal{S}_j^{\mathrm{AR}}} \hat{\boldsymbol{y}}_j^{(t,s)} + \sum_{s \in \mathcal{S}_j^{\mathrm{BR}}} \hat{\boldsymbol{y}}_{i_s j}^{(t,s)}, \tag{18}$$

where node $i_s \in \mathcal{N}_j$ is the node that transmits in broadcast mode in slot $s \in \mathcal{S}_j^{\mathrm{BR}}$. The initial estimate of the combined model parameters is given by $\hat{\boldsymbol{y}}_i^{(0)} = \boldsymbol{0}$, $\forall i \in \mathcal{V}$.

The power scaling parameters $\gamma_j^{(t,s)}$ (cf. (12)) and $\alpha_i^{(t,s)}$ (cf. (15)) for $s \in \mathcal{S}_i^{\mathrm{Tx}}$ need to be properly chosen such that the power consumed by device $i \in \mathcal{V}$ per communication block satisfies

$$\sum_{s \in \mathcal{S}_i^{\mathrm{AT}}} \mathbb{E}[\|\boldsymbol{x}_{ij_s}^{(t,s)}\|^2] + \sum_{s \in \mathcal{S}_i^{\mathrm{BT}}} \mathbb{E}[\|\boldsymbol{x}_i^{(t,s)}\|^2] \leq N P^{(t)}, \ \forall i \in \mathcal{V}, \tag{19}$$

where node j_s is the center node connected to node i in slot $s \in \mathcal{S}_i^{\mathrm{AT}}$. Applying a simple equal power policy across different transmission slots of a device $i \in \mathcal{V}$ for all communication blocks, we have (cf. (19))

$$\mathbb{E}[\|\boldsymbol{x}_{ij}^{(t,s)}\|^2] \leq N P^{(t)} / |\mathcal{S}_i^{\mathrm{Tx}}|, \ \forall s \in \mathcal{S}_i^{\mathrm{AT}}, \tag{20}$$

$$\mathbb{E}[\|\boldsymbol{x}_i^{(t,s)}\|^2] \leq N P^{(t)} / |\mathcal{S}_i^{\mathrm{Tx}}|, \ \forall s \in \mathcal{S}_i^{\mathrm{BT}}. \tag{21}$$

By substituting (12) for $\boldsymbol{x}_{ij}^{(t,s)}$ in (20), if follows that we have the inequality

$$\gamma_j^{(t,s)} \leq NP^{(t)} \frac{|h_{ij}'^{(t)}|^2}{|\mathcal{S}_i^{\mathrm{Tx}}| \|\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}\|^2 w_{ji}^2}, \tag{22}$$

which implies that the power scaling factor $\gamma_j^{(t,s)}$ for channel alignment at device j is chosen as $\gamma_j^{(t,s)} = \min_{i \in \mathcal{N}_j^{(s)}} v_{ij}^{(t)}$, where $v_{ij}^{(t)} = |h_{ij}'^{(t)}|^2 / (|\mathcal{S}_i^{\mathrm{Tx}}| \|\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}\|^2 w_{ji}^2)$. As a result, the power scaling factor $\gamma_j^{(t,s)}$ can be acquired in the initial channel estimation phase by finding the minimum of $\mathcal{N}_j^{(s)}$ of non-negative values at the center node j. When $|\mathcal{N}_j^{(s)}|$ is sufficiently small, each node $i \in \mathcal{N}_j^{(s)}$ can calculate $v_{ij}^{(t)}$ based on its estimated channel and send its value as feedback to the center node j. When the underlying connectivity is dense, and hence $|\mathcal{N}_j^{(s)}|$ is large, this approach may entail an excessive communication overhead. In such cases, an estimate of $\gamma_j^{(t,s)}$ can be found by AirComp-assisted wireless sensing following [37, Section 5]. Assuming a target estimation error of ϵ, the required communication overhead amounts to $\lceil \log_2(v_{\max}/2\epsilon) \rceil$ rounds of power detection, where $v_{\max}$ denotes the maximum range of $\gamma_j^{(t,s)}$. Then, the center node j shares $\gamma_j^{(t,s)}$ with all nodes in $\mathcal{N}_j^{(s)}$ by broadcasting.

Similarly, by substituting (15) for $\boldsymbol{x}_i^{(t,s)}$ in (21), the power scaling factor $\alpha_i^{(t,s)}$ is expressed as

$$\alpha_i^{(t,s)} = NP^{(t)} / \|\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}\|^2, \tag{23}$$

which requires negligible communication overhead.

Furthermore, each device $j \in \mathcal{V}$ needs to update the estimate of its own model parameter $\hat{\boldsymbol{\theta}}_j^{(t+1)}$ as

$$\hat{\boldsymbol{\theta}}_j^{(t+1)} = \hat{\boldsymbol{\theta}}_j^{(t)} + \frac{m}{d} (\boldsymbol{A}^{(t)})^T \boldsymbol{A}^{(t)} (\boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)}). \tag{24}$$

Finally, device $j \in \mathcal{V}$ approximates update (3) as

$$\boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(t+1/2)} + \zeta^{(t)} \left(w_{jj} \hat{\boldsymbol{\theta}}_j^{(t+1)} + \hat{\boldsymbol{y}}_j^{(t+1)} - \hat{\boldsymbol{\theta}}_j^{(t+1)} \right). \tag{25}$$

To sum up, the proposed analog implementation is presented in Appendix C.

IV. Convergence Analysis For Digital Transmission

In this section, we derive convergence properties of the general class of digital transmission protocols presented in Section III-A. The analysis holds for any fixed transmission schedule, which determines the number M of slots. We start by recalling that, at each iteration t, update (11) is carried out by device $i \in \mathcal{V}$ for all nodes $j \in \{i\} \cup \mathcal{N}_i$. In (11), the concatenation of

compression, quantization, and decompression yields an output vector $\frac{m_j^{(t)}}{d}(\boldsymbol{A}_j^{(t)})^T\ \mathcal{Q}_b(\boldsymbol{A}_j^{(t)}\boldsymbol{u}_j^{(t)})$ for the input vector $\boldsymbol{u}_j^{(t)} = \boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)}$. The number $m_j^{(t)} = \lfloor B_j^{(t)}/b \rfloor$ of rows of matrix $\boldsymbol{A}_j^{(t)}$ at iteration t depends on the current rate (10) supported by the fading channels between device j and its neighbors. Taking the randomness of the fading realizations into account, the counterpart of the compression operator (9) under digital transmission is given by the following lemma.

Lemma 4.1: On average over RLC, the MSE for the concatenation of compression, quantization, and decompression under digital transmission satisfies

$$\mathbb{E}\left\|\boldsymbol{u} - \frac{m_i^{(t)}}{d}(\boldsymbol{A}_i^{(t)})^T\mathcal{Q}_b(\boldsymbol{A}_i^{(t)}\boldsymbol{u})\right\|^2 \leq \left(1-\omega^{(t)}\right)\|\boldsymbol{u}\|^2, \tag{26}$$

for all $\boldsymbol{u} \in \mathbb{R}^{d\times 1}$ and for all $i \in \mathcal{V}$, where we have $\omega^{(t)} = \min_{i\in\mathcal{V}}\omega_i^{(t)}$ with $\omega_i^{(t)} = \frac{m_i^{(t)}}{d}$ and $m_i^{(t)}$ denoting the number of rows of $\boldsymbol{A}_i^{(t)}$.

Proof: Please refer to Appendix D. ■

By (26), the parameter $\omega = \min_{t\in\{0,\ldots,T-1\}}\{\omega^{(t)}\}$, where $\omega^{(t)} \in [0,1]$, is a measure of the quality of the reconstruction of the model difference used in update (11). Supposing static channel conditions in which the transmission rate (10) remains constant over iterations, we have $m_i^{(t)} = m_i$, $\forall i \in \mathcal{V}$. In these conditions, the right-hand side (RHS) of (26) can reduce to $(1-\frac{m_i}{d})\|\boldsymbol{u}\|^2$ (see [38, Appendix C]), which is exactly the RHS of (9) given $m = m_i$ and $\boldsymbol{A}^{(t)} = \boldsymbol{A}_i^{(t)}$. Assuming fading channel conditions, the following corollary quantifies the quality of model reconstruction on average over Rayleigh-fading channels.

Corollary 4.1: On average over RLC and the Rayleigh-fading based channel model described in section II, the MSE for the RLC compression operator (c.f. (9)) under digital transmission satisfies

$$\mathbb{E}\left\|\boldsymbol{u} - \frac{m_i^{(t)}}{d}(\boldsymbol{A}_i^{(t)})^T\mathcal{Q}_b(\boldsymbol{A}_i^{(t)}\boldsymbol{u})\right\|^2 \leq \left(1 - \frac{1}{d}\min_{i\in\mathcal{V}}\sum_{n=1}^{d} G_i^{(t)}(n)\right)\|\boldsymbol{u}\|^2, \tag{27}$$

for all $\boldsymbol{u} \in \mathbb{R}^{d\times 1}$ and for all $i \in \mathcal{V}$, where the function $G_i^{(t)}(n)\colon \mathbb{Z}_+ \mapsto \mathbb{R}$, $i \in \mathcal{V}$, defined as

$$G_i^{(t)}(n) = \exp\left(-\frac{N_0}{P^{(t)}}\frac{1}{MA_0}\left(2^{nb\frac{M}{N}}-1\right)\sum_{j\in\mathcal{N}_i}\left(\frac{d_{ij}}{d_0}\right)^r\right). \tag{28}$$

Proof: Please refer to Appendix E. ■

Note that the MSE for RLC given by the RHS of (27) decreases with $P^{(t)}$. In particular, when the transmission power $P^{(t)} \to \infty$, it is seen in (28) that $G_i(n) \to 1$, thus leading to zero mean-square estimation error.

With Lemma 4.1, the convergence properties of the digital protocol can be quantified in a manner similar to [20, Theorem 4]. To this end, we define the following topology-related parameters dependent on the mixing matrix $\boldsymbol{W}$: the spectral gap $\delta = 1 - \|\frac{\mathbf{1}\mathbf{1}^T}{K} - \boldsymbol{W}\|$; the parameter $\beta = \|\boldsymbol{I} - \boldsymbol{W}\|_2$; and the function $p(\delta, \omega) = \frac{\delta^2\omega}{2(16\delta+\delta^2+4\beta^2+2\delta\beta^2-8\delta\omega)}$ that depends on the spectral gap δ and on the model-difference estimation quality ω. Then, the convergence of the digital implementation is provided by the following theorem.

Theorem 4.2 (Optimality Gap for Digital Transmission [20, Theorem 19]): For learning rate $\eta^{(t)} = \frac{3.25}{\mu}\frac{1}{t+a}$ with $a \geq \max\{\frac{5}{p(\delta,\omega)}, \frac{13L}{\mu}\}$, consensus step size $\zeta^{(t)} = \frac{2p(\delta,\omega)}{\delta} \triangleq \zeta_0(\delta,\omega)$, and fixed fading realizations $\{h_{ij}'^{(t)}\}$, on average over SGD and RLC, Algorithm 2 yields an optimality gap satisfying

$$\begin{aligned}\mathbb{E}[F(\tilde{\boldsymbol{\theta}}_T)] - F^* &\leq \mathbb{E}\left[\frac{1}{S_T}\sum_{t=0}^{T-1} w^{(t)} F(\bar{\boldsymbol{\theta}}^{(t)})\right] - F^* \\ &\leq \underbrace{\frac{\mu}{3.25}\frac{a^3 - 3.25a^2}{S_T} v_e^{(0)} + \frac{1.625(2a+T)T}{\mu S_T}\frac{\bar{\sigma}^2}{K}}_{\text{centralized error}} + \underbrace{\frac{158.45 \times 24LT}{\mu^2 (p(\delta,\omega))^2 S_T} G^2}_{\text{consensus error}}, \end{aligned} \tag{29}$$

where $w^{(t)} = (a+t)^2$; $S_T = \sum_{t=0}^{T-1} w^{(t)}$; $\tilde{\boldsymbol{\theta}}_T = \frac{1}{S_T}\sum_{t=0}^{T-1} w^{(t)}\bar{\boldsymbol{\theta}}^{(t)}$ is the weighted sum of the average iterate $\bar{\boldsymbol{\theta}}^{(t)} = \frac{1}{K}\sum_{i\in\mathcal{V}} \boldsymbol{\theta}_i^{(t)}$ across all the communication rounds; F^* denotes the optimum objective value for problem (P0); μ is the parameter for μ-strongly convex function $F(\boldsymbol{\theta})$; $v_e^{(0)} = \|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\|^2$ measures the initial distance to the optimal model parameter; and $\bar{\sigma}^2 = \frac{1}{K}\sum_{i\in\mathcal{V}} \sigma_i^2$ is the average of the mini-batch gradient variance over all devices.

Remark 4.1: The labels "centralized error" and "consensus error" in (29) refer to decomposition of the upper bound on optimality gap into a term that accounts for the performance of the average model $\bar{\boldsymbol{\theta}}^{(t)} = \frac{1}{K}\sum_{i\in\mathcal{V}} \boldsymbol{\theta}_i^{(t)}$ and is inevitable even in centralized training due to the use of estimates of the gradients in SGD — the "centralized error", and a term that measures the disagreement among agents — the "consensus error". To gain insights into how wireless resources, channel conditions, and topology of the connectivity graph affect the performance of the digital wireless implementation, we can rewrite (29) using relevant parameters in "big O" notation as [20, Theorem 4]

$$\mathbb{E}[F(\tilde{\boldsymbol{\theta}}_T)] - F^* \leq \mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu K T}\right) + \mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu K T^2} + \frac{LG^2}{\mu^2 (p(\delta,\omega))^2 T^2}\right) + \mathcal{O}\left(\frac{\mu}{T^3}\right). \tag{30}$$

This result shows that when the total number of iterations T is sufficiently large, the optimality gap (30) behaves as $\mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu K T}\right)$, which recovers the convergence rate of centralized SGD with

ideal communications. However, when the wireless communication resources are limited, and hence $NT \ll \infty$, the second term, scaling as $\mathcal{O}\left(\frac{1}{T^2}\right)$, becomes equally important in (30), demonstrating the impact of the topology via δ and β, as well as the effect of the quality of digital transmission via ω. Since the function $p(\delta, \omega)$ is monotonically increasing with $\delta \in [0, 1]$ and $\omega \in [0, 1]$, the second term in (30) decreases with δ and ω. This implies that convergence is improved for more connected graphs with larger δ [39], and for smaller estimation errors with larger ω.

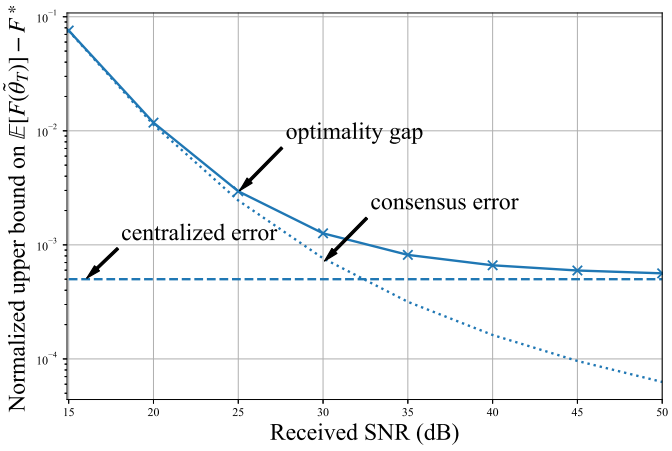
A. *Numerical Illustration*

In this subsection, we corroborate the analysis by numerically evaluating each constituent term in the upper bounds (29) on the optimality gap. We consider a setup consisting of $K = 20$ devices that are located at randomly and independently selected distances in the interval $[50, 200]$ m away from a center position, with all angles (in radius) uniformly distributed in the interval $[0, 2\pi]$. The connectivity graph, accounting for the impact of slow fading, is modelled as: *(i)* a complete graph; *(ii)* a planar 5×4 grid graph; *(iii)* a planar 5×4 grid graph with torus wrapping; or *(iv)* a star graph as in conventional FL. We set the strong-convexity parameter as $\mu = 0.0002$, and the smoothness factor as $L = 0.16$. We adopt the following standard choice of $\boldsymbol{W}$: $w_{ij} = \alpha$ for all $j \in \mathcal{N}_i$, $w_{ii} = 1 - |\mathcal{N}_i|\alpha$, and $w_{ij} = 0$ otherwise. We set $\alpha = 2/(\lambda_1(\boldsymbol{L}) + \lambda_{K-1}(\boldsymbol{L}))$, where $\boldsymbol{L}$ is the Laplacian of the connectivity graph. We plot the upper bound at iteration t normalized by the corresponding value at $t = 0$, hence evaluating the improvement in the expected optimality gap. The purpose of employing such normalized metrics is to standardize numerical values, and hence easier to compare. The SNR is defined as the received SNR, averaged over small-scale channel fading, at a distance of 125 meters (m) away from the deployment center.[2] Other parameters are set as $d_0 = 1$ m, $A_0 = 10^{-3.35}$, $\gamma = 3.76$ and $N_0 = -169$ dBm.

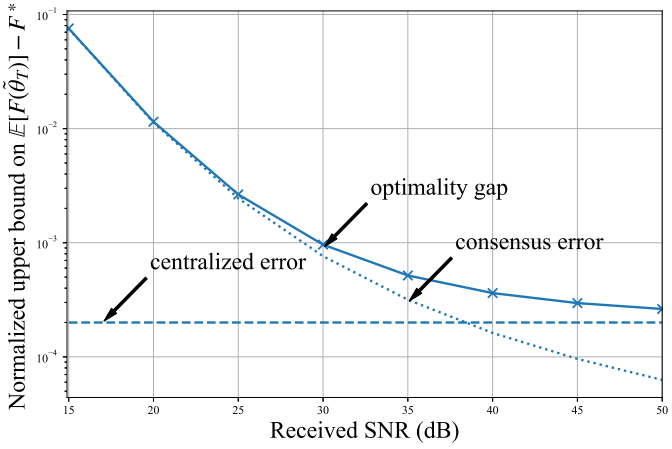
We apply the greedy vertex-coloring algorithm [35, Algorithm G] in order to determine the number of slots. Fig. 3 plots separately the centralized and the consensus errors, as well as the overall error in (29), for different number of iterations $t = 2000$ and $t = 5000$ under a planar topology with torus wrapping. The centralized error does not depend on the SNR, and it decreases at the fastest rate over iterations. The consensus error decreases with the received SNR due to the improvement in the parameter ω that characterizes the quality of model reconstruction

[2]We employ a benchmark distance of 125 m to make the received SNR fall into range of values of practical interest.

(cf. (29)). As a result, the consensus error dominates the optimality gap until the received SNR level increases to a sufficiently large value dependent on the iteration t. Furthermore, the overall optimality gap approaches the centralized error as the SNR increases.



(a) $t = 2000$.



(b) $t = 5000$.

Fig. 3. Normalized upper bounds (29) on the optimality gap versus the received SNR levels for a planar grid graph with torus wrapping and $N = 10^4$ ($a = 4 \times 10^7$).

We now turn to an analysis of the impact of the topology of the connectivity graph on the convergence for digital transmission. To this end, we set the same received SNR for all devices ignoring the impact of path loss in order to isolate the impact of different topologies. We employ TDMA-based scheduling that assigns only one device as the transmitter at one slot such that there are equal number $M = K$ of slots for all topologies. Fig. 4 reports the optimality gap,

along with the constituent errors in (29). The optimality gap decreases with the spectral gap δ of the connectivity graph, which equals $\delta = 1$, 0.31, 0.103, and 0.095, for complete graph, planar graph with and without torus wrapping, and star topology, respectively. Fig. 4 suggests that, under the constant-weight mixing strategy, if the underlying connectivity graph is denser, the nodes perform better model aggregation, thus achieving consensus faster (smaller "consensus error"). Fig. 4 shows that the consensus error contributes very little to the overall error for densely connected graphs, such as the complete graph and the planar grid with torus wrapping, while it becomes dominant in less densely connected graphs, such as the planar grid and the star graph.

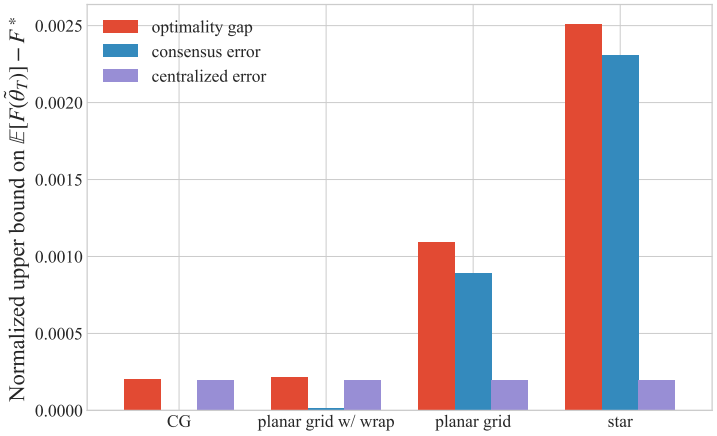


Fig. 4. Normalized upper bounds (29) on the optimality gap for different topologies of the connectivity graph with $M = K$, SNR$= 30$ dB, and $N = 10^4$ for $t = 5000$ ($a = 1.3 \times 10^8$).

V. Convergence Analysis For Analog Transmission

In this section, we derive convergence properties for the general class of analog transmission schemes described in Section III-B. The analysis holds for any fixed scheduling scheme operating in pairs of slots as described in Section III-B. We start by noting that, while for the digital implementation, the number $m_j^{(t)} = \lfloor B_j^{(t)}/b \rfloor$ rows of matrix $\boldsymbol{A}_j^{(t)}$ for device $j \in \mathcal{V}$ depends on the fading channels between device j and its neighbors, for the analog implementation, the number m of rows of matrix $\boldsymbol{A}^{(t)}$ is fixed as the number $m = N/M$ of available channel uses in each slot of the communication block. Therefore, for analog communication, we can directly use (9) to quantify the quality of the estimate of the model difference used in update (24). However, due to the presence of channel noise, update (18) for the combined parameters

of neighboring devices does not satisfy $\bar{\boldsymbol{\theta}}^{(t+1)} = \bar{\boldsymbol{\theta}}^{(t+1/2)}$. This calls for a novel derivation of convergence properties that does not follow from [20, Theorem 19].

Next, we relate the update in (25), which is subject to Gaussian noise, to the noiseless update in (3) in the following lemma.

Lemma 5.1: The consensus update (25) for analog implementation is equivalent to

$$\boldsymbol{\theta}_i^{(t+1)} = \boldsymbol{\theta}_i^{(t+1/2)} + \zeta^{(t)} \sum_{j \in \mathcal{N}_i \cup \{i\}} w_{ij} \left(\hat{\boldsymbol{\theta}}_j^{(t+1)} - \hat{\boldsymbol{\theta}}_i^{(t+1)} \right) + \zeta^{(t)} \sum_{\tau=0}^{t} \frac{m}{d} (\boldsymbol{A}^{(\tau)})^T \tilde{\boldsymbol{n}}_i^{(\tau)}, \tag{31}$$

where $\tilde{\boldsymbol{n}}_i^{(t)} \sim \mathcal{N}(0, \tilde{N}_{0i}^{(t)} \boldsymbol{I})$ is the effective noise with power

$$\tilde{N}_{0i}^{(t)} = \frac{1}{2} \frac{N_0}{N P^{(t)}} \left(\sum_{s \in \mathcal{S}_i^{\mathrm{AR}}} \max_{j \in \mathcal{N}_i^{(s)}} \left\{ |\mathcal{S}_j^{\mathrm{Tx}}| w_{ij}^2 \frac{\|\boldsymbol{u}_j^{(t)}\|^2}{\left|h_{ij}'^{(t)}\right|^2} \right\} + \sum_{s \in \mathcal{S}_i^{\mathrm{BR}}} |\mathcal{S}_{j_s}^{\mathrm{Tx}}| w_{ij_s}^2 \frac{\|\boldsymbol{u}_{j_s}^{(t)}\|^2}{\left|h_{ij_s}'^{(t)}\right|^2} \right), \tag{32}$$

where we denoted $\boldsymbol{u}_j^{(t)} = \boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)}$.

Proof: Please refer to Appendix F. ■

Comparing (31) with (3) reveals that consensus updates (25) for analog implementation are noisy approximation of those for ideal communication used by Algorithm 1 and digital implementation used by Algorithm 2. Note, in particular, that by (31), unlike Algorithms 1-2, we no longer have the preservation of the average of the model parameters across the network. In fact, the average $\bar{\boldsymbol{\theta}}^{(t+1)} = 1/K \sum_{i \in \mathcal{V}} \boldsymbol{\theta}_i^{(t+1)}$ of the model parameters, obtained by averaging over $i \in \mathcal{V}$ on both sides of (31), is corrupted by Gaussian noise as

$$\bar{\boldsymbol{\theta}}^{(t+1)} = \bar{\boldsymbol{\theta}}^{(t+1/2)} + \zeta^{(t)} \frac{1}{K} \sum_{i \in \mathcal{V}} \sum_{\tau=0}^{t} \frac{m}{d} (\boldsymbol{A}^{(\tau)})^T \tilde{\boldsymbol{n}}_i^{(\tau)}. \tag{33}$$

Next, we investigate how the noisy consensus updates given by (31) affect the convergence of Algorithm 4. In the proof of Theorem 19 in [20, Appendix D], the key step leading to the final convergence result is to construct an error sequence defined as

$$e^{(t)} = \underbrace{\sum_{i \in \mathcal{V}} \mathbb{E} \|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_i^{(t)}\|^2}_{\text{consensus error}} + \underbrace{\sum_{i \in \mathcal{V}} \mathbb{E} \|\hat{\boldsymbol{\theta}}_i^{(t+1)} - \boldsymbol{\theta}_i^{(t+1/2)}\|^2}_{\text{compression error}}, \tag{34}$$

where the first term measures the consensus error, while the second term accounts for the impact of compression. It is shown in [20, Appendix D] that the above two sub-terms are coupled with each other via a recursive relation that yields the main result summarized in the previous section. However, as discussed, their approach cannot be directly applied to analog transmission due to the fact that the consensus preserving property no longer holds, i.e., $\bar{\boldsymbol{\theta}}^{(t+1)} \neq \bar{\boldsymbol{\theta}}^{(t+1/2)}$, as seen

from (33). To address this challenge, we need a new upper bound on (34). To this end, we first introduce the function

$$p^{(t)}(\delta,\omega)=\min\{\tilde{p}^{(t)}(\delta,\omega),p(\delta,\omega)\}, \tag{35}$$

where

$$\tilde{p}^{(t)}(\delta,\omega)=\frac{\delta\zeta_0(\delta,\omega)}{\sqrt[4]{\tilde{N}_{0,T}}t/a'+1}-\left(\frac{\delta^2}{4}+\frac{2}{\omega}\beta^2\right)\frac{(\zeta_0(\delta,\omega))^2}{\left(\sqrt[4]{\tilde{N}_{0,T}}t/a'+1\right)^2}, \tag{36}$$

with $p(\delta,\omega)$ and $\zeta_0(\delta,\omega)$ defined in Theorem 4.2. We also denote

$$\tilde{N}_{0,T}=\max_{t\in\{0,\ldots,T-1\}}\left\{\sum_{i\in\mathcal{V}}\tilde{N}_{0i}^{(t)}\right\}, \tag{37}$$

and $\omega=m/d$. We will see that the function $p^{(t)}(\delta,\omega)$ plays an analogous role to $p(\delta,\omega)$ for digital transmission in that it contributes to the decaying rate of the error sequence defined in (34). Specifically, both $\tilde{p}^{(t)}(\delta,\omega)$ and $p(\delta,\omega)$ are increasing functions of δ and ω.

A key ingredient of the proposed approach is to design a consensus step size that is adaptive to the iteration index t. This is because a consensus step size $\zeta^{(t)}=\zeta_0$, as used in [20], [21] for ideal communication, causes the accumulated channel noise term in (33) to grow with t, leading to a possible divergence of the upper bound on $e^{(t)}$ (see (68) in [38, Appendix E] for details). To suppress the error growth induced by channel noise, we propose an adaptive consensus step size $\zeta^{(t)}=\frac{\zeta_0(\delta,\omega)}{\sqrt[4]{\tilde{N}_{0,T}}t/a'+1}$, based on which we have the following lemma.

Lemma 5.2: For learning rate $\eta^{(t)}=\frac{3.25}{\mu}\frac{1}{t+a}$ with $a\geq\max\{\frac{5}{p^{(t)}(\delta,\omega)},\frac{13L}{\mu}\}$, adaptive consensus step size $\zeta^{(t)}=\frac{\zeta_0(\delta,\omega)}{\sqrt[4]{\tilde{N}_{0,T}}t/a'+1}$ with $a'>a\sqrt[4]{\tilde{N}_{0,T}}$, and $T<\infty$, on average over RLC, the error sequence in (34) satisfies

$$e^{(t+1)}\leq\left(1-\frac{p^{(t)}(\delta,\omega)}{2}\right)e^{(t)}+\frac{1}{p^{(t)}(\delta,\omega)}(\eta^{(t)})^2\left(24KG^2+A(\delta,\omega)\sqrt[4]{\tilde{N}_0}\right), \tag{38}$$

where $A(\delta,\omega)=\delta(\zeta_0(\delta,\omega)a')^3(2-\omega)\omega^2d(\frac{\mu}{3.25})^2$ is a function depending on the spectral gap δ of graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ and the estimation quality ω.

Proof: We provide herein an outline of the proof highlighting the key steps. First, to understand the dynamics of the error sequence $e^{(t)}$, we need to revisit the respective upper bounds on the one-step dynamics of its constituents, i.e., the consensus error $\sum_{i\in\mathcal{V}}\mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}_i^{(t)}\|^2$ (c.f. [20, Lemma 17]) and the compression error $\sum_{i\in\mathcal{V}}\mathbb{E}\|\hat{\boldsymbol{\theta}}_i^{(t+1)}-\boldsymbol{\theta}_i^{(t+1/2)}\|^2$ (c.f. [20, Lemma 18]). Unlike the setting studied in [20], both terms are affected by the effective noise power $\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}$ accumulated up to iteration t. Specifically, this noise term appears in the bounds

multiplied by $(\zeta^{(t)})^2$. Therefore, the consensus rate $\zeta^{(t)}$ should be properly designed so as to balance the conflicting requirements of reducing the impact of the accumulated channel noise and accelerating the consensus updates. Combining the bounds on consensus and compression errors produces an upper bound on the one-step dynamics of the overall error sequence $e^{(t)}$, which encompasses the first term in the RHS of (38) and an additional term that is the sum of a function of $\eta^{(t)}$ and a function of $\zeta^{(t)}$. Finally, by judiciously designing the sequence $p^{(t)}(\delta,\omega)$ to transform the function of $\zeta^{(t)}$ into an upper bound on the function of $\eta^{(t)}$, we recover (38). For the full version of this proof, please refer to Appendix G. ■

Remark 5.1: When communication is ideal, i.e., $\tilde{N}_{0,T} \to 0$, the function $p^{(t)}(\delta,\omega)$ reduces to $p(\delta,\omega)$, and the upper bound (38) reduces to the result derived in [20, Lemma 21].

By leveraging Lemma 5.2, the convergence of the analog implementation scheme is revealed as follows.

Theorem 5.1 (Optimality Gap for Analog Transmission): For a given total number T of iterations, learning rate $\eta^{(t)} = \frac{3.25}{\mu}\frac{1}{t+a}$ with $a \geq \max\{\frac{5}{p^{(t)}(\delta,\omega)}, \frac{13L}{\mu}\}$, adaptive consensus step size $\zeta^{(t)} = \frac{\zeta_0(\delta,\omega)}{\sqrt[4]{\tilde{N}_{0,T}}t/a'+1}$ with $a' > a\sqrt[4]{\tilde{N}_{0,T}}$, and fixed fading realizations $\{h_{ij}'^{(t)}\}$, on average over SGD, RLC and channel noise, Algorithm 4 yields an optimality gap satisfying

$$\mathbb{E}[F(\tilde{\boldsymbol{\theta}}_T)] - F^* \leq \mathbb{E}\left[\frac{1}{S_T}\sum_{t=0}^{T-1} w_t F(\bar{\boldsymbol{\theta}}^{(t)})\right] - F^* \leq \underbrace{\frac{\mu}{3.25}\frac{a^3 - 3.25a^2}{S_T}v_e^{(0)} + \frac{1.625(2a+T)T}{\mu S_T}\frac{\bar{\sigma}^2}{K}}_{\text{centralized error}} +$$

$$\underbrace{\frac{158.45 \times 24G^2LT}{\mu^2(p^{(T)}(\delta,\omega))^2 S_T}}_{\text{noiseless consensus error}} + \underbrace{\frac{158.45\frac{A(\delta,\omega)}{K}\sqrt[4]{\tilde{N}_{0,T}}LT}{\mu^2(p^{(T)}(\delta,\omega))^2 S_T} + \frac{1}{K^2}D(\delta,\omega)\sqrt{\tilde{N}_{0,T}}}_{\text{AWGN error}}, \quad (39)$$

where $D(\delta,\omega) = \omega^2 d\frac{\mu}{3.25}(\zeta_0(\delta,\omega)a')^2$.

Proof: We provide herein an outline of the proof highlighting the key steps. First, we characterize the exact upper bound on the error sequence $e^{(t)}$ based on Lemma 5.2 by leveraging [20, Lemma 22]. Next, we provide an upper bound on the dynamics of the distance to the optimal solution $\mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^*\|^2$ in terms of the optimality gap $\mathbb{E}[F(\bar{\boldsymbol{\theta}}^{(t)})] - F^*$ and of polynomials of the learning rate $\eta^{(t)}$. Unlike [20], this bound is also corrupted by the product of $(\zeta^{(t)})^2$ and the effective noise power $\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}$ accumulated up to iteration t. Finally, by properly designing the consensus rate $\zeta^{(t)}$, we arrive at an upper bound on the modified time-average of the optimality gap $\mathbb{E}[F(\bar{\boldsymbol{\theta}}^{(t)})] - F^*$ by using a variant of [40, Lemma 3.3]. For the full version of this proof, please refer to Appendix H. ■

Remark 5.2: Similar to analysis given in Section IV, we decompose the upper bound on the optimality gap (39) into different terms. The "centralized error" carries the same meaning as that in (29); the "noiseless consensus error" quantifies the disagreement among agents in the absence of communication noise, i.e., $\tilde{N}_{0i}^{(t)} = 0$ (cf. (32)); and the "AWGN error" accounts for the impact of Gaussian noise on the consensus updates in (25). To gain insights into how wireless resources, channel conditions, and topology of the connectivity graph affect the performance of the analog wireless implementation, we can write (39) in the following form

$$\mathbb{E}[F(\tilde{\boldsymbol{\theta}}_T)] - F^* \leq \mathcal{O}\left(\frac{\sqrt{\tilde{N}_{0,T}}}{K^2}\right) + \mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu K T^2} + \frac{\left(24G^2 + \frac{A(\delta,\omega)}{K}\sqrt[4]{\tilde{N}_{0,T}}\right)L}{\mu^2 (p^{(T)}(\delta,\omega))^2 T^2}\right) + \mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu K T}\right) + \mathcal{O}\left(\frac{\mu}{T^3}\right). \quad (40)$$

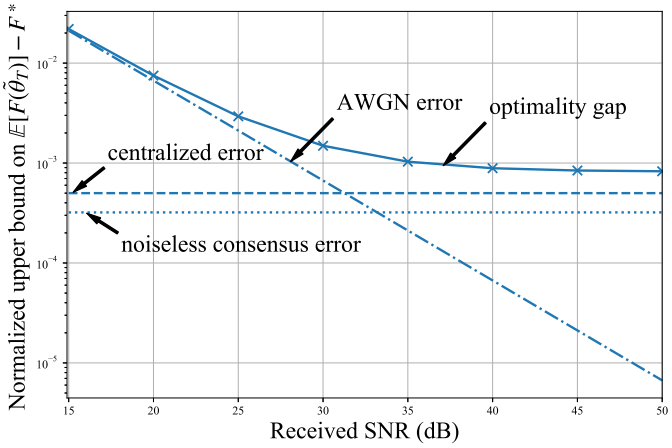
Compared with (30), the upper bound (40) reveals that, even if $T \to \infty$, there is a non-vanishing term $\mathcal{O}\left(\frac{\sqrt{\tilde{N}_{0,T}}}{K^2}\right)$, as well as a term scaling as $\mathcal{O}\left(\frac{\sqrt[4]{\tilde{N}_{0,T}}}{(p^{(T)}(\delta,\omega))^2 T^2}\right)$ that may not vanish either. We also note that $\tilde{N}_{0,T}$ is non-decreasing with T by its definition (37). This highlights the significant impact of the topology parameters δ and β, as well as of the quality of the reconstruction $\omega = m/d$ via function $p^{(T)}(\delta,\omega)$. Furthermore, as the effective noise power $\tilde{N}_{0i}^{(t)}$ in (32) for $i \in \mathcal{V}$ decreases with the transmitting power $P^{(t)}$, the convergence rate improves with $P^{(t)}$. In particular, when $P^{(t)} \to \infty$, we have $\tilde{N}_{0,T} = 0$, and therefore (40) reduces exactly to the corresponding expression for the noiseless case in (30).

Remark 5.3: Theorem 5.1 assumes smooth and strongly convex loss functions. Existing results such as [21, Theorem 4.1] on smooth and non-convex loss functions are based on constant learning step size, and they cannot be applied to our setting. This is because constant learning step sizes yield increasing upper bounds on the error sequence $e^{(t)}$ (c.f. (74)), thus leading to possible divergence of the time average of the expected gradient norms $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\|^2$, which is often adopted to analyze the convergence to stationary point for non-convex objectives. Accordingly, one of the key challenges in analyzing convergence bounds for analog wireless implementations in non-convex settings is to jointly design the diminishing learning step size $\eta^{(t)}$ and the decreasing consensus rate $\zeta^{(t)}$. This is left for future work.

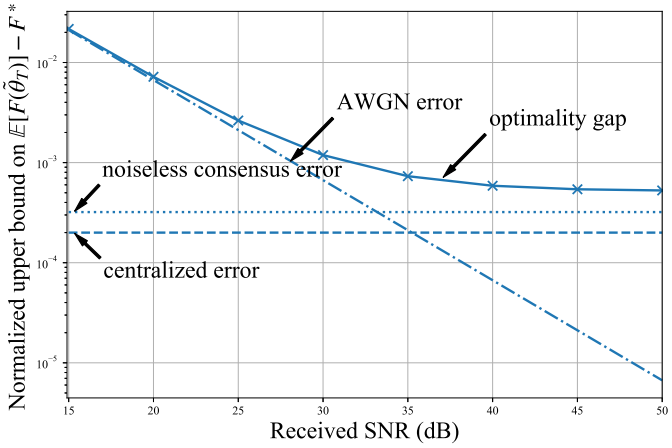
A. Numerical Illustration

In this subsection, we elaborate on the results obtained from the analysis above by following the approach in Section IV-A. For the scheduling scheme, we apply the sequential scheduling policy

proposed in Algorithm 3 (see Appendix B). Fig. 5 plots the upper bounds and the individual terms in (39) as in Fig. 3 for digital communication under the planar grid topology with torus wrapping. The centralized error and the noiseless consensus error are independent of SNR. Note that the latter depends on the parameter $\omega = \frac{N/M}{d}$ that quantifies the quality of the estimate in (9), which is independent of the SNR. As a result, the impact of SNR on the optimality gap is only through the AWGN error, which dominates the other terms when the SNR level is sufficiently small, here around 25 dB, while it becomes negligible when the SNR is large enough, here lager than 40 dB.



(a) $t = 2000$.



(b) $t = 5000$.

Fig. 5. Normalized upper bounds (39) on the optimality gap versus the received SNR levels for a planar grid graph with torus wrapping and $N = 10^4$ ($a = 7 \times 10^6$, $a' = a\sqrt[4]{\tilde{N}_{0,T}}$, and $\tilde{N}_{0,T} = 10^{-5}$).

Next, we study the impact of the topology of the connectivity graph on the convergence for analog transmission in Fig. 6. The general conclusions are analogous to Fig. 4, which illustrates the corresponding results for digital communication. In particular, the noiseless consensus error is shown to increase when the connectivity graph is less densely connected, i.e., with smaller δ. In contrast, the AWGN error is less sensitive to a change in connectivity, and it becomes dominant for sufficiently large number of iterations such as $t = 5000$.

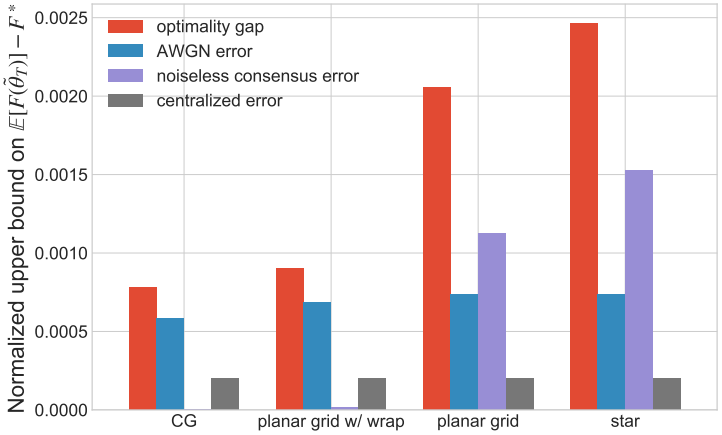


Fig. 6. Normalized upper bounds (39) on the optimality gap for different topologies of the connectivity graph with $M = K$, SNR$=30$ dB, and $N = 10^4$ for $t = 5000$ ($a = 3 \times 10^7$, $a' = 2a\sqrt[4]{\tilde{N}_{0,T}}$, and $\tilde{N}_{0,T} = 10^{-5}$ corresponding to 30 dB received SNR).

VI. Numerical Experiments

In this section, we corroborate the analysis developed in Sections IV and V by evaluating the empirical performance of the digital and analog wireless implementations over a wireless D2D network. We consider the learning task of image classification over the Fashion-MNIST dataset [41] that consists of 28×28 images divided into $C = 10$ classes. There are $60,000$ training data samples and $10,000$ test data samples, which are equally divided among classes. Each device $i \in \mathcal{V}$ has data samples from at least six different classes, with the number of missing classes being uniformly selected in the set $\{0, 1, 2, 3, 4\}$. An equal number x_i of data samples are then selected across the available classes at device i, such that the total number $\sum_{i\in\mathcal{V}} \sum_{n=1}^{10} \mathbb{1}_{n,i} x_i$ of training samples in use are maximized, where $\mathbb{1}_{n,i}$ is an indicator function denoting whether class $n \in \{1, \ldots, 10\}$ is available at device i ($\mathbb{1}_{n,i} = 1$) or not ($\mathbb{1}_{n,i} = 0$). All devices share a softmax regression model. We adopt the standard cross-entropy loss with

ℓ_2 regularization $f_i(\boldsymbol{\theta}) = -\frac{1}{|\mathcal{D}_i|}\sum_{\boldsymbol{\xi}\in\mathcal{D}_i}\sum_{n\in\Omega_i} y_{i,n}\log(\text{softmax}_n(\boldsymbol{a}_i(\boldsymbol{\theta},\boldsymbol{\xi}))) + \frac{\mu}{2}\|\boldsymbol{\theta}\|^2$, where Ω_i denotes the set of available classes on device $i \in \mathcal{V}$; $y_{i,n} \in \{0,1\}$ is the one-hot encoded label corresponding to the nth class for data sample $\boldsymbol{\xi} \in \mathcal{D}_i$; the $n$the entry of softmax is defined as $\text{softmax}_n(\boldsymbol{z}) = \frac{e^{z_n}}{\sum_{n=1}^{C} e^{z_n}}$, where $\boldsymbol{z} = (z_1, \ldots, z_C)^T$; and $\boldsymbol{a}_i(\boldsymbol{\theta},\boldsymbol{\xi})$ is the vector consisting of the logits for device i with its nth entry corresponding to class n. The SGD is executed with mini-batch size of $|\mathcal{D}_i^{(t)}| = 64$, and we add momentum to all updates with a factor of 0.9. In line with the local empirical risk function defined above, the strong-convexity parameter is set as $\mu = 0.002$. The smoothness factor L is numerically computed as the largest eigenvalue of the data Gramian matrix.

The optimal value F^* used for quantifying the empirical optimality gap is numerically obtained by the standard decentralized SGD (applying Algorithm 1 with $\mathcal{C} = \mathcal{D} = \boldsymbol{I}$ for sufficiently large T), with the hyper parameters $a > 0$ for the learning rate $\eta^{(t)} = \frac{3.25}{\mu}\frac{1}{t+a}$ and the (constant) consensus step size $\zeta_0 \in (0,1]$ optimized via grid search. We consider $K = 20$ devices with the connectivity graph modelled as a planar 5×4 grid graph with torus wrapping or a chain graph. All the other parameters for simulations are set as in Section IV-A unless specified otherwise. As benchmarks, we consider decentralized learning with ideal communications, i.e., applying Algorithm 1 with $\mathcal{C}^{(t)} = \mathcal{D}^{(t)} = \boldsymbol{I}$, as well as independent learning that carries out training based solely on local data with no communications among devices.[3]

We start by studying the impact of the proposed adaptive consensus rate on the convergence for analog transmission. We recall that the analysis in Section V has revealed that a consensus rate $\zeta^{(t)} = \frac{\zeta_0}{t/1000+1}$ is necessary in order to ensure convergence. In line with the analysis, Fig. 7 shows that the optimality gap with fixed consensus rates diverges as a function of the number of iteration t, while it converges with the proposed adaptive consensus rate. The figure also demonstrates the advantages brought by communications w.r.t local training with no communications.

Next, we provide a performance comparison between digital and analog wireless implementations in terms of the analytical and the empirical upper bounds on the optimality gap. We plot the optimality gap normalized by its value evaluated at iteration $t = 200$ and $t = 0$ for the analytical and the empirical results, respectively. For the empirical results, we set $\eta^{(t)} = \frac{3.25}{\mu}\frac{1}{t+200}$ for all schemes; $\zeta^{(t)} = 0.001$ for the schemes of ideal, digital and no communication; and $\zeta^{(t)} = \frac{0.001}{t/d+1}$ with $d = 35355,\ 42045,\ 50000,\ 62872,$ and 74767 for $N = 500,\ 1000,\ 2000,\ 5000,$ and $10000,$

[3]The source code implementing numerical experiments is available at https://github.com/Fuzzy-Face/JSAC_FL.

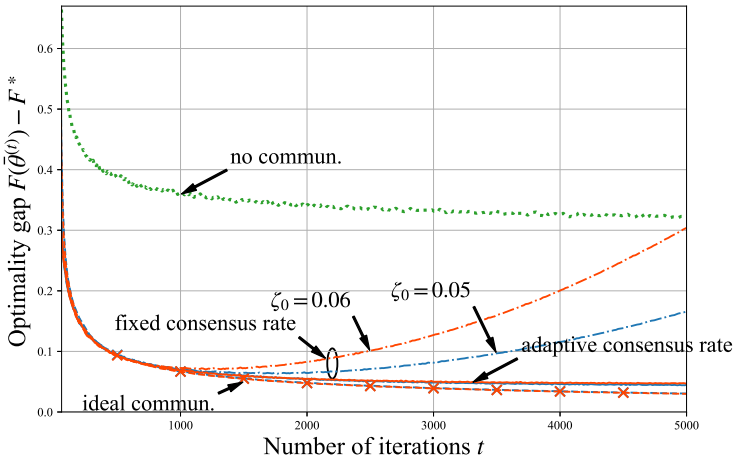
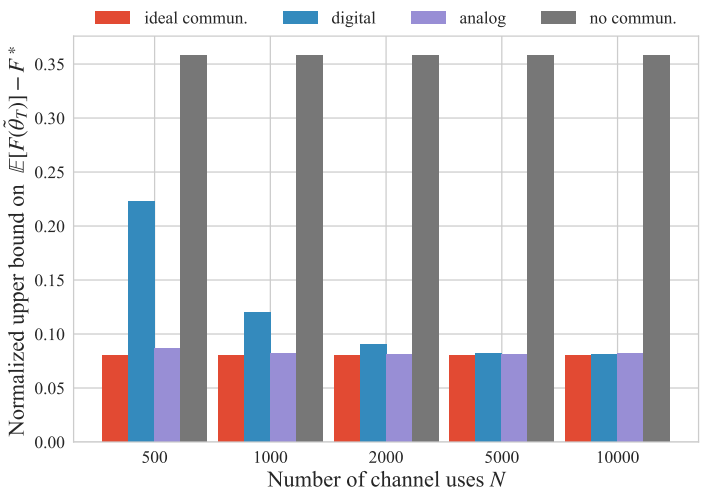


Fig. 7. Empirical optimality gap versus the number of iterations t for analog transmission over a planar grid graph with torus wrapping, with SNR$=30$ dB and $N = 8000$ ($a = 200$).

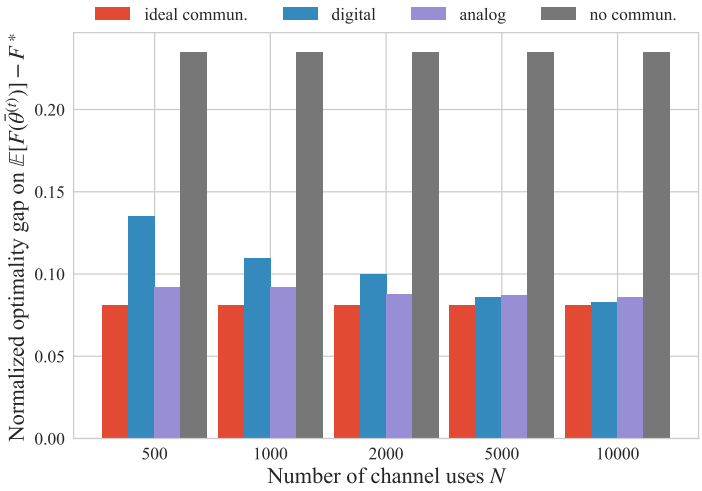
respectively, for the analog implementation. Fig. 8 compares digital and analog implementations, benchmarked by ideal communication and no communication. The analytical bounds shown in Fig. 8(a) are in general agreement with the empirical results shown in Fig. 8(b), demonstrating the practical relevance of the theory developed in this paper. In particular, the analysis correctly predicts the advantages of the analog implementation for sufficiently small number of channel uses N, and the marginal benefits of the digital implementation in the complementary regime of a large number of channel uses, e.g., when we have $N \geq 5000$.

VII. Conclusions

This paper has initiated studies on the communication-efficient implementations of DSGD algorithms for wireless FL in fully decentralized architectures. Specifically, we have proposed generic digital and analog transmission protocols tailored to FL over wireless D2D networks by enabling broadcasting for digital transmission, and both broadcasting and AirComp for analog transmission. We also adopted a practically favourable linear compression scheme, RLC, to reduce communication burden, and implemented a consensus step size adaptive to the training iteration. For both implementations, we developed rigorous analysis framework in term of their convergence properties, characterizing the impact of the connectivity topology, the quality of transmission, and/or the channel noise on the optimality gap. Empirical experiments on an image-



(a) Normalized upper bounds on the optimality gap for different schemes.



(b) Normalized empirical optimality gap for different schemes.

Fig. 8. Performance comparison of analog and digital implementations in terms of normalized (upper bounds on) optimality gap for iteration $t = 2500$ for different number N of channel uses with SNR set to 20 dB over a chain graph.

classification task verified the analytical results as well as the importance of an adaptive consensus step size.

There are several important research directions beyond the scope of this paper that may warrant more research. First, the channel-inversion based analog AirComp transmission can cause channel noise enhancement, which could be mitigated by designing optimal power control policies beyond channel inversion [25]. Secondly, it is also important to extend the results in this work to more general learning settings lifting the assumption of μ-L convexity. Moreover, it

would be interesting to apply more advanced (compression)-based gossip algorithms [18], [23] for decentralized wireless FL. It is worth emphasizing that the adaptations of these schemes to wireless FL in a D2D architecture would require novel design to cope with the presence of channel impairments. Last but not the least, to address the "stragglers" issues in decentralized wireless FL, effective device scheduling, possibly combined with asynchronous training, would be worth investigating along with the corresponding convergence analysis leveraging random graph theory [30].

APPENDIX A
ALGORITHM FOR DIGITAL IMPLEMENTATION

APPENDIX B
A SCHEDULING STRATEGY FOR THE ANALOG IMPLEMENTATION

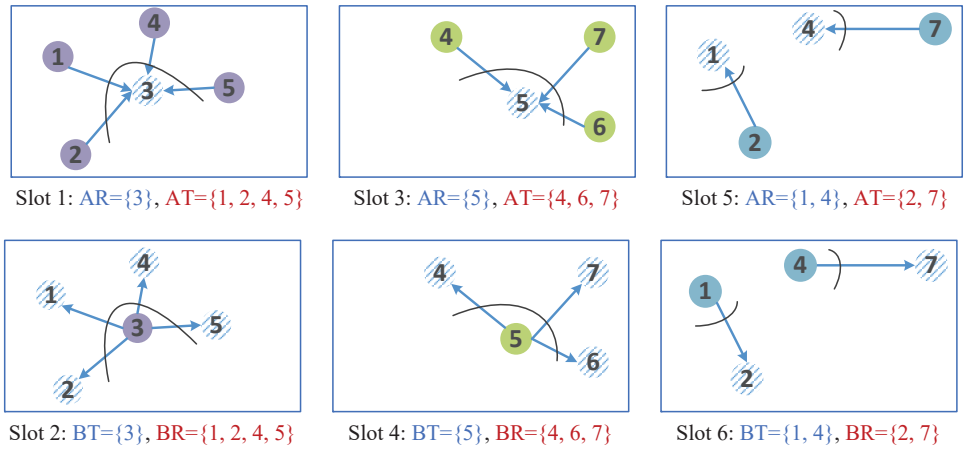


Fig. 9. The outcome of the scheduling policy in Algorithm 3, yields $M = 6$ transmission slots and the following sets: $\mathcal{S}_1^{\mathrm{AT}} = \{1\}$, $\mathcal{S}_1^{\mathrm{BT}} = \{6\}$, $\mathcal{S}_1^{\mathrm{AR}} = \{5\}$, $\mathcal{S}_1^{\mathrm{BR}} = \{2\}$; $\mathcal{S}_2^{\mathrm{AT}} = \{1, 5\}$, $\mathcal{S}_2^{\mathrm{BR}} = \{2, 6\}$; $\mathcal{S}_3^{\mathrm{BT}} = \{2\}$, $\mathcal{S}_3^{\mathrm{AR}} = \{1\}$; $\mathcal{S}_4^{\mathrm{AT}} = \{1, 3\}$, $\mathcal{S}_4^{\mathrm{BT}} = \{6\}$, $\mathcal{S}_4^{\mathrm{AR}} = \{5\}$, $\mathcal{S}_4^{\mathrm{BR}} = \{2, 4\}$; $\mathcal{S}_5^{\mathrm{AT}} = \{1\}$, $\mathcal{S}_5^{\mathrm{BT}} = \{4\}$, $\mathcal{S}_5^{\mathrm{AR}} = \{3\}$, $\mathcal{S}_5^{\mathrm{BR}} = \{2\}$; $\mathcal{S}_6^{\mathrm{AT}} = \{3\}$, $\mathcal{S}_6^{\mathrm{BR}} = \{4\}$; and $\mathcal{S}_7^{\mathrm{AT}} = \{3, 5\}$, $\mathcal{S}_7^{\mathrm{BR}} = \{4, 6\}$.

Algorithm 2: Digital Wireless Implementation

Input : Consensus step size $\zeta^{(t)}$, SGD learning step size $\eta^{(t)}$, connectivity graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and mixing matrix $\boldsymbol{W}$

Initialize at each node $i \in \mathcal{V}$: $\boldsymbol{\theta}_i^{(0)}$, $\hat{\boldsymbol{\theta}}_j^{(0)} = \mathbf{0}$, $\forall j \in \mathcal{N}_i \bigcup \{i\}$;

for $t = 0, 1, \ldots, T-1$ **do**

  for each device $i \in \mathcal{V}$ **do** in parallel

    update $\boldsymbol{\theta}_i^{(t+1/2)} = \boldsymbol{\theta}_i^{(t)} - \eta^{(t)} \hat{\nabla} f_i(\boldsymbol{\theta}_i^{(t)})$;

  end

  for slot $s = 1, \ldots, M$ **do**

    for each scheduled device i at slot s **do** in parallel

      broadcast $\mathcal{Q}_b\left(\boldsymbol{A}_i^{(t)}(\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)})\right)$;

    end

    for each neighboring device $j \in \mathcal{N}_i$ **do** in parallel

      receive $\mathcal{Q}_b\left(\boldsymbol{A}_j^{(t)}(\boldsymbol{\theta}_j^{(t+1/2)} - \hat{\boldsymbol{\theta}}_j^{(t)})\right)$;

    end

  end

  for each device $i \in \mathcal{V}$ **do** in parallel

    update (11) and (3).

  end

end

Output: $\boldsymbol{\theta}_i^{(T-1)}$, $\forall i \in \mathcal{V}$

APPENDIX C
ALGORITHM FOR ANALOG IMPLEMENTATION

APPENDIX D
PROOF OF LEMMA 4.1

First, as per the definition of the linear encoding matrix $\boldsymbol{A} = \frac{1}{\sqrt{m}}\boldsymbol{H}\boldsymbol{R}$, it follows that $\boldsymbol{A}\boldsymbol{A}^T = \frac{1}{m}\boldsymbol{H}\boldsymbol{R}\boldsymbol{R}\boldsymbol{H}^T = \frac{1}{m}\boldsymbol{H}\boldsymbol{H}^T = \frac{d}{m}\boldsymbol{I}$.[4] In addition, by denoting the lth row of the partial Hadamard

[4]We will omit the superscript $(\cdot)^{(t)}$ whenever it does not cause ambiguity throughout the appendices.

Algorithm 3: Scheduling Policy for Analog Transmission

Input : $n=0$, graph $\mathcal{G}^{(1)}=\mathcal{G}(\mathcal{V},\mathcal{E})$

Initialize for each node $i\in\mathcal{N}_i$: $\mathcal{S}_i^{\mathrm{AT}}=\mathcal{S}_i^{\mathrm{AR}}=\mathcal{S}_i^{\mathrm{BT}}=\mathcal{S}_i^{\mathrm{BR}}=\emptyset$;

repeat

  update $n=n+1$;

  color the auxilary graph associated with graph $\mathcal{G}^{(n)}$ as described in Section III-A;

  choose $c^*=\arg\max_c\{d_c^{(n)}\}$;

  Construct a set $\mathcal{N}_{c^*}^{(n)}$ of center nodes composed of all nodes in color c^*;

  foreach center node $i\in\mathcal{N}_{c^*}^{(n)}$ **do**

    update $\mathcal{S}_i^{\mathrm{AR}}\leftarrow\mathcal{S}_i^{\mathrm{AR}}\bigcup\{2n-1\}$;

    update $\mathcal{S}_i^{\mathrm{BT}}\leftarrow\mathcal{S}_i^{\mathrm{BT}}\bigcup\{2n\}$;

  end

  foreach center node i's neighbors $j\in\mathcal{N}_i$ in graph $\mathcal{G}^{(n)}$ **do**

    update $\mathcal{S}_j^{\mathrm{AT}}\leftarrow\mathcal{S}_j^{\mathrm{AT}}\bigcup\{2n-1\}$;

    update $\mathcal{S}_j^{\mathrm{BR}}\leftarrow\mathcal{S}_j^{\mathrm{BR}}\bigcup\{2n\}$;

  end

  update $\mathcal{V}=\mathcal{V}\setminus\left(\mathcal{N}_{c^*}^{(n)}\bigcup\{\text{ disconnected node from}\,\mathcal{G}^{(n)}\}\right)$;

  update $\mathcal{E}=\mathcal{E}\setminus\left(\bigcup_{i\in\mathcal{N}_{c^*}^{(n)}}\left(\bigcup_{j\in\mathcal{N}_i}\{(i,j)\}\right)\right)$;

  update graph $\mathcal{G}^{(n+1)}=\mathcal{G}(\mathcal{V},\mathcal{E})$.

until $\mathcal{G}^{(n)}=\emptyset$;

Output: $M=2n$, $\{\mathcal{S}_i^{\mathrm{AT}},\mathcal{S}_i^{\mathrm{AR}},\mathcal{S}_i^{\mathrm{BT}},\mathcal{S}_i^{\mathrm{BR}}\}$, $\forall i\in\mathcal{V}$

matrix $\boldsymbol{H}$ by $\boldsymbol{h}_l^T\in\mathbb{R}^{1\times d}$, $l=1,\ldots,m$, $\mathbb{E}[\boldsymbol{A}^T\boldsymbol{A}]$ is given by

$$\begin{aligned}\mathbb{E}[\boldsymbol{A}^T\boldsymbol{A}]&=\frac{1}{m}\mathbb{E}[\boldsymbol{R}[\boldsymbol{h}_1,\ldots,\boldsymbol{h}_m][\boldsymbol{h}_1,\ldots,\boldsymbol{h}_m]^T\boldsymbol{R}]\\&=\frac{1}{m}\mathbb{E}[[\boldsymbol{R}\boldsymbol{h}_1,\ldots,\boldsymbol{R}\boldsymbol{h}_m][\boldsymbol{R}\boldsymbol{h}_1,\ldots,\boldsymbol{R}\boldsymbol{h}_m]^T]\\&=\frac{1}{m}\sum_{l=1}^{m}\mathbb{E}[(\boldsymbol{R}\boldsymbol{h}_l)(\boldsymbol{R}\boldsymbol{h}_l)^T]\\&=\frac{1}{m}\sum_{l=1}^{m}\mathbb{E}\begin{bmatrix}r_1^2h_{l,1}^2&\ldots&r_1r_dh_{l,1}h_{l,d}\\\vdots&\ddots&\vdots\\r_1r_dh_{l,1}h_{l,d}&\ldots&r_d^2h_{l,d}^2\end{bmatrix}\overset{(a)}{=}\boldsymbol{I},\end{aligned}\tag{41}$$

Algorithm 4: Analog Wireless Implementation

Input : Consensus step size $\zeta^{(t)}$, SGD learning step size $\eta^{(t)}$, connectivity graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, mixing matrix $\boldsymbol{W}$, transmission slots $M = 2n$, and subsets $\{\mathcal{S}_i^{\rm BT}, \mathcal{S}_i^{\rm BR}, \mathcal{S}_i^{\rm AT}, \mathcal{S}_i^{\rm AR}\}$, $i \in \mathcal{V}$

Initialize at each node $i \in \mathcal{V}$: $\boldsymbol{\theta}_i^{(0)}$, $\hat{\boldsymbol{\theta}}_j^{(0)} = \mathbf{0}$, $\forall j \in \mathcal{N}_i \cup \{i\}$, and $\hat{\boldsymbol{y}}_i^{(0)} = \mathbf{0}$;

for $t = 0, 1, \ldots, T-1$ **do**

  perform Line 3 - 5 in Algorithm 2;

  for slot $s = 1, 3, \ldots, 2n-1$ **do** *AirComp transmission*

    for each node i in $\{i \in \mathcal{V} \mid s \in \mathcal{S}_i^{\rm AT}\}$ **do** in parallel

      transmit to a center node j (12);

    end

    for each center node j in $\{j \in \mathcal{V} \mid s \in \mathcal{S}_j^{\rm AR}\}$ **do** in parallel

      receive (13) and estimate (14);

    end

  end

  for slot $s = 2, 4, \ldots, 2n$ **do** *BC transmission*

    for each node i in $\{i \in \mathcal{V} \mid s \in \mathcal{S}_i^{\rm BT}\}$ **do** in parallel

      broadcast (15);

    end

    for each node j in $\{j \in \mathcal{V} \mid s \in \mathcal{S}_j^{\rm BR}\}$ **do** in parallel

      receive (16) and estimate (17) ;

    end

  end

  for each device $j \in \mathcal{V}$ **do** in parallel

    update (18), (24) and (25).

  end

end

Output: $\boldsymbol{\theta}_i^{(T-1)}$, $\forall i \in \mathcal{V}$

where $h_{l,n}$ denotes the nth entry of $\boldsymbol{h}_l$, $n = 1, \ldots, d$; and (a) is due to the fact that $r_n^2 h_{l,n}^2 = 1$ for all $l = 1, \ldots, m$, as well as $\mathbb{E}[r_{n_1} r_{n_2}] = 0$ for all $n_1 \neq n_2$. Then given the number of rows of $\boldsymbol{A}$, i.e., m, fixed, for all $\boldsymbol{u} \in \mathbb{R}^{d\times 1}$, on average over RLC, it follows that

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{u} - \frac{m}{d}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{u}\right\|^2 &= \mathbb{E}\left[\boldsymbol{u}^T\boldsymbol{u} - \frac{2m}{d}\boldsymbol{u}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{u} + \left(\frac{m}{d}\right)^2 \boldsymbol{u}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{u}\right] \\ &\overset{(a)}{=} \|\boldsymbol{u}\|^2 - \frac{m}{d}\boldsymbol{u}^T\mathbb{E}[\boldsymbol{A}^T\boldsymbol{A}]\boldsymbol{u} \\ &\overset{(b)}{=} \left(1 - \frac{m}{d}\right)\|\boldsymbol{u}\|^2, \end{aligned} \tag{42}$$

where (a) comes from $\boldsymbol{A}\boldsymbol{A}^T = \frac{d}{m}\boldsymbol{I}$ and (b) is obtained by plugging (41) into (a). As a result, denoting $\frac{m_i}{d}$ by ω_i, $i \in \mathcal{V}$, we have

$$\begin{aligned}&\mathbb{E}\left\|\boldsymbol{u} - \frac{m_i}{d}\boldsymbol{A}_i^T\mathcal{Q}_b(\boldsymbol{A}_i\boldsymbol{u})\right\|^2 = \mathbb{E}\left\|\left(\boldsymbol{u} - \frac{m_i}{d}\boldsymbol{A}_i^T\boldsymbol{A}_i\boldsymbol{u}\right) + \left(\frac{m_i}{d}\boldsymbol{A}_i^T\boldsymbol{A}_i\boldsymbol{u} - \frac{m_i}{d}\boldsymbol{A}_i^T\mathcal{Q}_b(\boldsymbol{A}_i\boldsymbol{u})\right)\right\|^2 \\ &\overset{(a)}{=} \mathbb{E}\left[\left\|\boldsymbol{u} - \frac{m_i}{d}\boldsymbol{A}_i^T\boldsymbol{A}_i\boldsymbol{u}\right\|^2 + 2\frac{m_i}{d}\left(\boldsymbol{u} - \frac{m_i}{d}\boldsymbol{A}_i^T\boldsymbol{A}_i\boldsymbol{u}\right)^T\boldsymbol{A}_i^T\boldsymbol{Q}_b\boldsymbol{A}_i\boldsymbol{u} + \frac{m_i^2}{d^2}\left\|\boldsymbol{A}_i^T\boldsymbol{Q}_b\boldsymbol{A}_i\boldsymbol{u}\right\|^2\right] \\ &\overset{(b)}{=} \mathbb{E}\left\|\boldsymbol{u} - \frac{m_i}{d}\boldsymbol{A}_i^T\boldsymbol{A}_i\boldsymbol{u}\right\|^2 + \frac{m_i^2}{d^2}\mathbb{E}\left\|\boldsymbol{A}_i^T\boldsymbol{Q}_b\boldsymbol{A}_i\boldsymbol{u}\right\|^2 \\ &\overset{(c)}{=} \left(1 - \frac{m_i}{d} + \frac{m_i}{d}\epsilon_b^2\right)\|\boldsymbol{u}\|^2 \\ &\overset{(d)}{\approx} \left(1 - \frac{m_i}{d}\right)\|\boldsymbol{u}\|^2 \\ &\leq (1 - \min_{i\in\mathcal{V}}\omega_i)\|\boldsymbol{u}\|^2 \end{aligned} \tag{43}$$

where (a) is because of $\boldsymbol{A}_i\boldsymbol{u} - \mathcal{Q}_b(\boldsymbol{A}_i\boldsymbol{u}) = \boldsymbol{Q}_b\boldsymbol{A}\boldsymbol{u}$ in accordance with IEEE 754 standard, in which $\boldsymbol{Q}_b \in \mathbb{R}^{m_i\times m_i}$ is a diagonal matrix with its ith diagonal entry satisfying $|[\boldsymbol{Q}_b]_{kk}| < \epsilon_b$, $k = 1, \ldots, m_i$; (b) is due to the fact $\boldsymbol{A}_i\boldsymbol{A}_i^T = \frac{d}{m_i}\boldsymbol{I}$; and (c) follows from (42) and $\mathbb{E}[\boldsymbol{A}_i^T(\boldsymbol{Q}_b)^2\boldsymbol{A}_i] = \frac{1}{m_i}\sum_{k=1}^{m_i}|[\boldsymbol{Q}_b]_{kk}|^2\boldsymbol{I} \leq \epsilon_b^2\boldsymbol{I}$. In addition, considering very little value of ϵ_b in common double-precision floating number systems ($\epsilon_b = 2^{-24}$ for $b = 32$ and $\epsilon_b = 2^{-53}$ for $b = 52$) [42, ch. 2], we safely approximate (c) by (d). Lemma 4.1 is thus proved.

APPENDIX E
PROOF OF COROLLARY 4.1

Note that (43) implies $m_i \leq d$, for all $i \in \mathcal{V}$. We thus assume that when the channel condition can support more than d rows for the linear encoding matrix $\boldsymbol{A}_i$, $\boldsymbol{A}_i \in \mathbb{R}^{d\times d}$ is kept. That said, the MSE corresponding to cases when $m_i \geq d$ are all equal to zero. As a result, when $m_i^{(t)}$ varies

with channel state at the tth iteration, on average over RLC and the Rayleigh-fading channels, according to the law of iterated expectation, we have

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{u}-\frac{m_i}{d}\boldsymbol{A}_i^T\mathcal{Q}_b(\boldsymbol{A}_i\boldsymbol{u})\right\|^2 &= \mathbb{E}_{m_i}\left[\mathbb{E}_{\mathrm{RLC}}\left[\left\|\boldsymbol{u}-\frac{m_i}{d}\boldsymbol{A}_i^T\mathcal{Q}_b(\boldsymbol{A}_i\boldsymbol{u})\right\|^2\middle| m_i\right]\right]\\ &\overset{(43)}{\approx} \mathbb{E}_{m_i<d}\left[\left(1-\frac{m_i}{d}\right)\right]\|\boldsymbol{u}\|^2\\ &= \sum_{n=0}^{d}\Pr(m_i=n)\left(1-\frac{m_i}{d}\right). \end{aligned} \tag{44}$$

Next, it remains to calculate $\Pr(m_i=n)$ for $n=0,\ldots,d-1$, $i\in\mathcal{V}$. For $n=0$, it follows that

$$\begin{aligned}\Pr(m_i=0) &= \Pr(B_i/b<1)\\ &= \Pr\left(\min_{j\in\mathcal{N}_i}|h'_{ij}|^2<\frac{N_0}{PM}\left(2^{b\frac{M}{N}}-1\right)\right)\\ &\overset{(a)}{=} 1-\exp\left(-\frac{N_0}{P}\frac{1}{MA_0}\left(2^{b\frac{M}{N}}-1\right)\sum_{j\in\mathcal{N}_i}\left(\frac{d_{ij}}{d_0}\right)^r\right)\\ &\overset{(28)}{=} 1-G_i(1), \end{aligned} \tag{45}$$

where (a) follows from the fact that $|h'_{ij}|^2$, $j\in\mathcal{N}_i$, are exponentially and independently distributed with mean value $A_0(\frac{d_0}{d_{ij}})^r$ such that $\Pr(\min_{j\in\mathcal{N}_i}|h'_{ij}|^2<x)=1-\exp(-\frac{x}{A_0}\sum_{j\in\mathcal{N}_i}(\frac{d_{ij}}{d_0})^r)$. Moreover, for $n=1,\ldots,d-1$, it is easy to show that $\Pr(m_i=n)=\Pr(n\le B_i/b<n+1)=G_i(n)-G_i(n+1)$. To sum up, the probability mass function (PMF) of m_i, $i\in\mathcal{V}$, is given by

$$\Pr(m_i=n)=\begin{cases}1-G_i(1), & n=0,\\ G_i(n)-G_i(n+1), & n=1,\ldots,d-1,\\ G_i(d), & n\ge d.\end{cases} \tag{46}$$

Finally, Plugging (46) into (44), we arrive at

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{u}-\frac{m_i}{d}\boldsymbol{A}_i^T\mathcal{Q}_b(\boldsymbol{A}_i\boldsymbol{u})\right\|^2 &\approx \sum_{n=0}^{d}\Pr(m_i=n)\left(1-\frac{m_i}{d}\right)\\ &= 1-G_i(1)+\sum_{n=1}^{d-1}(G_i(n)-G_i(n+1))\left(1-\frac{n}{d}\right), \end{aligned} \tag{47}$$

which, after some manipulation employing telescoping sums, implies (27) in Corollary 4.1.

APPENDIX F
PROOF OF LEMMA 5.1

We substitute (13) and (16) into (14) and (17), respectively, and rewrite (18) as

$$\hat{\boldsymbol{y}}_j^{(t+1)} = \hat{\boldsymbol{y}}_j^{(t)} + \frac{m}{d}(\boldsymbol{A}^{(t)})^T\boldsymbol{A}^{(t)}\Big(\sum_{s\in\mathcal{S}_j^{\mathrm{AR}}}\sum_{i\in\mathcal{N}_j^{(s)}} w_{ji}(\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)}) + \sum_{s\in\mathcal{S}_j^{\mathrm{BR}}} w_{ji_s}(\boldsymbol{\theta}_{i_s}^{(t+1/2)} - \hat{\boldsymbol{\theta}}_{i_s}^{(t)})\Big) + \frac{m}{d}(\boldsymbol{A}^{(t)})^T\tilde{\boldsymbol{n}}_j^{(t)}, \quad (48)$$

in which $\tilde{\boldsymbol{n}}_j^{(t)}$ is the effective noise defined as

$$\tilde{\boldsymbol{n}}_j^{(t)} = \sum_{s\in\mathcal{S}_j^{\mathrm{AR}}} \frac{\Re\{\boldsymbol{n}_j^{(t,s)}\}}{\sqrt{\gamma_j^{(t,s)}}} + \sum_{s\in\mathcal{S}_j^{\mathrm{BR}}} \frac{w_{ji_s}}{\sqrt{\alpha_{i_s}^{(t,s)}}}\Re\left\{\frac{\boldsymbol{n}_j^{(t,s)}}{h_{i_sj}^{\prime(t)}}\right\}, \quad (49)$$

which is also Gaussian with zero mean and the covariance matrix calculated as

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{n}}_j^{(t)}(\tilde{\boldsymbol{n}}_j^{(t)})^T] &= \sum_{s\in\mathcal{S}_j^{\mathrm{AR}}}\frac{1}{\gamma_j^{(t,s)}}\mathbb{E}[\Re\{\boldsymbol{n}_j^{(t,s)}\}(\Re\{\boldsymbol{n}_j^{(t,s)}\})^T] + \sum_{s\in\mathcal{S}_j^{\mathrm{BR}}}\frac{w_{ji_s}^2}{\alpha_{i_s}^{(t,s)}}\mathbb{E}\left[\Re\Big\{\frac{\boldsymbol{n}_j^{(t,s)}}{h_{i_sj}^{\prime(t)}}\Big\}\Big(\Re\Big\{\frac{\boldsymbol{n}_j^{(t,s)}}{h_{i_sj}^{\prime(t)}}\Big\}\Big)^T\right]\\ &\overset{(a)}{=}\frac{N_0}{2}\Bigg(\sum_{s\in\mathcal{S}_j^{\mathrm{AR}}}\frac{1}{\gamma_j^{(t,s)}} + \sum_{s\in\mathcal{S}_j^{\mathrm{BR}}}\frac{w_{ji_s}^2}{\alpha_{i_s}^{(t,s)}\big|h_{i_sj}^{\prime(t)}\big|^2}\Bigg)\boldsymbol{I}, \end{aligned} \quad (50)$$

where (a) is due to the facts that the entries $\boldsymbol{n}_j^{(t,s)}$ are $i.i.d.$ with each denoted by $n_{ji}^{(t,s)} \sim \mathcal{CN}(0, N_0)$, and the real and the image part of $n_{ji}^{(t,s)}$ are also independent Gaussian random variables with zero mean and variance $\frac{N_0}{2}$, thus leading to $\sum_{s\in\mathcal{S}_j^{\mathrm{BR}}}\frac{w_{ji_s}^2}{\alpha_{i_s}^{(t,s)}}\mathbb{E}[\Re\{\frac{\boldsymbol{n}_j^{(t,s)}}{h_{i_sj}^{\prime(t)}}\}(\Re\{\frac{\boldsymbol{n}_j^{(t,s)}}{h_{i_sj}^{\prime(t)}}\})^T] = \frac{N_0}{2|h_{i_sj}^{\prime(t)}|^2}\boldsymbol{I}$. Then specifying $\alpha_{i_s}^{(t,s)}$ and $\gamma_j^{(t,s)}$ by (21) and (20), respectively, we obtain (32).

Furthermore, the scheduling scheme in 3 suggests that each node is scheduled to receive as a center node for *at most one* slot (cf. Line 13 in Algorithm 3), i.e., $|\mathcal{S}_j^{\mathrm{AR}}| \le 1$, while it may be scheduled to receive from a broadcast Tx for *multiple* slots. In all cases, the estimate update (48) always aggregates model parameters from all of node j's neighbors. Hence, (48) can be simplified as

$$\hat{\boldsymbol{y}}_j^{(t+1)} = \hat{\boldsymbol{y}}_j^{(t)} + \frac{m}{d}(\boldsymbol{A}^{(t)})^T\boldsymbol{A}^{(t)}\sum_{i\in\mathcal{N}_j} w_{ji}\boldsymbol{u}_i^{(t)} + \frac{m}{d}(\boldsymbol{A}^{(t)})^T\tilde{\boldsymbol{n}}_j^{(t)}, \quad (51)$$

where $\boldsymbol{\theta}_i^{(t+1/2)} - \hat{\boldsymbol{\theta}}_i^{(t)} \triangleq \boldsymbol{u}_i^{(t)}$. By recursively applying (51), it follows, for any device $i \in \mathcal{V}$, that

$$\hat{\boldsymbol{y}}_i^{(t+1)} = \hat{\boldsymbol{y}}_i^{(0)} + \sum_{\tau=0}^{t}\left(\frac{m}{d}(\boldsymbol{A}^{(\tau)})^T\boldsymbol{A}^{(\tau)}\sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{u}_j^{(\tau)}\right) + \sum_{\tau=0}^{t}\frac{m}{d}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{n}}_i^{(\tau)}. \quad (52)$$

On the other hand, applying weighted sum over $j \in \mathcal{N}_i$ on both sides of (24), it follows that

$$\sum_{j\in\mathcal{N}_i} w_{ij}\hat{\boldsymbol{\theta}}_j^{(t+1)} = \sum_{j\in\mathcal{N}_i} w_{ij}\hat{\boldsymbol{\theta}}_j^{(t)} + \frac{m}{d}(\boldsymbol{A}^{(t)})^T\boldsymbol{A}^{(t)}\sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{u}_j^{(t)}. \tag{53}$$

Recursively applying (53) leads to

$$\sum_{j\in\mathcal{N}_i} w_{ij}\hat{\boldsymbol{\theta}}_j^{(t+1)} = \sum_{j\in\mathcal{N}_i} w_{ij}\hat{\boldsymbol{\theta}}_i^{(0)} + \sum_{\tau=0}^{t}\left(\frac{m}{d}(\boldsymbol{A}^{(\tau)})^T\boldsymbol{A}^{(\tau)}\sum_{j\in\mathcal{N}_i} w_{ij}\boldsymbol{u}_j^{(\tau)}\right). \tag{54}$$

Then, combining (52) and (54) with the fact $\hat{\boldsymbol{y}}_i^{(0)} = \sum_{j\in\mathcal{N}_i} w_{ij}\hat{\boldsymbol{\theta}}_i^{(0)} = \mathbf{0}$, we have

$$\hat{\boldsymbol{y}}_i^{(t+1)} = \sum_{j\in\mathcal{N}_i} w_{ij}\hat{\boldsymbol{\theta}}_j^{(t+1)} + \sum_{\tau=0}^{t}\frac{m}{d}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{n}}_i^{(\tau)}. \tag{55}$$

By substituting (55) for $\hat{\boldsymbol{y}}_i^{(t+1)}$ in device i's consensus update for analog implementation, (25) can be recast as (31).

APPENDIX G
PROOF OF LEMMA 5.2

To prove Lemma 5.2, we first provide variants of [20, Lemma 17 and Lemma 18], where matrix notations are used for the simplicity of notation.

Lemma G.1 (Variant of [20, Lemma 17]): Denoting $[\boldsymbol{\theta}_1^{(t)},\ldots,\boldsymbol{\theta}_K^{(t)}] \in \mathbb{R}^{d\times K}$, $[\hat{\boldsymbol{\theta}}_1^{(t)},\ldots,\hat{\boldsymbol{\theta}}_K^{(t)}] \in \mathbb{R}^{d\times K}$, $[\bar{\boldsymbol{\theta}}^{(t)},\ldots,\bar{\boldsymbol{\theta}}^{(t)}] \in \mathbb{R}^{d\times K}$ and $[\tilde{\boldsymbol{n}}_1^{(t)},\ldots,\tilde{\boldsymbol{n}}_K^{(t)}] \in \mathbb{R}^{d\times K}$, by $\boldsymbol{\Theta}^{(t)}$, $\hat{\boldsymbol{\Theta}}^{(t)}$, $\bar{\boldsymbol{\Theta}}^{(t)}$ and $\tilde{\boldsymbol{N}}^{(t)}$, respectively, then for consensus step size $\zeta^{(t)} \geq 0$, mixing matrix $\boldsymbol{W}$ and any parameter $\alpha_1 > 0$, on average over RLC and AWGN, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\Theta}^{(t+1)} - \bar{\boldsymbol{\Theta}}^{(t+1)}\|_F^2 \leq (1+\alpha_1)\left(1-\delta\zeta^{(t)}\right)^2\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)} - \bar{\boldsymbol{\Theta}}^{(t+1/2)}\|_F^2 + \\ (1+\alpha_1^{-1})\beta^2(\zeta^{(t)})^2\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)} - \hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2 + (\zeta^{(t)})^2\omega^2 d\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)},\end{aligned} \tag{56}$$

where $\tilde{N}_0^{(\tau)} = \sum_{i\in\mathcal{V}}\tilde{N}_{0i}^{(\tau)}$ is the sum-variance of effective noise over all devices.

Proof: With the fact $(\boldsymbol{W}-\boldsymbol{I})\mathbf{1}\mathbf{1}^T/K=\mathbf{0}$, rewrite (31) and (33) in matrix forms as $\boldsymbol{\Theta}^{(t+1)}=\boldsymbol{\Theta}^{(t+1/2)}+\zeta^{(t)}\hat{\boldsymbol{\Theta}}^{(t+1)}(\boldsymbol{W}-\boldsymbol{I})+\zeta^{(t)}\omega\sum_{\tau=0}^{t}(A^{(\tau)})^T\tilde{\boldsymbol{N}}^{(\tau)}$ and $\bar{\boldsymbol{\Theta}}^{(t+1)}=\bar{\boldsymbol{\Theta}}^{(t+1/2)}+\zeta^{(t)}\omega\sum_{\tau=0}^{t}(A^{(\tau)})^T\tilde{\boldsymbol{N}}^{(\tau)}\mathbf{1}\mathbf{1}^T/K$, respectively. Then, we have

$$
\begin{aligned}
&\mathbb{E}\|\boldsymbol{\Theta}^{(t+1)}-\bar{\boldsymbol{\Theta}}^{(t+1)}\|_F^2\\
=&\mathbb{E}\Big\|\boldsymbol{\Theta}^{(t+1/2)}-\bar{\boldsymbol{\Theta}}^{(t+1/2)}+\zeta^{(t)}\hat{\boldsymbol{\Theta}}^{(t+1)}(\boldsymbol{W}-\boldsymbol{I})+\zeta^{(t)}\omega\sum_{\tau=0}^{t}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{N}}^{(\tau)}\Big(\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^T}{K}\Big)\Big\|_F^2\\
\overset{(a)}{=}&\mathbb{E}\Big\|\boldsymbol{\Theta}^{(t+1/2)}-\bar{\boldsymbol{\Theta}}^{(t+1/2)}+\zeta^{(t)}\hat{\boldsymbol{\Theta}}^{(t+1)}(\boldsymbol{W}-\boldsymbol{I})\Big\|_F^2+(\zeta^{(t)})^2\omega^2\mathbb{E}\Big\|\sum_{\tau=0}^{t}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{N}}^{(\tau)}\Big(\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^T}{K}\Big)\Big\|_F^2\\
\overset{(b)}{\leq}&\mathbb{E}\Big\|(\boldsymbol{\Theta}^{(t+1/2)}-\bar{\boldsymbol{\Theta}}^{(t+1/2)})(\boldsymbol{I}+\zeta^{(t)}(\boldsymbol{W}-\boldsymbol{I}))+\zeta^{(t)}(\hat{\boldsymbol{\Theta}}^{(t+1)}-\boldsymbol{\Theta}^{(t+1/2)})(\boldsymbol{W}-\boldsymbol{I})\Big\|_F^2\\
&+(\zeta^{(t)})^2\omega^2 d\left(1-\frac{1}{K}\right)\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}\overset{(c)}{\leq}\ \text{RHS of (56)},
\end{aligned}
$$

where (a) is due to the expectation taken over Gaussian noise and RLC, i.e., $\mathbb{E}[(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{N}}^{(\tau)}]$, is zero; the first term in (b) is as a result of $\bar{\boldsymbol{X}}(\boldsymbol{W}-\boldsymbol{I})=\mathbf{0}$ for any $\boldsymbol{X}\in\mathbb{R}^{d\times K}$; the second term in (b) is based on $\boldsymbol{A}^{(\tau)}(\boldsymbol{A}^{(\tau)})^T=\frac{d}{m}\boldsymbol{I}$ and $i.i.d.$ entries of $\tilde{\boldsymbol{n}}_i^{(\tau)}$ such that $\mathbb{E}\|\tilde{\boldsymbol{n}}_i^{(\tau)}\|^2=m\tilde{N}_{0i}^{(\tau)}$ with $\sum_{i\in\mathcal{V}}\tilde{N}_{0i}^{(\tau)}\triangleq\tilde{N}_0^{(\tau)}$; and (c) follows [20, Lemma 17]. ■

Lemma G.2 (Variant of [20, Lemma 18]): In addition to $\boldsymbol{\Theta}^{(t)}$, $\hat{\boldsymbol{\Theta}}^{(t)}$, $\bar{\boldsymbol{\Theta}}^{(t)}$ and $\tilde{\boldsymbol{N}}^{(t)}$ denoted as in Lemma G.1, denoting $[\hat{\nabla}f_1(\boldsymbol{\theta}_1^{(t)}),\ldots,\hat{\nabla}f_K(\boldsymbol{\theta}_K^{(t)})]\in\mathbb{R}^{d\times K}$ by $\hat{\nabla}F^{(t)}$, then for consensus step size $\zeta^{(t)}\geq 0$, mixing matrix $\boldsymbol{W}$ and any parameter $\alpha_2>0$, on average over RLC and AWGN, we have

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{\Theta}^{(t+3/2)}-\hat{\boldsymbol{\Theta}}^{(t+2)}\|_F^2\leq&(1+\alpha_2^{-1})(1-\omega)\beta^2(\zeta^{(t)})^2\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}-\bar{\boldsymbol{\Theta}}^{(t+1/2)}\|_F^2+\\
&(1+\alpha_2)(1-\omega)\left(1+\beta\zeta^{(t)}\right)^2\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}-\hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2+\\
&(\zeta^{(t)})^2(1-\omega)\omega^2 d\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}.
\end{aligned}
\tag{57}
$$

Proof: Given the constant $\omega=m/d$ in analog implementation, (24) implies that, for $i\in\mathcal{V}$,

$$
\hat{\boldsymbol{\Theta}}_i^{(t+1)}=\hat{\boldsymbol{\Theta}}_i^{(t)}+\omega(\boldsymbol{A}^{(t)})^T\boldsymbol{A}^{(t)}(\boldsymbol{\Theta}_i^{(t+1/2)}-\hat{\boldsymbol{\Theta}}_i^{(t)}).
\tag{58}
$$

Therefore, substituting (58) for $\hat{\boldsymbol{\Theta}}^{(t+2)}$, if follows that

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{\Theta}^{(t+3/2)}-\hat{\boldsymbol{\Theta}}^{(t+2)}\|_F^2=\mathbb{E}\left\|(\boldsymbol{\Theta}^{(t+3/2)}-\hat{\boldsymbol{\Theta}}_i^{(t+1)})-\omega(\boldsymbol{A}^{(t+1)})^T\boldsymbol{A}^{(t+1)}(\boldsymbol{\Theta}_i^{(t+3/2)}-\hat{\boldsymbol{\Theta}}_i^{(t+1)})\right\|_F^2\\
&\overset{(a)}{=}(1-\omega)\mathbb{E}\|\boldsymbol{\Theta}^{(t+3/2)}-\hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2\overset{(b)}{=}(1-\omega)\mathbb{E}\|\boldsymbol{\Theta}^{(t+1)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}-\hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2\\
&\overset{(c)}{=}(1-\omega)\mathbb{E}\left\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}+\zeta^{(t)}\hat{\boldsymbol{\Theta}}^{(t+1)}(\boldsymbol{W}-\boldsymbol{I})-\hat{\boldsymbol{\Theta}}^{(t+1)}+\zeta^{(t)}\omega\sum_{\tau=0}^{t}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{N}}^{(\tau)}\right\|_F^2\\
&\overset{(d)}{\leq}(1-\omega)\Big(\mathbb{E}\big\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}+\zeta^{(t)}\hat{\boldsymbol{\Theta}}^{(t+1)}(\boldsymbol{W}-\boldsymbol{I})-\hat{\boldsymbol{\Theta}}^{(t+1)}\big\|_F^2+\\
&\quad\mathbb{E}\big\|\zeta^{(t)}\omega\sum_{\tau=0}^{t}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{N}}^{(\tau)}\big\|_F^2\Big)\\
&\overset{(e)}{\leq}(1-\omega)\Big((1+\alpha_2^{-1})\beta^2(\zeta^{(t)})^2\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}-\bar{\boldsymbol{\Theta}}^{(t+1/2)}\|_F^2+\\
&\quad(1+\alpha_2)\left(1+\beta\zeta^{(t)}\right)^2\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}-\hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2+(\zeta^{(t)})^2\omega^2 d\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}\Big),
\end{aligned}$$

where (a) is based on (9); (b) executes one SGD step (cf. (1)); (c) is due to the analog consensus update (cf. (31)); (d) is as a results of zero-mean AWGN; and (e) follows [20, Lemma 18] combining with the noise variance as derived similarly in the proof for Lemma G.2. ■

We elaborate on non-trivial modifications made based on [20, Appendix C] in the sequel.[5] First, we need to modify two auxiliary functions to be adaptive as below:

$$\eta_1(\zeta^{(t)})=(1+\alpha_1)(1-\delta\zeta^{(t)})^2+(1+\alpha_2^{-1})(1-\omega)\beta^2(\zeta^{(t)})^2, \tag{59}$$

$$\xi_1(\zeta^{(t)})=(1+\alpha_1^{-1})\beta^2(\zeta^{(t)})^2+(1+\alpha_2)(1-\omega)\left(1+\beta\zeta^{(t)}\right)^2, \tag{60}$$

and introduce another auxiliary function defined as

$$\eta_2(x)=\left(\frac{\delta^2}{4}+\frac{2}{\omega}\beta^2\right)x^2-\delta x+1=1-\tilde{p}(x), \tag{61}$$

where

$$\tilde{p}(x)=\delta x-\left(\frac{\delta^2}{4}+\frac{2}{\omega}\beta^2\right)x^2. \tag{62}$$

According to the definition of function $\eta_2(x)$, which is a quadratic convex function decreasing over $[0,x^*]$ with $x^*=\min_x\eta_2(x)$, it follows that $\tilde{p}(x)$ increases over $[0,x^*]$.

[5]We will omit the dependence of functions on δ and ω but the iteration index t throughout the appendices, as long as it does not cause any ambiguity in the context.

Next, we substitute $\zeta^{(t)} = \frac{\zeta_0}{\sqrt[4]{\tilde{N}_0 t/a'+1}}$ for x in $\tilde{p}(x)$ to obtain function $\tilde{p}^{(t)} : \mathbb{R}_+ \mapsto \mathbb{R}$ in (36). Note that since ζ_0 given in Theorem 4.2 is given by $\zeta_0 = \lambda' x^*$ for a specific choice $\lambda' = \frac{8\beta^2+\delta^2\omega}{2(16\delta+\delta^2+4\beta^2+2\delta\beta^2-8\delta\omega)} \in (0,1]$ (see [20, (20), (24)] for detail), it follows that $\zeta^{(t)} \leq \zeta^{(0)} = \zeta_0 \leq x^*$. As we establish the facts that $\zeta^{(t)} \in [0, x^*]$ decreases over $t \geq 0$ and $\tilde{p}(x)$ increases over $[0, x^*]$, $\tilde{p}^{(t)}$ turns out to be decreasing over $t \geq 0$ according to the composition of monotonic functions. Furthermore, based on the relation of $\eta_1(x) \leq \eta_2(x)$ derived in [20, Apppendix C], which holds for any x and $\omega > 0$, since

$$\eta_1(\zeta_0) \leq \eta_2(\zeta_0) = 1 - \tilde{p}(\zeta_0) \overset{(36)}{=} 1 - \tilde{p}^{(0)} \overset{[20,\ (24)]}{\leq} 1 - p(\delta, \omega), \tag{63}$$

it implies that $p(\delta, \omega) \leq \tilde{p}^{(0)}$.

In addition, we have

$$\eta_1(\zeta^{(t)}) \leq \eta_2(\zeta^{(t)}) = 1 - \tilde{p}(\zeta^{(t)}) \overset{(36)}{=} 1 - \tilde{p}^{(t)}, \tag{64}$$

and

$$\xi_1(\zeta^{(t)}) \overset{(a)}{\leq} \xi_1(\zeta_0) \overset{[20,\ (26)]}{\leq} 1 - p(\delta, \omega), \tag{65}$$

where (a) is due to the fact that function $\xi_1(\zeta^{(t)})$ increases over $\zeta^{(t)} > 0$ and $\zeta^{(t)} \leq \zeta_0$. Finally, combining (64) and (65), we arrive at the following variant of [20, (21)]

$$\max\left\{\eta_1(\zeta^{(t)}), \xi_1(\zeta^{(t)})\right\} \leq 1 - \min\left\{\tilde{p}^{(t)}, p(\delta, \omega)\right\} \overset{(35)}{=} 1 - p^{(t)}. \tag{66}$$

Now, we are ready to derive the recursive upper-bound given by (38). The following inequalities are frequently recalled in the sequel. For given matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ of the same size, and a square matrix $\boldsymbol{Z}$,

$$\|\boldsymbol{X} + \boldsymbol{Y}\|_F^2 \leq (1+\alpha)\|\boldsymbol{X}\|_F^2 + (1+\alpha^{-1})\|\boldsymbol{Y}\|_F^2, \ \forall \alpha > 0, \tag{67}$$

and

$$\|\boldsymbol{X}\boldsymbol{Z}\|_F \leq \|\boldsymbol{X}\|_F \|\boldsymbol{Z}\|_2. \tag{68}$$

By applying one SGD step (cf. (1)), it follows that, for any given $\alpha_3 > 0$,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)} - \bar{\boldsymbol{\Theta}}^{(t+1/2)}\|_F^2 &= \mathbb{E}\left\|(\boldsymbol{\Theta}^{(t)} - \bar{\boldsymbol{\Theta}}^{(t)}) - \eta^{(t)}\hat{\nabla}F^{(t)}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{K}\right)\right\|_F^2 \\
&\overset{(67),(68)}{\leq} (1+\alpha_3^{-1})\mathbb{E}\|\boldsymbol{\Theta}^{(t)} - \bar{\boldsymbol{\Theta}}^{(t)}\|_F^2 + (1+\alpha_3)(\eta^{(t)})^2\mathbb{E}\|\hat{\nabla}F^{(t)}\|_F^2\left\|\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{K}\right\| \\
&\overset{(7)}{\leq} (1+\alpha_3^{-1})\mathbb{E}\|\boldsymbol{\Theta}^{(t)} - \bar{\boldsymbol{\Theta}}^{(t)}\|_F^2 + (1+\alpha_3)K(\eta^{(t)})^2G^2.
\end{aligned} \tag{69}$$

Similarly, for the same choice of α_3, it is also true that

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}-\bar{\boldsymbol{\Theta}}^{(t+1/2)}\|_F^2\\
&\overset{(67)}{\leq}(1+\alpha_3^{-1})\mathbb{E}\|\boldsymbol{\Theta}^{(t)}-\bar{\boldsymbol{\Theta}}^{(t)}\|_F^2+(1+\alpha_3)\mathbb{E}\left\|\eta^{(t)}\hat{\nabla}F^{(t)}\left(\boldsymbol{I}-\frac{\mathbf{1}\mathbf{1}^T}{K}\right)+\eta^{(t+1)}\hat{\nabla}F^{(t+1)}\right\|_F^2\\
&\overset{(7)}{\leq}(1+\alpha_3^{-1})\mathbb{E}\|\boldsymbol{\Theta}^{(t)}-\bar{\boldsymbol{\Theta}}^{(t)}\|_F^2+6(1+\alpha_3)K(\eta^{(t)})^2G^2,
\end{aligned}\tag{70}$$

and

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\Theta}^{(t+1/2)}-\eta^{(t+1)}\hat{\nabla}F^{(t+1)}-\hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2 \overset{(67),(7)}{\leq}&(1+\alpha_3^{-1})\|\boldsymbol{\Theta}^{(t+1/2)}-\hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2+\\
&(1+\alpha_3)K(\eta^{(t)})^2G^2.
\end{aligned}\tag{71}$$

Then, combining (56) and (57), where relevant terms are substituted by (69)-(71), after some manipulations, we have

$$\begin{aligned}
e^{(t+1)}\overset{(34)}{\leq}&\ \eta_1(\zeta^{(t)})(1+\alpha_3^{-1})\mathbb{E}\|\boldsymbol{\Theta}^{(t)}-\bar{\boldsymbol{\Theta}}^{(t)}\|_F^2+\xi_1(\zeta^{(t)})(1+\alpha_3^{-1})\|\boldsymbol{\Theta}^{(t+1/2)}-\hat{\boldsymbol{\Theta}}^{(t+1)}\|_F^2+\\
&(\eta_1(\zeta^{(t)})+\xi_1(\zeta^{(t)}))(1+\alpha_3)(\eta^{(t)})^2 6KG^2+(\zeta^{(t)})^2(2-\omega)\omega^2 d\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}\\
\leq&\underbrace{\max\left\{\eta_1(\zeta^{(t)}),\xi_1(\zeta^{(t)})\right\}(1+\alpha_3^{-1})e^{(t)}+12\max\left\{\eta_1(\zeta^{(t)}),\xi_1(\zeta^{(t)})\right\}(1+\alpha_3)(\eta^{(t)})^2KG^2}_{\text{part I}}+\\
&\underbrace{(\zeta^{(t)})^2(2-\omega)\omega^2 d\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}}_{\text{part II}}.
\end{aligned}\tag{72}$$

Note that by using (66) and choosing $\alpha_1=\frac{\delta\zeta^{(t)}}{2}$, $\alpha_2=\frac{\omega}{2}$ and $\alpha_3=\frac{p^{(t)}}{2}$ [20, (20), Lemma 21], part I in (72) proves to be $(1-\frac{p^{(t)}}{2})e^{(t)}+\frac{2}{p^{(t)}}(\eta^{(t)})^2 12KG^2$. Furthermore, looking into part II, we have

$$\begin{aligned}
(\zeta^{(t)})^2\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)}\overset{(a)}{\leq}&\frac{(\zeta^{(t)})^2}{p^{(t)}}p^{(t)}t\tilde{N}_0\overset{(b)}{\leq}\frac{(\zeta^{(t)})^2}{p^{(t)}}\tilde{p}^{(t)}\left(\sqrt[4]{\tilde{N}_0}t/a'\right)a'\tilde{N}_0^{\frac{3}{4}}\\
\overset{(c)}{\leq}&\frac{\delta\zeta_0}{p^{(t)}}(\zeta^{(t)})^2\frac{1}{\sqrt[4]{\tilde{N}_0}t/a'+1}\left(\sqrt[4]{\tilde{N}_0}t/a'+1\right)a'\tilde{N}_0^{\frac{3}{4}}\overset{(d)}{=}\frac{\delta\zeta_0}{p^{(t)}}\left(\frac{\zeta_0 a'/\sqrt[4]{\tilde{N}_0}}{t+a'/\sqrt[4]{\tilde{N}_0}}\right)^2a'\tilde{N}_0^{\frac{3}{4}}\\
=&\frac{1}{p^{(t)}}\left(\frac{\mu}{3.25}\right)^2\left(\frac{3.25/\mu}{t+a'/\sqrt[4]{\tilde{N}_0}}\right)^2\delta(\zeta_0 a')^3\sqrt[4]{\tilde{N}_0}\\
\overset{(e)}{\leq}&\frac{1}{p^{(t)}}\left(\frac{\mu}{3.25}\right)^2(\eta^{(t)})^2\delta(\zeta_0 a')^3\sqrt[4]{\tilde{N}_0},
\end{aligned}\tag{73}$$

where (a) is due to $\tilde{N}_{0,T} = \max_{t\in\{0,\ldots,T-1\}}\{\tilde{N}_0^{(t)}\}$; (b) is as a result of $p^{(t)} \leq \tilde{p}^{(t)}$ (cf. (35)); (c) is because of $\tilde{p}^{(t)} \leq \frac{\delta\zeta_0}{\sqrt[4]{\tilde{N}_0}t/a'+1}$ (cf. (36)); (d) is by definition of $\zeta^{(t)} = \frac{\zeta_0}{\sqrt[4]{\tilde{N}_0}t/a'+1}$; and (e) results from $\eta^{(t)} = \frac{3.25}{\mu}\frac{1}{t+a}$ with $a < a'/\sqrt[4]{\tilde{N}_0}$.

Finally, plugging the above results for part I and part II into (72), we complete the proof for Lemma 5.2.

Appendix H
Proof of Theorem 5.1

Since $\tilde{p}^{(t)}$ is decreasing over $t \geq 0$, by definition of $p^{(t)}(\delta,\omega)$ in (35), $p^{(t)}(\delta,\omega)$ proves to be non-increasing over $t \geq 0$. Hence, $p^{(t)}(\delta,\omega) \geq p^{(T)}(\delta,\omega)$, and we can further upper bound the RHS of (38) by replacing $p^{(t)}(\delta,\omega)$ with $p^{(T)}(\delta,\omega)$. Based on Lemma 5.2, we have the exact upper bound for $e^{(t)}$ given by [20, Lemma 22]

$$e^{(t)} \leq \frac{10}{(p^{(T)}(\delta,\omega))^2}(\eta^{(t)})^2\left(24KG^2 + A(\delta,\omega)\sqrt[4]{\tilde{N}_0}\right). \tag{74}$$

Then, by definition of $e^{(t)}$ (cf. (34)), we have

$$\sum_{i\in\mathcal{V}}\mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_i^{(t)}\|^2 \leq \frac{10}{(p^{(T)}(\delta,\omega))^2}(\eta^{(t)})^2\left(24KG^2 + A(\delta,\omega)\sqrt[4]{\tilde{N}_0}\right). \tag{75}$$

Next, to prove Theorem 5.1, we need to revisit [20, Lemma 20] as follows.

$$\mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^*\|^2 \overset{(1),(33)}{=} \mathbb{E}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\frac{1}{K}\sum_{i\in\mathcal{V}}\hat{\nabla}f_i(\boldsymbol{\theta}_i^{(t)}) + \zeta^{(t)}\frac{1}{K}\sum_{i\in\mathcal{V}}\sum_{\tau=0}^{t}\frac{m}{d}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{n}}_i^{(\tau)} - \boldsymbol{\theta}^*\right\|^2$$

$$= \underbrace{\mathbb{E}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\frac{1}{K}\sum_{i\in\mathcal{V}}\hat{\nabla}f_i(\boldsymbol{\theta}_i^{(t)}) - \boldsymbol{\theta}^*\right\|^2}_{\text{part I}} + \underbrace{\mathbb{E}\left\|\zeta^{(t)}\frac{1}{K}\sum_{i\in\mathcal{V}}\sum_{\tau=0}^{t}\frac{m}{d}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{n}}_i^{(\tau)}\right\|^2}_{\text{part II}}.$$

As for Part I, [20, Lemma 20] can be directly applied, and for part II, we need to relate it to learning rate $\eta^{(t)}$ as

$$\mathbb{E}\left\|\zeta^{(t)}\frac{1}{K}\sum_{i\in\mathcal{V}}\sum_{\tau=0}^{t}\frac{m}{d}(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{n}}_i^{(\tau)}\right\|^2 = \frac{1}{K^2}\frac{m^2}{d^2}(\zeta^{(t)})^2\sum_{\tau=0}^{t}\sum_{i\in\mathcal{V}}\mathbb{E}\|(\boldsymbol{A}^{(\tau)})^T\tilde{\boldsymbol{n}}_i^{(\tau)}\|^2$$

$$\overset{(a)}{=}\frac{1}{K^2}\frac{m^2}{d}(\zeta^{(t)})^2\sum_{\tau=0}^{t}\tilde{N}_0^{(\tau)} \overset{(37)}{\leq} \frac{1}{K^2}\frac{m^2}{d}(\zeta^{(t)})^2 t\tilde{N}_{0,T} \leq$$

$$\frac{1}{K^2}\frac{m^2}{d}\zeta^{(t)}\frac{\zeta_0(\delta,\omega)}{\sqrt[4]{\tilde{N}_0}t/a'+1}\left(\sqrt[4]{\tilde{N}_{0,T}}t/a'+1\right)a'\tilde{N}_{0,T}^{\frac{3}{4}} = \frac{1}{K^2}\frac{m^2}{d}\frac{\mu}{3.25}\frac{3.25/\mu}{t+a'/\sqrt[4]{\tilde{N}_0}}(\zeta_0(\delta,\omega)a')^2\sqrt{\tilde{N}_{0,T}}$$

$$\overset{(b)}{\leq}\frac{1}{K^2}\omega^2 d\frac{\mu}{3.25}\eta^{(t)}(\zeta_0(\delta,\omega)a')^2\sqrt{\tilde{N}_{0,T}}, \tag{76}$$

where (a) follows from $\boldsymbol{A}^{(\tau)}(\boldsymbol{A}^{(\tau)})^T = \frac{d}{m}\boldsymbol{I}$ and $\mathbb{E}\|\tilde{\boldsymbol{n}}_i^{(\tau)}\|^2 = m\tilde{N}_{0i}^{(\tau)}$; and (b) is due to $a < a'/\sqrt[4]{\tilde{N}_0}$ and $\omega = m/d$, as well as definition of $\eta^{(t)} = \frac{3.25}{\mu}\frac{1}{t+a}$. With part II replaced by the RHS of (76), we obtain the following lemma.

Lemma H.1 (Variant of [20, Lemma 20]): Denoting the optimal solution to (P0) by $\boldsymbol{\theta}^*$ and the corresponding objective value $F(\boldsymbol{\theta}^*)$ by F^*, the average of iterates $\bar{\boldsymbol{\theta}}^{(t+1)}$ satisfies

$$\mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^*\|^2 \leq (1-\eta^{(t)}\mu)\mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|^2 + \eta^{(t)}\frac{1}{K^2}\omega^2 d\frac{\mu}{3.25}(\zeta_0(\delta,\omega)a')^2\sqrt{\tilde{N}_{0,T}} + (\eta^{(t)})^2\frac{\bar{\sigma}^2}{K} + 2\eta^{(t)}(2\eta^{(t)}L-1)(\mathbb{E}[F(\bar{\boldsymbol{\theta}}^{(t)})] - F^*) + \eta^{(t)}\frac{2\eta^{(t)}L^2+L}{K}\sum_{i\in\mathcal{V}}\mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_i^{(t)}\|^2. \quad (77)$$

Since $\eta^{(t)} \leq \eta^{(0)} = \frac{3.25}{\mu a}$, and $a \geq \frac{13L}{\mu}$, it follows that $\eta^{(t)} \leq \frac{1}{4L}$ thus leading to $2L\eta^{(t)} - 1 \leq -0.5$ and $2\eta^{(t)}L^2 + L \leq 1.5L$. With this fact and the upper bound on the consensus error given by (75), (77) implies that

$$v_e^{(t+1)} \leq (1-\eta^{(t)}\mu)v_e^{(t)} - \eta^{(t)}f_e^{(t)} + \eta^{(t)}\frac{1}{K^2}\omega^2 d\frac{\mu}{3.25}(\zeta_0(\delta,\omega)a')^2\sqrt{\tilde{N}_{0,T}} + (\eta^{(t)})^2\frac{\bar{\sigma}^2}{K} + (\eta^{(t)})^3\frac{15L}{(p^{(T)}(\delta,\omega))^2}\left(24G^2 + \frac{A(\delta,\omega)}{K}\sqrt[4]{\tilde{N}_{0,T}}\right), \quad (78)$$

where $v_e^{(t)} = \mathbb{E}\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|^2$ and $f_e^{(t)} = \mathbb{E}[F(\bar{\boldsymbol{\theta}}^{(t)})] - F^*$ measure, on average over RLC, the distance to the optimal solution and the optimality gap to to the objective value for problem (P0), respectively. Note that the standard result in [40, Lemma 3.3] is not applicable to (78) to characterize the optimality gap due to the absence of linear terms w.r.t $\eta^{(t)}$. To capture the performance of the optimality-gap sequence $\{f_e^{(t)}\}$, we need the following lemma.

Lemma H.2 (Variant of [40, Lemma 3.3]): For non-negative sequences $\{v_e^{(t)}\}$ and $\{f_e^{(t)}\}$, $\eta^{(t)} = \frac{3.25}{\mu}\frac{1}{t+a}$ with $\mu > 0$ and $a > 1$, and constants $A,\ B,\ C \geq 0$, where $t = 0, \ldots, T-1$, such that

$$v_e^{(t+1)} \leq (1-\eta^{(t)}\mu)v_e^{(t)} + \eta^{(t)}A + (\eta^{(t)})^2 B + (\eta^{(t)})^3 C - \eta^{(t)}f_e^{(t)}, \quad (79)$$

we have

$$\frac{1}{S_T}\sum_{t=0}^{T-1} w^{(t)} f_e^{(t)} \leq \frac{\mu}{3.25}\frac{a^3 - 3.25a^2}{S_T}v_e^{(0)} + A + \frac{1.625(2a+T)T}{\mu S_T}B + \frac{3.25^2 T}{\mu^2 S_T}C, \quad (80)$$

where $w^{(t)} = (a+t)^2$ and $S_T = \sum_{t=0}^{T-1} w^{(t)}$.

Proof: Following similar procedures as that for proving [40, Lemma 3.3], first, multiplying $\frac{w^{(t)}}{\eta^{(t)}}$ with both sides of (79), we have

$$v_e^{(t+1)}\frac{w^{(t)}}{\eta^{(t)}} \leq (1-\eta^{(t)}\mu)\frac{w^{(t)}}{\eta^{(t)}}v_e^{(t)} + w^{(t)}A + w^{(t)}\eta^{(t)}B + w^{(t)}(\eta^{(t)})^2C - w^{(t)}f_e^{(t)}. \tag{81}$$

To obtain the same relation $(1-\eta^{(t)}\mu)\frac{w^{(t)}}{\eta^{(t)}} \leq \frac{w^{(t-1)}}{\eta^{(t-1)}}$ as in the original proof, it is equivalent to have $(c-3)(a+t)^2 + 3(a+t) - 1 \geq 0$ for all t. It is thus sufficient to have $(c-3)(a+t)^2 + 3(a+t) - 1|_{t=0} = (c-3)a^2 + 3a - 1 \geq 0$ for a choice of $c > 3$, which is satisfied by, e.g., $c = 3.25$ and any parameter $a \geq \frac{13L}{\mu}$. (By definition, $L \geq \mu$ and thus $a \geq 13$.) Next, by letting $t = T-1$ and recursively applying (81), it follows that

$$\sum_{t=0}^{T-1} w^{(t)}f_e^{(t)} \leq (1-\eta^{(0)}\mu)\frac{w^{(0)}}{\eta^{(0)}}v_e^{(0)} + \sum_{t=0}^{T-1} w^{(t)}A + \sum_{t=0}^{T-1} w^{(t)}\eta^{(t)}B + \sum_{t=0}^{T-1} w^{(t)}(\eta^{(t)})^2C. \tag{82}$$

As a result, plugging $\eta^{(0)} = \frac{3.25}{\mu a}$ and dividing both sides of (82) by S_T, we obtain (80). ■

According to Lemma H.2, for constants $A = \frac{1}{K^2}\omega^2 d\frac{\mu}{3.25}(\zeta_0(\delta,\omega)a')^2\sqrt{\tilde{N}_{0,T}}$, $B = \frac{\bar{\sigma}^2}{K}$ and $C = \frac{15L}{(p^{(T)}(\delta,\omega))^2}(24G^2 + \sqrt[4]{\tilde{N}_{0,T}}A(\delta,\omega)/K)$, as well as $f_e^{(t)} = \mathbb{E}[F(\bar{\boldsymbol{\theta}}^{(t)})] - F^*$, (78) implies that

$$\frac{1}{S_T}\sum_{t=0}^{T-1} w^{(t)}\mathbb{E}[F(\bar{\boldsymbol{\theta}}^{(t)})] - F^* \leq \frac{\mu}{3.25}\frac{a^3 - 3.25a^2}{S_T}v_e^{(0)} + \frac{158.45\left(24G^2 + \frac{A(\delta,\omega)}{K}\sqrt[4]{\tilde{N}_{0,T}}\right)LT}{\mu^2(p^{(T)}(\delta,\omega))^2 S_T} + \frac{1.625T(T+2a)}{\mu S_T}\frac{\bar{\sigma}^2}{K} + \frac{1}{K^2}C(\delta,\omega)\sqrt{\tilde{N}_{0,T}}. \tag{83}$$

Combined (83) with $F(\tilde{\boldsymbol{\theta}}_T) = F(\frac{1}{S_T}\sum_{t=0}^{T-1} w^{(t)}\bar{\boldsymbol{\theta}}^{(t)}) \leq \frac{1}{S_T}\sum_{t=0}^{T-1} w^{(t)}F(\bar{\boldsymbol{\theta}}^{(t)})$, which is due to Jensen's inequality, Theorem 5.1 is thus proved.

REFERENCES

[1] H. Xing, O. Simeone, and S. Bi, "Decentralized federated learning via SGD over wireless D2D networks," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, May 2020.

[2] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling up machine learning: Parallel and distributed approaches*. Cambridge Univ. Press, 2011.

[3] T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu, "Distributed learning in the nonconvex world: from batch to streaming data, and beyond," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 26–38, May 2020.

[4] P. Kairouz, H. B. McMahan *et al.*, "Advances and open problems in federated learning," *to appear in Foundations and Trends® in Machine Learning*, vol. 14, no. 1, Mar. 2021.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[6] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Jun. 2019.

[7] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

[8] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, Dec. 2017.

[9] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018.

[10] A. Abdi, Y. M. Saidutta, and F. Fekri, "Analog compression and communication for federated learning over wireless MAC," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, May 2020.

[11] J. Wu, W. Huang, J. Huang, and T. Zhang, "Error compensated quantized SGD and its applications to large-scale distributed optimization," in *International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018.

[12] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[13] A. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "Braintorrent: A peer-to-peer environment for decentralized federated learning," *arXiv preprint arXiv:1905.06731*, 2019.

[14] B. Research, "Bringing HPC techniques to deep learning," 2017. [Online]. Available: https://andrew.gibiansky.com/blog/machine-learning/baidu-allreduce/

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

[16] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proc. IEEE*, vol. 106, no. 5, pp. 953–976, Apr. 2018.

[17] R. Xin, S. Kar, and U. A. Khan, "An introduction to decentralized stochastic optimization with gradient tracking," *arXiv preprint arXiv:1907.09648v2*, 2019.

[18] T. Sun, D. Li, and B. Wang, "Decentralized federated averaging," *arXiv preprint arXiv:2104.11375*, 2021.

[19] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2: Decentralized training over decentralized data," in *International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018.

[20] A. Koloskova, S. U. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, Jun. 2019.

[21] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, "Decentralized deep learning with arbitrary communication compression," in *International Conference on Learning Representations (ICLR)*, Apr. 2020.

[22] N. Singh, D. Data, J. George, and S. Diggavi, "SPARQ-SGD: Event-triggered and compressed communication in decentralized stochastic optimization," in *IEEE Conference on Decision and Control (CDC)*, Dec. 2020.

[23] T. Vogels, S. Karimireddy, and M. Jaggi, "Practical low-rank communication compression in decentralized deep learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2020.

[24] 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.

[25] C. Xiaowen, Z. Guangxu, X. Jie, W. Zhiqin, and C. Shuguang, "Optimized power control design for over-the-air federated edge learning," *arXiv preprint arXiv:2106.09316*, 2021.

[26] 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.

[27] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *to appear in IEEE Tran. Commun.*, 2021.

[28] J.-H. Ahn, O. Simeone, and J. Kang, "Wireless federated distillation for distributed edge learning with heterogeneous data," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Istanbul, Turkey, Sep. 2019.

[29] H. Guo, A. Liu, and V. K. N. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 197–210, Jan. 2021.

[30] J. Wang, A. K. Sahu, Z. Yang, G. Joshi, and S. Kar, "MATCHA: Speeding up decentralized SGD via matching decomposition sampling," in *Indian Control Conference*, Hyderabad, India, Dec. 2019.

[31] S. Savazzi, S. Kianoush, V. Rampa, and M. Bennis, "A joint decentralized federated learning and communications framework for industrial networks," in *IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Sep. 2020.

[32] E. Ozfatura, S. Rini, and D. Gunduz, "Decentralized SGD with over-the-air computation," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2020.

[33] Y. Shi, Y. Zhou, and Y. Shi, "Over-the-air decentralized federated learning," in *IEEE International Symposium on Information Theory (ISIT)*, Jul. 2021.

[34] R. Saha, S. Rini, M. Rao, , and A. Goldsmith, "Decentralized optimization over noisy, rate-constrained networks: Achieving consensus by communicating differences," *arXiv preprint arXiv:2010.11292*, 2021.

[35] T. Husfeldt, "Graph colouring algorithms," 2015. [Online]. Available: https://arxiv.org/abs/1505.05825

[36] D. Chakrabarty and P. Siddiqui, "On a decentralized $(\delta + 1)$-graph coloring algorithm," in *SIAM Symposium on Simplicity in Algorithms*, Salt Lake City, Utah, U.S., Jan. 2020.

[37] O. Abari, H. Rahul, and D. Katabi, "Over-the-air function computation in sensor networks," *arXiv preprint arXiv:1612.02307*, 2016.

[38] H. Xing, O. Simeone, and S. Bi, "Federated learning over wireless device-to-device networks: Algorithms and convergence analysis," *arXiv preprint arXiv:2101.12704*, 2021.

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

[40] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Advances in Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, Dec. 2018.

[41] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[42] N. J. Higham, *Accuracy and stability of numerical algorithms*. Philadelphia, PA, USA: SIAM, 2002.