Massive Digital Over-the-Air Computation for Communication-Efficient Federated Edge Learning

Li Qiao, *Graduate Student Member, IEEE*, Zhen Gao, *Member, IEEE*,
Mahdi Boloursaz Mashhadi, *Senior Member, IEEE*, and Deniz Gündüz, *Fellow, IEEE*

***Abstract*—Over-the-air computation (AirComp) is a promising technology converging communication and computation over wireless networks, which can be particularly effective in model training, inference, and more emerging edge intelligence applications. AirComp relies on uncoded transmission of individual signals, which are added naturally over the multiple access channel thanks to the superposition property of the wireless medium. Despite significantly improved communication efficiency, how to accommodate AirComp in the existing and future digital communication networks that are based on discrete modulation schemes remains a challenge. This paper proposes a massive digital AirComp (MD-AirComp) scheme that leverages an unsourced massive access protocol to enhance compatibility with both current and next-generation wireless networks. MD-AirComp utilizes vector quantization to reduce the uplink communication overhead, and employs shared quantization and modulation codebooks. At the receiver, we propose a near-optimal approximate message passing-based algorithm to compute the model aggregation results from the superposed sequences, which relies on estimating the number of devices transmitting each code sequence, rather than trying to decode the messages of individual transmitters. We apply MD-AirComp to federated edge learning (FEEL), and show that it significantly accelerates FEEL convergence compared to state-of-the-art while using the same amount of communication resources.**

***Index Terms*—Artificial intelligence of things (AIoT), digital over-the-air computation, unsourced massive access, federated edge learning, distributed optimization.**

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U2233216 and Grant 62071044, in part by the Shandong Province Natural Science Foundation under Grant ZR2022YQ62, and in part by the Beijing Nova Program. The work of L. Qiao was supported by the China Scholarship Council, and was carried out while he was a visiting student at Imperial College London. This work received funding from the UKRI for the projects AI-R (ERC Consolidator Grant, EP/X030806/1) and SONATA (EPSRC-EP/W035960/1), and the SNS JU project 6G-GOALS under the EU's Horizon program (Grant Agreement No. 101139232.) Part of the results were presented at the 2023 IEEE International Symposium on Information Theory (ISIT) [1]. We release our code at *https://github.com/liqiao19/MD-AirComp* and *https://gaozhen16.github.io/* to facilitate further research and reproducibility. For the purpose of open access, the authors have applied a Creative Commons Attribution (CCBY) license to any Author Accepted Manuscript version arising from this submission. (Corresponding author: Zhen Gao)

L. Qiao is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China, also with 5GIC & 6GIC, the Institute for Communication Systems (ICS), University of Surrey, GU2 7XH Guildford, U.K. (e-mails: qiaoli@bit.edu.cn).

Z. Gao is with Beijing Institute of Technology (Zhuhai), Zhuhai 519088, China, also with the MIIT Key Laboratory of Complex-Field Intelligent Sensing, Beijing Institute of Technology, Beijing 100081, China, also with the Yangtze Delta Region Academy of Beijing Institute of Technology, Beijing Institute of Technology, Jiaxing 314000, China, also with the Advanced Technology Research Institute, Beijing Institute of Technology, Jinan 250307, China. (e-mails: gaozhen16@bit.edu.cn).

M. Boloursaz Mashhadi is with 5GIC & 6GIC, Institute for Communication Systems (ICS), University of Surrey, GU2 7XH Guildford, United Kingdom (email: m.boloursazmashhadi@surrey.ac.uk).

D. Gündüz is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (email: d.gunduz@imperial.ac.uk).

I. Introduction

With the advancement of the artificial intelligence of things (AIoT), a growing volume of data is being generated by a massive number of Internet-of-things (IoT) devices, which fuels the ongoing revolution in machine learning (ML) and artificial intelligence (AI) [2]–[4]. Traditionally, data collected by IoT devices is offloaded to the cloud or data centers to train massively large models [5]. However, due to concerns over data privacy and increasing computational capabilities of edge devices, AIoT in wireless networks is transitioning away from centralized learning towards distributed learning frameworks [5]–[9]. One of these emerging frameworks is federated edge learning (FEEL), where a central server coordinates multiple AIoT devices in its coverage area to train an ML model using their respective local datasets and computing resources [10]. As a result, FEEL requires solving the parallel and distributed optimization problem while considering the limited shared wireless resources and interference between devices. Model aggregation process in FEEL involves repeated uplink (UL) transmission of high-dimensional local gradients or model updates from tens to hundreds of devices, which causes a heavy burden on the multiple access networks [11].

A. Related Work

To address this problem, the concept of over-the-air computation (AirComp) has been introduced as a means of achieving communication-efficient model aggregation [10]–[12]. In the context of model aggregation, the objective is to compute the average of local model updates, rather than recovering each update individually. In analog AirComp, the fundamental principle involves embracing inter-user interference over the multiple-access channel (MAC) rather than mitigating it. More specifically, every element of the local model update at a device goes through pre-equalization using the inverse of the UL channel gain from the device to the BS, which acts as the parameter server. Subsequently, the pre-equalized elements are modulated onto the transmit waveform's amplitude. Through simultaneous transmission from multiple devices, the receiver can directly obtain the sum of the local model updates from multiple devices by leveraging the superposed waveform over the MAC. To further reduce the communication overhead, by exploiting sparsification and error accumulation, the dimension of the local model updates are reduced before analog transmission in [12]. After analog AirComp, the receiver reconstructs the average of local model updates by using the approximate message passing (AMP) algorithm. A higher FEEL accuracy is achieved with limited communication resources in [13], by exploiting time-correlated sparsification on the local model updates. Multi-input multi-output (MIMO)-based AirComp is

studied in [14]. Random projection technique proposed in [12] is extended to MIMO channel in [15]. Note that all the participating devices are synchronized in time in [10]–[15]. If the synchronization is imperfect, [16] and [17] considered misaligned analog AirComp to solve the problem. Dynamic scheduling of devices is proposed in [18] to optimize the energy efficiency of AirComp in FEEL. We refer the readers to [19] and [20] for a comprehensive overview of FEEL and AirComp. We should emphasize that AirComp is a general technology that turns wireless medium into a joint communication and computation platform. Its benefits can also be applied to other parallel and distributed optimization applications [21]. For instance, AirComp has recently been used for efficient inference at the edge [22], [23].

Despite the significant potential benefits of AirComp, most of the existing wireless networks adopt digital communication protocols (e.g., 3rd Generation Partnership Project standards) as well as hardware [24]. In other words, both existing and next-generation wireless networks may not have the flexibility to support any arbitrary modulation scheme, which is a fundamental requirement for the analog AirComp methods described earlier. As for digital communication, quantization, i.e., the process of mapping continuous values to a finite set of discrete values, is necessary. Quantization of model parameters in federated learning (FL) has been studied extensively, considering rate-limited error-free communication links from devices to the parameter server [25]–[30]. Specifically, the authors of [25] analyzed the convergence of FL with stochastic uniform quantization. To optimize the convergence speed of FL with quantization, adaptive quantization levels were investigated in [26] and [27]. In addition, joint optimization of bandwidth and quantization levels was considered in [28]. To further reduce the UL communication overhead, vector quantization (VQ) have shown a significant improvement on the communication efficiency in [29] and [30]. Different from the orthogonal UL transmission considered in [25]–[30], the authors of [31] proposed one-bit digital aggregation (OBDA) scheme for FEEL, which combines the ideas of AirComp and sign stochastic gradient descent (signSGD). Specifically, one-bit gradient quantization and digital modulation are adopted at devices, and majority-vote based gradient-decoding via AirComp is employed at the central server. Furthermore, non-coherent detector was proposed in [32] and [33], enabling scalar quantization-based digital AirComp without the need of channel state information (CSI). Despite its potential, this approach needs to allocate multiple frequencies to represent the quantization levels, thus raising concerns regarding spectral efficiency. In addition, a series of coding schemes recently proposed in [34]–[36] enables function computation from superposed digital constellations. However, these emerging digital AirComp schemes, i.e., [31]–[36], only enables scalar quantization, and it is not clear how/if these approaches can be extended to VQ to further reduce the communication overhead. Hence, this necessities a novel communication-efficient digital AirComp scheme that is suitable for any scalar or vector quantization scheme and can be deployed in existing digital communication systems.

To fill in this gap, we propose a digital scheme that jointly enables random access and computation among massive AIoT devices. We first revisit grant-free random access (GFRA) protocols, which have been recently studied to enable massive IoT access with low signaling overhead and latency [37]. Since devices can directly transmit their signals without the permission of the base station (BS), GFRA protocols reduce the signaling overhead. There are mainly two types of GFRA schemes, i.e., sourced massive access (SMA) [38]–[45] and unsourced massive access (UMA) [46]–[49]. Specifically, the SMA schemes proposed in [39]–[41] have demonstrated their superiority in both access latency and detection performance with the help of the elegant AMP algorithm framework. Furthermore, authors in [42] and [43] extend the SMA framework to non-terrestrial networks, which efficiently improve the transmission efficiency and reduce the detection error at the same time. Both SMA and UMA schemes adopt non-orthogonal random access codebook for reducing the communication overhead. In SMA, each device has its unique non-orthogonal preamble indicating its identity. The BS performs active device detection and channel estimation, followed by data decoding [38]–[45]. In UMA, instead, all the devices adopt the same non-orthogonal codebook. Transmitted bits are modulated by the index of the UMA codewords and decoded at the BS. The goal of the BS is to decode the list of transmitted messages, but not the identities of the active devices [46]–[49]. We remark that the identities of active devices are not necessary for model aggregation in FEEL either. Inspired by UMA, our target is to redesign the modulation and decoding modules of conventional UMA for efficient AirComp across a large number of devices.

B. Contributions

To improve the communication efficiency of FEEL, we propose a massive digital AirComp (MD-AirComp) scheme that can be deployed in existing and next-generation digital communication systems. Specifically, consider a scenario with multiple devices, each with its own signal vector. The goal of the receiver is to recover the average of these vectors. In MD-AirComp, individual signal vectors are quantized using the common VQ codebook. Then, the active devices select their transmit sequences from the common modulation codebook based on their quantization indices. Note that there is a one-to-one mapping between the quantization and modulation codebooks. After pre-equalization, the non-orthogonal sequences from active devices are transmitted and overlap at the BS. Then, an AMP-based digital aggregation (AMP-DA) algorithm is proposed at the BS to estimate the number of devices employing each modulation codeword. We remark that in most FEEL and edge inference applications, active devices are likely to transmit the same message, especially when their datasets/models are statistically similar [15], [22]. Our main contributions are summarized as follows:

- **Communication-efficient MD-AirComp method:** By exploiting the superposition property of MAC and VQ, an MD-AirComp scheme is proposed for communication-efficient FEEL across digital communication systems. Since the receiver is interested in the sum of the signal vectors, the MD-AirComp scheme is user-collision

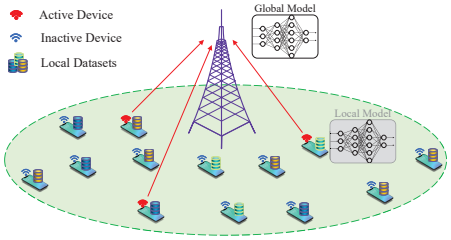


Fig. 1. Illustration of the FEEL scenario in cellular networks.

friendly. In addition, it can be deployed with any scalar or vector quantization scheme to exploit the distribution of the signal vectors for better training accuracy. Since our proposed MD-AirComp scheme is a general lossy computing scheme over a MAC, it can also be applied to other edge computing tasks, such as collaborative inference [22].

- **Proposed AMP-DA algorithm for signal aggregation:** Due to the sporadic activity of AIoT devices and the potentially high-resolution VQ in FEEL, the signal to be detected is inherently sparse. Moreover, due to similarities in local dataset, some devices are likely to transmit the same message, and user-collisions will make the signal even more sparse. In addition, non-zero elements in the signal are integers representing the numbers of devices selecting a specified codeword. By modeling these properties in the prior distribution, the proposed AMP-DA algorithm realizes efficient model aggregation. Furthermore, the AMP-DA algorithm uses majority voting to estimate the number of active devices and exploits the multiple observations from multi-antenna BS, to further enhance the detection accuracy.
- **Convergence analysis of MD-AirComp-based FEEL:** We analyze the convergence rate of FEEL when the proposed MD-AirComp scheme is employed. Specifically, we analyze the influence of errors, originating from VQ and wireless communication, on the convergence rate. The wireless communication error relates to the detection accuracy of the AMP-DA algorithm. Our findings indicate that the MD-AirComp-based FEEL achieves a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$, where T is the number of global training rounds.

Notation: Boldface lower and upper-case symbols denote column vectors and matrices, respectively. For a matrix $\mathbf{A}$, $\mathbf{A}^T$, $\mathbf{A}^*$, $\mathbf{A}^H$, $\|\mathbf{A}\|_F$, $[\mathbf{A}]_{m,n}$ denote the transpose, conjugate, Hermitian transpose, Frobenius norm, and the m-th row and n-th column element of $\mathbf{A}$, respectively. $[\mathbf{A}]_{:,n}$ denotes the n-th column of $\mathbf{A}$. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_p$ and $[\mathbf{x}]_m$ denote the l_p norm and m-th element of $\mathbf{x}$, respectively. $|\Gamma|$ denotes the cardinality of the ordered set Γ. $\lfloor\cdot\rfloor$ ($\lceil\cdot\rceil$) rounds each element to the nearest integer smaller (larger) than or equal to that element. The marginal distribution $p([\mathbf{x}]_m)$ is denoted as $p([\mathbf{x}]_m) = \int_{\setminus[\mathbf{x}]_m} p(\mathbf{x})$. $\mathcal{CN}(x;\mu,\nu)$ or $x \sim \mathcal{CN}(\mu,\nu)$ denotes the complex Gaussian distribution of random variable x with mean μ and variance ν. $[K]$ denotes the set $\{1,2,...,K\}$. $\mathbf{0}_{m\times 1}$ denotes the all-zero vector with m rows.

II. System Model

As shown in Fig. 1, we consider K single-antenna edge devices served by an M-antenna BS in a cellular system. Each device k, $k \in [K]$, has its own local dataset $\mathcal{B}_k$, which consists of labeled data samples. For convenience, we assume uniform sizes for local datasets, that is, $|\mathcal{B}_k| = B$, $\forall k$. In FEEL, K devices are coordinated by the the BS to solve the following distributed optimization problem in parallel

$$\min_{\mathbf{w}\in\mathbb{R}^W} f(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K} F_k(\mathbf{w}), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^W$ is the parameter vector of a common neural network model and $F_k(\mathbf{w})$ is the local loss function.

We follow the seminal federated averaging (FedAvg) [50] algorithm[1]. Specifically, in the $(t-1)$-th global training round of FEEL, $\forall t \in [T]$, the BS broadcasts the current global model $\mathbf{w}^{t-1}$ to the devices. Based on $\mathbf{w}^{t-1}$ and its local dataset $\mathcal{B}_k$, the k-th device performs T_l local iterations of stochastic gradient descent (SGD) to obtain $\mathbf{w}_k^t$, which is expressed as

$$\mathbf{w}_{k,t_l}^t = \mathbf{w}_{k,t_l-1}^t - \frac{\eta_l}{|\mathcal{B}_k|}\sum_{\boldsymbol{\varepsilon}\in\mathcal{B}_k} \nabla F_k(\mathbf{w}_{k,t_l-1}^t, \boldsymbol{\varepsilon}), \tag{2}$$

where $t_l \in [T_l]$, $\mathbf{w}_{k,0}^t = \mathbf{w}^{t-1}$, $\mathbf{w}_k^t = \mathbf{w}_{k,T_l}^t$, ∇ and η_l denote the gradient operator and the local learning rate, respectively, $F_k(\mathbf{w}_{k,t_l-1}^t, \boldsymbol{\varepsilon})$ denotes the local loss of model $\mathbf{w}_{k,t_l-1}^t$ on training sample $\boldsymbol{\varepsilon}$ with respect to its true label.

According to (2), the local model update can be obtained as $\Delta_k^t = \mathbf{w}_k^t - \mathbf{w}^{t-1}$, $\forall k,t$. Ideally, the local model updates will be transmitted to the BS for global model update. In practice, to reduce the communication cost of Δ_k^t, compression[2] and error accumulation are commonly adopted in FEEL [12]. In particular, we denote $\bar{\mathbf{s}}_k^t = \Delta_k^t + \mathbf{e}_k^t$, $\forall k,t$, where the local model update Δ_k^t is compensated with the accumulated error vector $\mathbf{e}_k^t$. In addition, $\mathbf{e}_k^t$ is defined as $\mathbf{e}_k^t = \bar{\mathbf{s}}_k^{t-1} - \mathcal{C}(\bar{\mathbf{s}}_k^{t-1})$, where $\mathcal{C}(\cdot)$ denotes compression on its argument and the output is the element-wise compressed values. It can be seen that the accumulated error vector keeps track of the values that are not updated to the BS due to compression, helping to accelerate the training [54]. Hence, the compressed local model update $\mathbf{s}_k^t = \mathcal{C}(\bar{\mathbf{s}}_k^t)$ is transmitted to the BS rather than Δ_k^t, $\forall k,t$, to reduce the communication cost while maintaining good training accuracy.

Furthermore, it is commonly adopted that only K_a ($K_a \ll K$) devices are simultaneously active in AIoT networks [12], [20], [28]. We assume the set of active devices $\mathcal{S}_a$, where $|\mathcal{S}_a| = K_a$, send their compressed local model updates $\mathbf{s}_k^t$, $k \in \mathcal{S}_a$, to the BS in each round. If $\mathbf{s}_k^t$ can be perfectly obtained at the BS, the global model update $\mathbf{s}^t \in \mathbb{R}^W$ can be calculated as

$$\mathbf{s}^t = \frac{1}{K_a}\sum_{k\in\mathcal{S}_a} \mathbf{s}_k^t. \tag{3}$$

Then, the global model can be updated as

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta\mathbf{s}^t, \tag{4}$$

[1]Although we consider FedAvg in this paper, an extension of the results to other federated optimization algorithms [51]–[53] is straightforward.

[2]Note that the typically used quantization and sparsification are two kinds of compression methods.

where η is the global learning rate. Finally, the updated parameter vector $\mathbf{w}^{t+1} \in \mathbb{R}^W$ is sent back to the devices, and steps (2), (3), and (4) are iterated until a convergence condition is met.

It is clear from (3) that only the sum of the local model updates, rather than the individual values, is needed at the BS. The BS does not need to identify the active devices either. However, due to the large dimension of the parameter vector and the large number of devices that can potentially participate in the training, there is a communication bottleneck in FEEL. Thereby, we propose a communication-efficient MD-AirComp scheme as explained in Section III.

Remark 1. *One reasons to have $K_a \ll K$ in FEEL could be due to heterogeneous computation speeds across edge devices [12], [20], [28], [53]. For a given time slot, only the devices that have finished their computation task will transmit. Alternatively devices can join or leave the learning process at different frames based on their energy or data availability as considered in typical IoT scenarios [2].*

III. The Proposed MD-AirComp Scheme

As shown in Fig. 2, the proposed MD-AirComp scheme consists of several modules, i.e., a quantization module, a codebook-based modulation module, and a pre-equalization module at the transmitter, a detection module and a digital aggregation module at the receiver. In addition, we use dotted boxes to describe the local training module and the model update module that are specific to FEEL. In the following, we will describe each of these modules in detail. For simplicity, we omit the index t unless otherwise specified. We reemphasize that MD-AirComp is a general approach for lossy computation over a MAC. In this paper, we motivate it through its application to FEEL.

A. Transmitter Design

At the transmitter, each active device first quantizes $\bar{\mathbf{s}}_k \in \mathbb{R}^W$, and then modulates the quantized value to a common modulation codebook. All active devices transmit their modulation codewords simultaneously. Next, we explain each component in detail.

1) Quantization Design: We will employ a single quantization codebook to be used by all the devices, denoted by $\mathbf{U} \in \mathbb{R}^{Q\times N}$, where $N = 2^J$ corresponds to J-bit quantization with N quantization codewords, $Q \geq 1$ ($Q \in \mathbb{N}$) denotes the length of each quantization codeword, i.e., $\mathbf{u}_n \in \mathbb{R}^Q$, $\forall n \in [N]$. Note that $Q = 1$ and $Q > 1$ indicate scalar quantization and VQ, respectively.

The vector $\bar{\mathbf{s}}_k \in \mathbb{R}^W$ at device k is divided into $D = W/Q$ blocks. For simplicity, here we assume that W can be divided by Q. Each block is then quantized independently using codebook $\mathbf{U}$. Let $\bar{\mathbf{s}}_k \in \mathbb{R}^W$ be quantized into $\mathbf{b}_k \in [N]^D$, where the d-th element of $\mathbf{b}_k$, $[\mathbf{b}_k]_d$, denotes the quantization index of the corresponding block. In particular, $[\mathbf{b}_k]_d$ is identified by finding the quantization codeword with the minimum Euclidean distance, which can be expressed as

$$[\mathbf{b}_k]_d = \arg\min_{i\in[N]} \left\| [\mathbf{U}]_{:,i} - [\bar{\mathbf{s}}_k]_{(d-1)Q+1:dQ} \right\|_2. \quad (5)$$

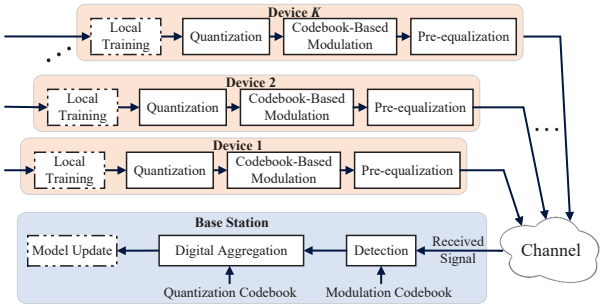


Fig. 2. The schematic diagram of the proposed MD-AirComp scheme, where the modules with dotted boxes are specific to FEEL.

To determine the quantization codebook, we can use any VQ method. From (5), it can be seen that the size of each local model update can be reduced by using VQ, thereby decreasing the communication overhead in the UL by a factor of Q.

Furthermore, we introduce a one-hot vector $\mathbf{x}_k^d \in \{0,1\}^N$, namely the selection vector, where $\|\mathbf{x}_k^d\|_0 = 1$ and $\left[\mathbf{x}_k^d\right]_{[\mathbf{b}_k]_d} = 1$. Hence, we can denote the quantized version of $\bar{\mathbf{s}}_k$ as

$$\mathbf{s}_k = \mathcal{C}(\bar{\mathbf{s}}_k) = (\mathbf{I}_D \otimes \mathbf{U}) \left[(\mathbf{x}_k^1)^T, ..., (\mathbf{x}_k^D)^T\right]^T, \quad (6)$$

where $\otimes$ and $\mathbf{I}_D$ denote the Kronecker product and the identity matrix of size D, respectively.

2) Codebook-Based Modulation: We consider the same random access codebook shared by all the devices, denoted by $\mathbf{P} \in \mathbb{C}^{L\times N}$, where each column of $\mathbf{P}$ is a codeword of length L, and there are N codewords in total. In the proposed scheme, there is a one-to-one mapping between $\mathbf{P}$ and $\mathbf{U}$. To be specific, if the device acquires the quantized value $[\mathbf{U}]_{:,n}$, then the modulation codeword (or sequence) $[\mathbf{P}]_{:,n}$ with the same index n is transmitted. The transmission process is similar to the preamble transmission of random access procedure in 5G communication systems [24]. Hence, the proposed codebook-based modulation is more suitable for the current wireless networks compared to analog modulation typically employed in AirComp.

3) UL Massive Access Model: We consider a time division duplex system, where the UL and downlink (DL) channels are commonly considered the same due to channel reciprocity [40], [48]. Before each UL communication round of FEEL, the BS broadcasts a pilot signal and each active device estimates its DL channel based on this pilot signal. Then, the active devices start their grant-free UL transmissions with pre-equalization in a synchronized manner [40]. For the d-th element of $\mathbf{b}_k$, $\forall k, d$, the received signal $\mathbf{Y}_d \in \mathbb{C}^{L\times M}$ at the BS can be expressed as

$$\mathbf{Y}_d = \sum_{k\in\mathcal{S}_a} \frac{1}{[\mathbf{h}_k]_1} \mathbf{P}\mathbf{x}_k^d (\mathbf{h}_k)^T + \mathbf{Z}_d = \mathbf{P}\mathbf{X}_d + \mathbf{Z}_d, \quad (7)$$

where the elements of the UL channel vector $\mathbf{h}_k \in \mathbb{C}^M$ are assumed to follow a complex Gaussian distribution with zero mean and unit variance, while channel inversion is considered for the first antenna of the BS[3]. $\mathbf{X}_d = \sum_{k\in\mathcal{S}_a} \frac{1}{[\mathbf{h}_k]_1} \mathbf{x}_k^d (\mathbf{h}_k)^T \in$

[3]Since each device has a single antenna, it is only possible to apply channel inversion for one of the antennas at the BS. For simplicity, we consider channel inversion for the first antenna of the BS. Also, we assume the channel reciprocity and perfect DL channel estimation at the devices [40]. Imperfect channel estimation will be discussed in Section VI-E.

$\mathbb{C}^{N\times M}$ is the equivalent transmit signal, and elements of the noise vector $\mathbf{Z}_d \in \mathbb{C}^{L\times M}$ obeys the independent and identically distributed (i.i.d.) complex Gaussian distribution with zero mean and variance σ_n^2. We define the signal-to-noise ratio (SNR) SNR $\triangleq 10\log_{10}\frac{\|\mathbf{P}\mathbf{X}_d\|_F^2}{LM\sigma_n^2}$. Hence, the first column of $\mathbf{X}_d$ obeys the following properties,

$$\left\|[\mathbf{X}_d]_{:,1}\right\|_1 = K_a,\ \ \left\|[\mathbf{X}_d]_{:,1}\right\|_0 \le K_a,\ \ [\mathbf{X}_d]_{n,1} \in \Omega, \forall n, \quad (8)$$

where $\Omega = \{0, [K_a]\}$. It can be seen that the n-th column of $\mathbf{P}$ is transmitted by $[\mathbf{X}_d]_{n,1}$ active devices, or equivalently the n-th quantization codeword appears at $[\mathbf{X}_d]_{n,1}$ devices, which indicates how to conduct the model aggregation. The non-zero elements of $[\mathbf{X}_d]_{:,m}$, $m \in \{2, 3, ..., M\}$, obey the ratio distribution of two uncorrelated complex Gaussian variables as discussed in [55]. Furthermore, due to the multiple observations at multiple antennas at the BS, $\mathbf{X}_d$ has row sparsity [39], which can be expressed as

$$\text{supp}\left\{[\mathbf{X}_d]_{:,1}\right\} = \text{supp}\left\{[\mathbf{X}_d]_{:,2}\right\} = \cdots = \text{supp}\left\{[\mathbf{X}_d]_{:,M}\right\}, \quad (9)$$

where supp $\{\cdot\}$ denotes the index of non-zero elements (rows) of a vector (matrix). Both (8) and (9) will be exploited to reconstruct $[\mathbf{X}_d]_{:,1}$, $\forall d$, in Section IV.

Remark 2. *As shown in (7), the noise power can be defined by the SNR and the average power of the received signal, which is commonly used in massive access system models [37]. If the wireless channel $[\mathbf{h}_k^{\text{UL}}]_1$ is in deep fading, the power of $[\mathbf{h}_k^{\text{UL}}]_1$ can be extremely small. In this case, pre-equalization might increase the received signal power at the rest of the antennas. As a result, SNR at the first antenna can be too small to conduct proper signal detection. In addition, if the channel power is too small, the transmit power needed for pre-equalization might be too large. Hence, we set a threshold ϵ_h on the channel condition, such that, if $|[\mathbf{h}_k]_1| < \epsilon_h$, the device does not participate in this training round.*

B. Receiver Design

According to (3), the receiver's target in MD-AirComp is to compute the aggregated global model update $\mathbf{s} \in \mathbb{R}^W$, which can be expressed as

$$\begin{aligned}\mathbf{s} &= \frac{1}{K_a}(\mathbf{I}_D \otimes \mathbf{U})\left[\left(\sum_{k\in\mathcal{S}_a}\mathbf{x}_k^1\right)^T, ..., \left(\sum_{k\in\mathcal{S}_a}\mathbf{x}_k^D\right)^T\right]^T \\ &= \frac{1}{K_a}(\mathbf{I}_D \otimes \mathbf{U})\left[\left([\mathbf{X}_1]_{:,1}\right)^T, ..., \left([\mathbf{X}_D]_{:,1}\right)^T\right]^T. \quad (10)\end{aligned}$$

It can be seen that the main differences between (10) and the decoding problem of UMA are as follows

- **The proposed MD-AirComp scheme is collision friendly:** In particular, the superposed equivalent signal vector $[\mathbf{X}_d]_{:,1}$ is used for the computing task (10) as a whole. Hence, it is unnecessary to decode each device's data from $[\mathbf{X}_d]_{:,1}$, $\forall d$. As a result, it is unnecessary to avoid collisions of multiple devices.

Algorithm 1 Proposed MD-AirComp Scheme

1: **Initialization:** Initialize model $\mathbf{w}^0 \in \mathbb{R}^W$, accumulated error vector $\mathbf{e}_k^0 = \mathbf{0}_{W\times 1}$, $\forall k \in [K]$, and global training round $t = 1$.
2: **while** $t < T$ **do**
3: BS broadcasts $\mathbf{w}^t$ to all the devices.
4: Local training via (2) to obtain Δ_k^t;
5: Error accumulation $\overline{\mathbf{s}}_k^t = \Delta_k^t + \mathbf{e}_k^t$;
6: Quantization via (5) and (6) to obtain $(\mathbf{x}_k^d)^t$ and $\mathbf{s}_k^t$;
7: Refine the accumulated error vector $\mathbf{e}_k^{t+1} = \Delta_k^t + \mathbf{e}_k^t - \mathbf{s}_k^t$;
8: Based on the quantization results in line 6, each device selects modulation codewords $\mathbf{P}(\mathbf{x}_k^d)^t$, $\forall d, k$;
9: BS broadcasts pilot signal for synchronization and channel pre-equalization. Active devices transmit, as shown in (7);
10: BS estimates overlapped selection vector $[\mathbf{X}_d]_{:,1}$, $\forall d$, and number of active devices K_a;
11: Model aggregation via (10);
12: Model update via (11).
13: $t = t + 1$.
14: **end while**

- **Detection algorithms for UMA do not apply to MD-AirComp:** In UMA, since data is embedded in the index of the transmitted codeword, detection algorithms focus on estimating the support of $\mathbf{X}_d$ rather than its values, $\forall d$. Furthermore, the number of active users K_a is trivial in UMA, but its detection is crucial for (10). In addition, due to pre-equalization in MD-AirComp, the received signal distributions on the first antenna and the remaining antennas are different.

1) CS-Based Detection: To achieve higher quantization accuracy, especially for VQ, the number of quantization bits J can be set very large, e.g., more than 8 bits. In this case, $N = 2^J$ can be far greater than K_a (i.e., $K_a \ll N$), which motivates us to employ non-orthogonal codebook $\mathbf{P}$ ($L < N$) and CS-based detection algorithms to reduce the communication overhead. Thanks to the sparsity of the equivalent signal $[\mathbf{X}_d]_{:,1}$, i.e., the number of non-zero elements $\|[\mathbf{X}_d]_{:,1}\|_0 \le K_a \ll N$, we can employ CS-based algorithms to estimate $\mathbf{x}_d$. Note that the codebook $\mathbf{P}$ is known at the BS. Our design target is to estimate the elements $[\mathbf{X}_d]_{:,1}$ as accurate as possible. The details of the proposed CS-based detection algorithm will be given in Section IV.

Furthermore, since the number of non-zero elements $\|[\mathbf{X}_d]_{:,1}\|_0$ will decrease if more users transmit the same codeword, the interference improves sparsity thereby becomes helpful for the CS-based detection. Hence, unlike the traditional UMA [47], user collision avoidance is unnecessary.

2) Digital Aggregation: After CS-based detection, we can obtain an estimate of the equivalent transmit signals, denoted by $[\widehat{\mathbf{X}}_d]_{:,1}$, $\forall d$. According to (8), estimated number of active devices $\widehat{K}_a$ can be obtained by calculating $\|[\widehat{\mathbf{X}}_d]_{:,1}\|_1$, $\forall d$, which will be illustrated in Section IV-C in detail. Then, by substituting $\widehat{K}_a$ and $[\widehat{\mathbf{X}}_d]_{:,1}$ into (10), we obtain the estimated global model update $\widehat{\mathbf{s}}$.

3) Model Update: According to (4) and (10), BS updates the weight of the ML model as

$$\widehat{\mathbf{w}}^{t+1} = \widehat{\mathbf{w}}^t + \eta \cdot \widehat{\mathbf{s}}^t, \quad (11)$$

and broadcasts it to all the devices and the next training round starts. Finally, the proposed MD-AirComp based FEEL procedure is summarized in **Algorithm** 1.

IV. Proposed CS-Based Digital Aggregation Algorithm

In this section, we will first introduce the CS-based problem formulation. Furthermore, we propose the AMP-based algorithm for efficient detection and K_a estimation. Finally, we summarize the proposed AMP-DA algorithm and analyse its computational complexity. For simplicity, we omit the subscript d in this section unless otherwise specified.

A. Problem Formulation

According to (10), since $\mathbf{U}$ and $\mathbf{P}$ are known at the receiver, the target of the BS is to estimate $[\mathbf{X}]_{:,1}$ and K_a from the observation $\mathbf{Y}$, as shown in (7). Hence, the detection problem at the BS can be formulated as follows

$$\begin{aligned} \min_{\mathbf{X},K_a} \quad & \|\mathbf{Y}-\mathbf{P}\mathbf{X}\|_F^2 \\ s.t. \quad & (8), \quad (9), \quad \text{and} \quad K_a \ll N. \end{aligned} \tag{12}$$

The optimization problem (12) minimizes the mean square error (MSE) between $\mathbf{Y}$ and $\mathbf{P}\mathbf{X}$, which is equivalent to calculate the posterior mean of $\mathbf{X}$ under the Bayesian framework [56]. In the following, we denote $[\mathbf{X}]_{n,m}$ as $x_{n,m}$, where $n \in [N]$ and $m \in [M]$. Furthermore, the posterior mean of $x_{n,m}$, $\forall n, m$, can be expressed as

$$\widehat{x}_{n,m} = \int x_{n,m} p\left(x_{n,m}|[\mathbf{Y}]_{:,m}\right) dx_{n,m}, \tag{13}$$

where $p(x_{n,m}|[\mathbf{Y}]_{:,m})$ is the marginal distribution of $p([\mathbf{X}]_{:,m}|[\mathbf{Y}]_{:,m})$ and it can be written as

$$p\left(x_{n,m}|[\mathbf{Y}]_{:,m}\right) = \int_{\setminus x_{n,m}} p\left([\mathbf{X}]_{:,m}|[\mathbf{Y}]_{:,m}\right). \tag{14}$$

As discussed in Section III-B, to achieve higher quantization accuracy, more quantization bits are needed for higher dimensional VQ. This can result in a large dimension of N, which demands a computationally efficient algorithm to calculate the marginal distribution $p(x_n|\mathbf{y})$. To address this issue, the AMP algorithm can be adopted to obtain the approximate marginal distributions with relatively low complexity, while achieving near minimum mean square error (MMSE) performance [57].

B. Approximate Message Passing (AMP) Update Rules

According to the AMP algorithm [57], in the large system regime, the estimation problem (7) can be approximately decoupled into N scalar problems as

$$\mathbf{Y} = \mathbf{P}\mathbf{X} + \mathbf{Z} \;\rightarrow\; r_{n,m} = x_{n,m} + z_{n,m}, \;\forall n, m, \tag{15}$$

where $z_{n,m} \sim \mathcal{CN}(0, \varphi_{n,m})$, $r_{n,m}$ is the estimation of $x_{n,m}$. Hence, according to Bayes's theorem, the posterior distribution of $x_{n,m}$ can be expressed as

$$\begin{aligned} p\left(x_{n,m}|[\mathbf{Y}]_{:,m}\right) &\approx p\left(x_{n,m}|r_{n,m}\right) \\ &= \frac{1}{p\left(r_{n,m}\right)} p\left(r_{n,m}|x_{n,m}\right) p\left(x_{n,m}\right), \end{aligned} \tag{16}$$

where symbol "$\approx$" is due to the approximation. The likelihood function and the normalization term are, respectively, denoted as

$$p\left(r_{n,m}|x_{n,m}\right) = \mathcal{CN}\left(r_{n,m}; x_{n,m}, \varphi_{n,m}\right), \tag{17}$$

$$p\left(r_{n,m}\right) = \int p\left(r_{n,m}|x_{n,m}\right) p\left(x_{n,m}\right) dx_{n,m}. \tag{18}$$

During the i-th iterations of AMP, $(r_{n,m})^i$ and $(\varphi_{n,m})^i$, $\forall n, m$, are updated at the variable nodes of the factor graph as [58], [59]

$$(\varphi_{n,m})^i = \left(\sum\nolimits_{l=1}^{L} \frac{|[\mathbf{P}]_{l,n}|^2}{\sigma^2 + (V_{l,m})^i} \right)^{-1}, \tag{19}$$

$$(r_{n,m})^i = (\widehat{x}_{n,m})^i + (\varphi_{n,m})^i \sum_{l=1}^{L} \frac{[\mathbf{P}]_{l,n}^* \left([\mathbf{Y}]_{l,m} - (Z_{l,m})^i \right)}{\sigma^2 + (V_{l,m})^i}, \tag{20}$$

where $(V_{l,m})^i$ and $(Z_{l,m})^i$, $\forall l, m$, are updated at the factor nodes as

$$(V_{l,m})^i = \sum\nolimits_{n=1}^{N} |[\mathbf{P}]_{l,n}|^2 \, (\widehat{v}_{n,m})^i, \tag{21}$$

$$(Z_{l,m})^i = \sum_{n=1}^{N} [\mathbf{P}]_{l,n} (\widehat{x}_{n,m})^i - (V_{l,m})^i \frac{[\mathbf{Y}]_{l,m} - (Z_{l,m})^{i-1}}{\sigma^2 + (V_{l,m})^{i-1}}. \tag{22}$$

Note that $(\cdot)^i$ denotes its argument in the i-th AMP iteration, $\widehat{x}_{n,m}$ and $\widehat{v}_{n,m}$ denote the posterior mean and variance of $x_{n,m}$, respectively.

Furthermore, we delve into the prior distribution $p(x_{n,m})$ needed for the posterior estimation (16). According to (8), we model the prior distribution of $x_{n,1}$ and $x_{n,m}$, $m \in \{2, 3, ..., M\}$, as

$$p\left(x_{n,1}\right) = (1 - a_{n,1}) \, \delta(x_{n,1}) + \frac{a_{n,1}}{K_a} \sum_{s \in [K_a]} \delta(x_{n,1} - s), \tag{23}$$

$$p\left(x_{n,m}\right) = (1 - a_{n,m}) \delta\left(x_{n,m}\right) + a_{n,m} \mathcal{CN}\left(x_{n,m}; \mu_0, \tau_0\right), \tag{24}$$

respectively, where the sparsity indicator $a_{n,m}$ ($a_{n,1}$) equals to zero if $x_{n,m}$ ($x_{n,1}$) equals to zero, otherwise $a_{n,m}$ ($a_{n,1}$) equals to one. Also, in (23), if $a_{n,1} = 1$, we assume that $x_{n,1}$ can be any element from set $[K_a]$ with equal probability. Note that this assumption is an approximation since the BS cannot obtain the actual distribution of each element $x_{n,1}$, $\forall n \in [N]$. By adopting (23), the algorithm tends to obtain estimates of $x_{n,1}$ (if $a_{n,1} = 1$) that are integers from $[K_a]$.

Substituting (17), (18), and (23) into (16), the posterior mean and variance of $x_{n,1}$, $\forall n \in [N]$, can be evaluated as

$$\widehat{x}_{n,1} = \sum\nolimits_{x_{n,1} \in [K_a]} x_{n,1} p(x_{n,1}|r_{n,1}) dx_{n,1}, \tag{25}$$

$$\widehat{v}_{n,1} = \sum\nolimits_{x_{n,1} \in [K_a]} |x_{n,1}|^2 p(x_{n,1}|r_{n,1}) dx_{n,1} - |\widehat{x}_{n,1}|^2. \tag{26}$$

Proposition 1. *Substituting (17), (18), and (24) into (16), the posterior distribution of $x_{n,m}$, $m \in \{2, 3, ..., M\}$ can be reformulated as*

$$\begin{aligned} p(x_{n,m}|r_{n,m}) = (1 - \pi_{n,m}) \delta(x_{n,m}) \\ + \pi_{n,m} \mathcal{CN}\left(x_{n,m}; \mu_{n,m}, \tau_{n,m}\right), \end{aligned} \tag{27}$$

where we have

$$\mu_{n,m} = \left(\mu_0\varphi_{n,m} + \tau_0 r_{n,m}\right) \big/ \left(\varphi_{n,m} + r_{n,m}\right), \tag{28}$$

$$\tau_{n,m} = \left(\tau_0\varphi_{n,m}\right) \big/ \left(\varphi_{n,m} + r_{n,m}\right), \tag{29}$$

$$\mathcal{L}_{n,m} = \ln\frac{\varphi_{n,m}}{\tau_0+\varphi_{n,m}} - \frac{\left(r_{n,m}-\mu_0\right)^2}{\tau_0+\varphi_{n,m}} + \frac{\left|r_{n,m}\right|^2}{\varphi_{n,m}}, \tag{30}$$

$$\pi_{n,m} = a_{n,m} \big/ \left[a_{n,m} + (1-a_{n,m})\exp\left(-\mathcal{L}_{n,m}\right)\right]. \tag{31}$$

Proof. Please refer to the Appendix A. ■

Furthermore, according to (27), we can obtain the posterior mean and variance of $x_{n,m}$, $m \in \{2,3,...,M\}$, respectively, denoted by

$$\widehat{x}_{n,m} = \pi_{n,m}\mu_{n,m}, \tag{32}$$

$$\widehat{v}_{n,m} = \pi_{n,m}\left(\left|\mu_{n,m}\right|^2 + \tau_{n,m}\right) - \left(\widehat{x}_{n,m}\right)^2. \tag{33}$$

C. Parameter Estimation

1) Estimation of K_a: According to (25), we can obtain the estimated equivalent transmit signal vector $[\widehat{\mathbf{X}}_d]_{:,1}$, where the superscript d denotes the d-th block, $\forall d \in [D]$. In addition, as indicated in (8), $\|[\widehat{\mathbf{X}}_d]_{:,1}\|_1$ should equal to K_a, in the case of perfect detection. To improve the robustness, we propose to acquire $\widehat{K}_a$ as follows

$$\widehat{K}_a = \Psi\left\{\left\lfloor\left\|[\widehat{\mathbf{X}}_d]_{:,1}\right\|_1 + \frac{1}{2}\right\rfloor, d \in [D]\right\}, \tag{34}$$

where function $\Psi\{\cdot\}$ calculates the most frequently occurring element of its argument. The intuition of (34) is from majority voting [31], which improves the detection accuracy of K_a. We note that once $\widehat{K}_a$ is estimated, it is possible to go back and update the estimates of those blocks for which the estimated $\widehat{\mathbf{X}}_d$ did not correspond to $\widehat{K}_a$ active devices. We do not carry out this refinement in this paper.

2) Estimation of the activity indicators: The unknown parameter, denoted by $\boldsymbol{\theta} = \{a_{n,m}, \sigma^2, \tau_0, \mu_0, n \in [N], m \in [M]\}$, can be obtained by using the expectation maximization (EM) algorithm [60]. The EM algorithm is an iterative approach that effectively finds the maximum likelihood solutions for probabilistic models with unknown parameters [60]. Here, the EM update rules are as follows

$$\boldsymbol{\theta}^{i+1} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left\{\ln\, p\left(\mathbf{X},\mathbf{Y}\right)|\mathbf{Y};\boldsymbol{\theta}^i\right\}, \tag{35}$$

where $\boldsymbol{\theta}^i$ denotes the unknown parameter in the i-th iteration, $\mathbb{E}\{\cdot|\mathbf{Y};\boldsymbol{\theta}^i\}$ represents the expectation conditioned on the received signal $\mathbf{Y}$ under $\boldsymbol{\theta}^i$.

For example, to estimate the activity indicator $a_{n,m}$, $\forall n,m$, we partially differentiate $\mathbb{E}\left\{\ln\, p\left(\mathbf{X},\mathbf{Y}\right)|\mathbf{Y}\right\}$ with respect to $a_{n,m}$, and let it equal zero [60]. According to (16), the posterior distribution $p\left(x_{n,m}|[\mathbf{Y}]_{:,m}\right)$ can be approximated as $p\left(x_{n,m}|r_{n,m}\right)$, thereby simplifying the complexity of EM estimation. Hence, according to (23) and (24), the sparsity indicators $a_{n,1}$ and $a_{n,m}$, $m \in \{2,3,...,M\}$, can be obtained as follows:

$$\left(a_{n,1}\right)^{i+1} = \sum\nolimits_{x_{n,1}\in[K_a]} p\left(x_{n,1}|r_{n,1};\left(a_n\right)^i\right), \tag{36}$$

$$\left(a_{n,m}\right)^{i+1} = \pi_{n,m}, \tag{37}$$

Algorithm 2 Proposed AMP-DA Algorithm

Input: Received signals $\mathbf{Y}_d \in \mathbb{C}^{L\times M}$, modulation codebook $\mathbf{P} \in \mathbb{C}^{L\times N}$, quantization codebook $\mathbf{U} \in \mathbb{R}^{Q\times N}$, maximum iteration number I_0, and damping parameter τ.
Output: Estimated global model update $\widehat{\mathbf{s}}$.
1: $\forall d,n,l$: Initialize iteration index $i=1$, sparsity indicators $(a_n^d)^1 = 0.5$, noise variance $(\sigma^2)^1 = 100$, residual $R^0 = 100$, $(Z_{l,m}^d)^0 = [\mathbf{Y}_d]_{l,m}$, $(V_{l,m}^d)^0 = 1$, $(\widehat{x}_{n,m}^d)^1 = 0$, and $(\widehat{v}_{n,m}^d)^1 = 1$;
2: **for** $i = 1$ to I_0 **do**
3: Compute $(V_{l,m}^d)^i$, $(Z_{l,m}^d)^i$, $(\varphi_{n,m}^d)^i$, $(r_{n,m}^d)^i$ using (19)-(22);
%**Decoupling steps**
4: Damping: $(V_{l,m}^d)^i = \tau(V_{l,m}^d)^{i-1} + (1-\tau)(V_{l,m}^d)^i$, $(Z_{l,m}^d)^i = \tau(Z_{l,m}^d)^{i-1} + (1-\tau)(Z_{l,m}^d)^i$;
% **Denoising steps**
5: Compute $(\widehat{x}_{n,1}^d)^i$ and $(\widehat{v}_{n,1}^d)^i$ using (25) and (26);
6: Compute $(\widehat{x}_{n,m}^d)^i$ and $(\widehat{v}_{n,m}^d)^i$ using (32) and (33), $m \in \{2,...,M\}$;
%Parameter update:
7: Update $(a_n^d)^i$ using (38), update $(\sigma^2)^i$, $(\mu_0^d)^i$, and $(\tau_0^d)^i$;
8: Calculate residual: $R^i = \frac{1}{LD}\sum\limits_{d\in[D]}\left\|\mathbf{Y}_d - \mathbf{P}(\widehat{\mathbf{X}}_d)^i\right\|_F$;
9: $i = i+1$;
10: **if** $i > 15$ and $R^{i+1} \geq R^i$ **then**
11: **break**; {End iterations}
12: **end if**
13: **end for**
14: $\forall d$:Estimated equivalent transmit signal $\widehat{\mathbf{x}}_d = [(\widehat{\mathbf{X}}_d)^i]_{:,1}$;
%Estimate the number of active devices:
15: Estimate $\widehat{K}_a$ using (34);
%Model aggregation:
16: Obtain global model update $\widehat{\mathbf{s}} = \frac{1}{\widehat{K}_a}\left(\mathbf{I}_D \otimes \mathbf{U}\right)\left[\widehat{\mathbf{x}}_1^T, ..., \widehat{\mathbf{x}}_D^T\right]^T$.

respectively. To exploit the row sparsity as indicated in (9), we further average the activity indicators on M antennas as

$$\left(a_n\right)^{i+1} = \frac{1}{M}\sum\nolimits_{m=1}^{M}\left(a_{n,m}\right)^{i+1}, \forall n. \tag{38}$$

3) Estimation of other parameters: Furthermore, we apply the EM update rules (35) to estimate other unknown parameters. In particular, to estimate the noise variance, we partially differentiate $\mathbb{E}\left\{\ln\, p\left(\mathbf{X},\mathbf{Y}\right)|\mathbf{Y}\right\}$ with respect to σ^2 and let it equal zero [60]. Similarly, calculations can be done to obtain τ_0 and μ_0. Note that only (24) includes τ_0 and μ_0, suggesting that the messages from the first antenna do not contribute to the estimation of τ_0 and μ_0. Finally, we obtain the EM update rules of the noise variance σ^2, the unknown τ_0, and μ_0 as $(\sigma^2)^{i+1} = \frac{1}{MN}\sum\limits_{l,m=1}^{L,M}\left[\frac{\left([\mathbf{Y}]_{l,m} - (Z_{l,m})^i\right)^2}{\left(1+\frac{(V_{l,m})^i}{(\sigma^2)^i}\right)^2} + \frac{(\sigma^2)^i(V_{l,m})^i}{(V_{l,m})^i+(\sigma^2)^i}\right]$, $(\tau_0)^{i+1} = \frac{\sum_{m=2}^{M}\sum_{n=1}^{N}(\pi_{n,m})^i\left[\left((\mu_0)^{i+1} - (\mu_{n,m})^i\right)^2 + (\tau_{n,m})^i\right]}{\sum_{m=2}^{M}\sum_{n=1}^{N}(\pi_{n,m})^i}$, and $(\mu_0)^{i+1} = \frac{\sum_{m=2}^{M}\sum_{n=1}^{N}(\pi_{n,m})^i(\mu_{n,m})^i}{\sum_{m=2}^{M}\sum_{n=1}^{N}(\pi_{n,m})^i}$, respectively.

D. Proposed AMP-DA Algorithm

Based on (19)-(38), we summarize the proposed AMP-DA algorithm in **Algorithm** 2. The details are explained as follows. In line 1, we initialize the sparsity indicators a_n^d, the variables $V_{l,m}^d$, $Z_{l,m}^d$, the posterior mean $\widehat{x}_{n,m}^d$, and posterior variance $\widehat{v}_{n,m}^d$, $\forall d,n,l,m$. The iteration starts in line 2. Specifically, lines 3-6 correspond to the AMP operation. In the i-th iteration of the AMP decoupling step (line 3), $V_{l,m}^d$,

$Z_{l,m}^d$, $\varphi_{n,m}^d$, and $r_{n,m}^d$, are calculated according to (21), (22), (19), and (20), respectively. Furthermore, a damping parameter τ is adopted in line 4 to prevent the algorithm from diverging [61]. In addition, in the AMP denoising steps (line 5 and line 6), we calculate the posterior mean $\widehat{x}_{n,1}^d$ and the corresponding posterior variance $\widehat{v}_{n,1}^d$, of the i-th iteration, by using (25) and (26), respectively. In addition, for $m \in \{2, ..., M\}$, $\widehat{x}_{n,m}^d$ and $\widehat{v}_{n,m}^d$ are calculated by using (32) and (33). Accordingly, EM operation is used to update the sparsity indicators a_n^d, the noise variance σ^2, and the mean and variance of the prior distribution μ_0^d and τ_0^d, respectively, in line 7. The residual of the i-th iteration is calculated in line 8. Then, the iteration restarts in line 3 until the maximum iteration number I_0 is reached. Otherwise, if line 10 is triggered, i.e., the residual does not goes down after at least 15 steps, we stop iterations. After the iteration, we acquire the estimated equivalent transmit signal $\widehat{\mathbf{x}}_d = [(\widehat{\mathbf{X}}_d)^i]_{:,1}$, $\forall d$, in line 14. According to (34), in line 15, we estimate the number of active devices $\widehat{K}_a$ based on $\widehat{\mathbf{x}}_d$, $\forall d$. Finally, according to (10) and Section III-B, we obtain the global model update $\widehat{\mathbf{s}}$ in line 16 by using $\widehat{\mathbf{x}}_d$ and $\widehat{K}_a$, $\forall d$, and the weight of the ML model will be updated using $\widehat{\mathbf{s}}$.

Remark 3. *Since only $[\widehat{\mathbf{X}}_d]_{:,1}$ matters for the computing task (10), the digital model aggregation can be done by only using the received signal at the first antenna of the BS, i.e., running Algorithm 2 except line 6. If multiple antennas are available at the BS, the proposed AMP-DA algorithm improves the estimation accuracy of $[\mathbf{X}_d]_{:,1}$ by exploiting the row sparsity, as described in (9). As a result, the training accuracy of the proposed MD-AirComp improves with multi-antenna BS, which will be verified in the simulations.*

E. Computational Complexity

The computational complexity of the proposed AMP-DA algorithm mainly comes from the complex-valued matrix multiplications, i.e., (19)-(22), (25), (26), (32), and (33), at each iteration. The overall computational complexity at one FEEL training round is of the order of $\mathcal{O}(I_0 LNMD)$, which equals to $\mathcal{O}(I_0 LM2^J W/Q)$. It can be seen that the overall complexity scales linearly with the quantization level N, which is appealing for efficient detection with high-dimensional VQ along with a large quantization level. The computational complexity of our proposed AMP-DA algorithm can be reduced further by employing deep unfolding schemes [62], which is left as a future research direction.

V. Convergence Analysis

In this section, we present the theoretical convergence results of our proposed MD-AirComp. We first introduce some assumptions needed for the proof. For simplicity, $[\widehat{\mathbf{X}}_d]_{:,1}$ is denoted by $\mathbf{x}_d$ in this and the following sections.

Assumption 1 (Smoothness)**.** *Each loss function on the k-th device, $F_k(\cdot)$ is $\overline{L}$-smooth, i.e., $\forall \mathbf{w}, \overline{\mathbf{w}} \in \mathbb{R}^W$,*

$$|F_k(\mathbf{w}) - F_k(\overline{\mathbf{w}}) - \langle \nabla F_k(\overline{\mathbf{w}}), \mathbf{w} - \overline{\mathbf{w}} \rangle| \leq \frac{\overline{L}}{2} \|\mathbf{w} - \overline{\mathbf{w}}\|^2. \tag{39}$$

This indicates the $\overline{L}$-gradient Lipschitz condition, i.e., $\|\nabla F_k(\mathbf{w}) - \nabla F_k(\overline{\mathbf{w}})\| \leq \overline{L} \|\mathbf{w} - \overline{\mathbf{w}}\|$, where $\overline{L}$ is a constant.

Assumption 2 (Bounded Gradient)**.** *Each loss function on the k-th device $F_k(\mathbf{w})$ has G-bounded stochastic gradient on ℓ_2 norm, i.e., for all $\boldsymbol{\varepsilon}$, we have $\|\nabla F_k(\mathbf{w}, \xi)\| \leq G$.*

Assumption 3 (Bounded Variance)**.** *Each stochastic gradient on the k-th device has a bounded local variance, i.e., for all $\mathbf{w}, k \in [K]$,we have $\mathbb{E}\left[\|\nabla F_k(\mathbf{w}, \boldsymbol{\varepsilon}) - \nabla F_k(\mathbf{w})\|^2\right] \leq \sigma_l^2$, and the loss function on each device has a global variance bound, $\frac{1}{K}\sum_{k=1}^K \|\nabla F_k(\mathbf{w}) - \nabla f(\mathbf{w})\|^2 \leq \sigma_g^2$.*

Note that $\boldsymbol{\varepsilon}$ denotes a random local data sample. The bounded local variance represents the randomness of stochastic gradients, and the bounded global variance represents data heterogeneity among devices. Specifically, $\sigma_g = 0$ corresponds to the i.i.d setting, in which the datasets of all the clients have the same distribution. Assumptions 1-3 are widely adopted in federated optimization problems [25], [28], [63], [64].

Assumption 4 (Biased Quantizer)**.** *Consider a biased operator $\mathcal{C} : \mathbb{R}^Q \to \mathbb{R}^Q$: $\forall \mathbf{z} \in \mathbb{R}^Q$, there exists a constant $0 \leq q \leq 1$ such that*

$$\|\mathcal{C}(\mathbf{z}) - \mathbf{z}\| \leq q\|\mathbf{z}\|, \forall \mathbf{z} \in \mathbb{R}^Q. \tag{40}$$

Assumption 4 is a standard assumption for biased compressors, which was adopted to model scaled-sign and top-k compressors in the literature. For such biased quantizer, there exists a constant γ such that, for global training round $t \in [T]$, we have

$$\left\| \mathcal{C}\left(\tfrac{1}{K}\sum_{k=1}^K [\Delta_k^t + \mathbf{e}_k^t] \right) - \tfrac{1}{K}\sum_{k=1}^K \mathcal{C}\left(\Delta_k^t + \mathbf{e}_k^t\right) \right\| \leq \gamma \left\| \tfrac{1}{K}\sum_{k=1}^K \Delta_k^t \right\|, \tag{41}$$

which bounds the difference between the average of the quantized values and the quantized value of the averages [64].

Assumption 5 (AMP Estimation)**.** *We assume that the number of non-zero elements in $\mathbf{x}_d$, i.e., $\|\mathbf{x}_d\|_0$, $\forall d \in [D]$, is smaller than the length of the modulation codeword L. This enables AMP algorithm to reconstruct $\mathbf{x}_d$ with an estimation MSE of $\mathbb{E}\|\widehat{\mathbf{x}}_d - \mathbf{x}_d\|^2 = N\sigma_n^2$, $\forall d$ [57]. In addition, we can assume the perfect estimation of the number of active devices K_a.*

As described in (8), $\|\mathbf{x}_d\|_0 \leq K_a$. In practice, the communication system can have prior knowledge about the maximum number of active devices K_a at each iteration, and can select an L that satisfies Assumption 5. Note that this assumption simplifies the analysis, while we will describe how the estimation error affects the performance later.

In the following, we will show the convergence results[4] of the proposed MD-AirComp-based FEEL scheme.

Theorem 1. *Under Assumptions 1-5, if the local learning rate η_l satisfies the following condition: $\eta_l \leq \min\left\{\frac{1}{8K\overline{L}}, \frac{1}{\eta\overline{L}T_l(3+2C_1)}\right\}$, then the iterates of MD-AirComp-*

[4]For simplicity, we only present the convergence guarantee for full participation of devices. Note that the theoretical analysis can be easily extended to partial participation according to [64].

based FEEL satisfy

$$\min_{t\in[T]} \mathbb{E}[\|\nabla f(\mathbf{w}^t)\|^2] \leq 4\left[\frac{f_0 - f_*}{\eta\eta_l T_l T} + \Xi + \Psi\right], \tag{42}$$

where $f_0 \triangleq f(\mathbf{w}^0)$, $f_* \triangleq \min_{\mathbf{w}} f(\mathbf{w})$, $\Xi \triangleq (\frac{3}{2}+C_1)\eta\overline{L}\sigma_l^2\frac{\eta_l}{K} + \frac{5\eta_l^2 T_l \overline{L}^2}{2}(\sigma_l^2 + 6T_l\sigma_g^2)$, $\Psi \triangleq \frac{\eta\eta_l \overline{L} C_0 T_l G^2 N^2 \sigma_n^2}{K^2}$, $C_1 \triangleq \frac{2(q+\gamma)^2}{(1-q^2)^2}$, *and* $C_0 \triangleq \frac{2(1+q^2)^2}{(1-q^2)^2}$.

Proof. Please refer to the Appendix B. ■

Remark 4. *It can be seen that the upper bound for* $\min_{t\in[T]} \mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2]$ *contains three terms: 1) The first term, is due to the initialization error, i.e.* $f_0 - f_*$. *2) The second term* Ξ *relates to the local stochastic variance* σ_l *and global variance* σ_g. *In the i.i.d setting, we have zero global variance, i.e.,* $\sigma_g = 0$, *and* Ξ *will be smaller and less dependent on the number of local training steps* T_l. *Note that* C_1 *and* C_0 *are two constants related to the quantization error. If the quantization error increases, i.e.,* q *(*$0 < q < 1$*) increases,* C_1 *and* C_0 *will increase as well. 3) Finally, the third term* Ψ *is due to the wireless communication errors, i.e.,* $\|\widehat{\mathbf{x}}_d^t - \mathbf{x}_d^t\|^2$, $\forall t, d$. *Note that if the wireless channel is noiseless, we have* $\Psi = 0$. *In practice, a lower detection MSE at the BS will reduce* Ψ, *resulting in a tighter convergence bound.*

Corollary 1. *Suppose we choose the global learning rate* $\eta = \Theta(\sqrt{K})$ *and local learning rate* $\eta_l = \Theta(\frac{1}{T_l\sqrt{T}})$, *when* T *is sufficiently large. With this choice of the learning rates, the convergence rate for MD-AirComp-based FEEL is independent of the communication errors and quantization effects, and satisfies*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla f(\mathbf{w}_t)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \tag{43}$$

VI. Simulation Results

A. Parameter setting

In this section, we evaluate the performance of the proposed MD-AirComp scheme by considering FEEL for the image classification task on the commonly adopted CIFAR-10 dataset [5]. The CIFAR-10 dataset has a training set of 50,000 and a test set of 10,000 colour images of size 32 × 32 in 10 classes. We consider $K = 40$ devices (unless otherwise specified) in the cell collaboratively training a neural network (the structure follows the ResNet [65]) with 269722 (i.e., $W \approx 2.7 \times 10^5$) parameters. We consider non-i.i.d. local training data across the devices as follows [17]: 1) Each one of the 40 devices is randomly assigned 250 samples from the dataset; 2) The remaining 40,000 samples are sorted by their labels and grouped into 40 shards of size 1000. Then, each device is assigned one shard[5]. Furthermore, we set the number of local training iterations to $E = 3$, the mini-batch size to 20, the local learning rate to $\eta_l = 0.01$, and the global learning rate to $\eta = 1$. The metric to evaluate the performance is the *test accuracy* of the model on the test dataset.

We adopt the K-means++ algorithm for VQ to learn the quantization codebook [67]. Since the distribution of the model updates changes from one training round to the next [27], we also update the VQ codebook to achieve better performance. We assume that the BS has available a small local dataset and performs local training as well. In each communication round, the BS obtains a local model update with error accumulation $\mathbf{s}' \in \mathbb{R}^W$ as described in Section III-A. Then the BS uses its own local update $\mathbf{s}'$ to determine the VQ codebook. In particular, the D data samples $\{[\mathbf{s}']_{(d-1)Q+1:dQ}\}_{d=1}^D$ are used by the VQ algorithm to obtain the quantization codebook $\mathbf{U} \in \mathbb{R}^{Q\times N}$, where $N = 2^J$ is the number of clusters. Then, the BS broadcasts[6] codebook $\mathbf{U}$ to the devices together with the global model. Note that we pad zeros at the end of $\mathbf{s}'$ to ensure that $\mathbf{s}'$ is divisible by Q. At the receiver, those padded elements will be deleted. Since W is much larger than Q, the effects of padding can be negligible. Different quantization parameters J and Q are tested in Fig. 3. We remark that the VQ quantizer is a biased quantizer as prescribed by Assumption 4.

As for the wireless transmission, we assume that 30% of the devices are active in each round. Since the BS does not exactly know K_a, we use $0.4 \times K$ instead of the actual K_a in (23) as a prior information. We set SNR $= 20$ dB unless otherwise specified. For Algorithm 2, the damping parameter is set to $\tau = 0.3$, and the maximum iteration number to $I_0 = 50$. The elements of modulation codebook $\mathbf{P} \in \mathbb{C}^{L\times N}$ obey i.i.d. complex Bernoulli distribution, which satisfies the restricted isometry property and is commonly adopted in CS [42]. In addition, as mentioned in Remark 2, to avoid extremely small SNR at the first receive antenna, we set $\epsilon_h = 0.14$ as the threshold of pre-equalization. We note that in practice, this threshold can be determined by the transmit power constraint.

For comparison, we consider the following schemes. **IFed**: IFed is the acronym for *ideal federated learning*, where the communication channels are assumed ideal and there is no quantization distortion. The IFed algorithm adopted for the simulations is FedAvg[7] [50]. **PA** (J, Q): Here, PA is the acronym for *perfect aggregation* at the receiver, while the transmitter is the same as the proposed MD-AirComp scheme. The number of quantization bits and the VQ dimension are denoted by (J, Q). **MD-AirComp** (J, Q, L): The proposed MD-AirComp scheme, where the number of quantization bits, the VQ dimension, and the length of modulation codewords are denoted by (J, Q, L). Different numbers of antennas M at the BS are also tested in the simulations. **OBDA**: The state-of-the-art digital AirComp scheme with one-bit quantization introduced in [31]. Note that the main parameters of the OBDA scheme are adopted as released in [31]. However, we observed that learning rate has a significant impact on the convergence of OBDA, especially when dealing with non-i.i.d. local training data. To achieve a better test accuracy of

[5]Note that the earth mover's distance (EMD) for i.i.d. local datasets equals to zero [66]. In our simulation, we have EMD $= 1.44$, which represents highly non-i.i.d. local datasets [66].

[6]Note that the number of real values in $\mathbf{U}$ is NQ. Considering $N = 256$ and $Q = 20$, we can obtain $\frac{NQ}{W} \approx 1.9\%$, which implies that the communication cost of broadcasting $\mathbf{U}$ is very small in comparison to broadcasting the updated model parameters.

[7]Note that all the schemes are built upon FedAvg in our simulation for the sake of fairness, while FedAvg can be replaced by other optimization algorithms, e.g., adaptive federated optimizer (FedAdam) [63], for better convergence.

OBDA, we carefully choose the cosine annealing scheduler [68] for learning rate adaptation, where the initial learning rate is 0.001, the minimal learning rate is 0.0001, and the cosine function follows a period of 1400. **FSK-MV**: FSK-MV is the acronym for *frequency shift keying-based majority vote* proposed in [32], which is the best performing FSK-based method in terms of communication overhead [32], [33]. Based on SignSGD, the learning parameters of FSK-MV are the same as OBDA. At the transmitters, each binary element of the model update vector is transmitted by sending a fixed power pilot signal on either of the two pre-allocated subcarrier frequencies. No pre-equalization needed at the transmitter. The receiver detects the signal power on the two subcarriers for majority voting.

B. Communication Overhead

Furthermore, we compare the communication overhead of different schemes. We consider an orthogonal frequency-division multiplexing (OFDM) system, where P subcarriers are available. Due to the large number of devices, the communication overhead is dominated by the UL communications. We consider the UL communication overhead of one training round, i.e., K_a active devices quantize and transmit their model update (W elements) to the BS. With one-bit quantization, the number of symbols transmitted by OBDA[8] and FSK-MV is W and $2W$, respectively. It can be seen that FSK-MV does not require CSI at a scarifies of larger communication overhead. Furthermore, the proposed MD-AirComp transmits $\lceil \frac{W}{Q} \rceil L$ symbols [9]. Our simulation results in the following show that L can be set smaller than Q, making the communication overhead of MD-AirComp smaller than that of the OBDA and FSK-MV schemes. In addition, for the above digital AirComp schemes, since all the devices share P subcarriers, the number of time slots needed equals to the number of transmitted symbols divided by P.

As for conventional digital communication, we consider the orthogonal frequency-division multiple access (OFDMA) and the quadrature amplitude modulation (QAM). For fair comparison, the same VQ at the transmitter is considered. After quantization, the index of each block (J bits) is modulated to a QAM symbol. Hence, each device needs to transmit $\lceil \frac{W}{Q} \rceil$ symbols. We assume each device is uniformly allocated with $\frac{P}{K}$ subcarriers. Hence, the number of time slots needed is $\lceil \frac{W}{Q} \rceil \frac{K}{P}$. We denote the conventional digital communication scheme as 'VQ + OFDMA', and we compare the number of time slots (i.e., transmission delay), of different schemes in Table I. As mentioned in Remark 1, $K_a \ll K$, while L has values close to, or possibly smaller (due to devices collision) than, K_a, the transmission delay of MD-AirComp is obviously far smaller than the 'VQ + OFDMA' scheme. Numerical results in Table I verify the lower communication overhead of the proposed MD-AirComp scheme. Note that the best performance that "VQ + OFDMA" scheme can achieve is the PA benchmark, if considering perfect QAM demodulation.

[8]Considering binary phase shift keying (BPSK) symbols as used in [31].

[9]In our conference paper [1], real-value modulation codebook $\mathbf{P} \in \mathbb{R}^{L \times N}$ was considered. In this case, if using I/Q channels, transmitting L-real values only costs $\lceil L/2 \rceil$ symbol. While, in this work, we extend the setting to a more general case, i.e., $\mathbf{P} \in \mathbb{C}^{L \times N}$.

TABLE I. UL communication overhead comparison

Schemes	Number of Time Slots	Numerical Values[1]
VQ + OFDMA	$\lceil \frac{W}{Q} \rceil \frac{K}{P}$	527
FSK-MV [32]	$\frac{2W}{P}$	527
OBDA [31]	$\frac{W}{P}$	264
MD-AirComp	$\lceil \frac{W}{Q} \rceil \frac{L}{P}$	264 ($L = 20$) 198 ($L = 15$)

[1] These values are calculated and rounded up, referring to the simulation parameters, i.e., $W = 269722$, $Q = 20$, $L = 20$ (or $L = 15$) and $K = 40$. The number of subcarriers $P = 1024$.

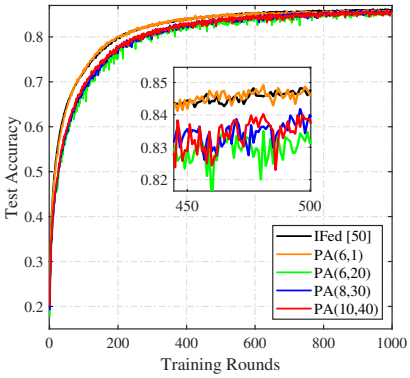
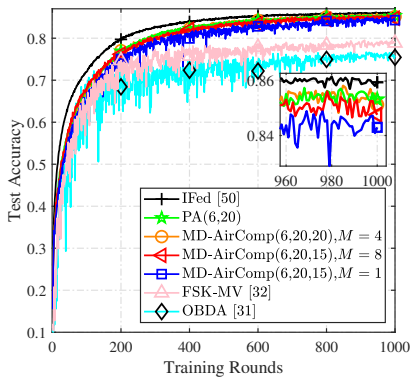


Fig. 3. Test accuracy vs training rounds under different VQ parameters.

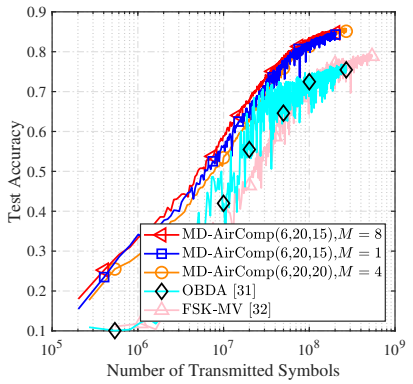
C. Results of FEEL

In Fig. 3, we present how the test accuracy changes with the training rounds for IFed and PA (J, Q) for different J and Q values. We observe that scalar quantization (i.e., $Q = 1$) with $J = 6$ achieves nearly the same test accuracy as the IFed benchmark. In addition, the test accuracy under other VQ parameters nearly overlap. Note that when the VQ dimension Q increases, we need more quantization bits J to convey the corresponding update with reasonable accuracy [60]. In the remaining simulations, we set $J = 6$ and $Q = 20$ to test the performance of the proposed MD-AirComp scheme.

As shown in Fig. 4(a), the test accuracy of the proposed MD-AirComp scheme converges to the IFed benchmark. We see slower convergence and some performance loss in the final test accuracy of OBDA and FSK-MV compared with MD-AirComp, which is expected considering the one-bit quantization loss in these schemes. FSK-MV converges faster and improves the final accuracy compared with OBDA, while requiring a larger communication overhead. Furthermore, it can be seen that "MD-AirComp (6, 20, 15), $M = 1$" performs worse than the PA (6, 20) due to the communication error under limited observation L. Fortunately, by exploiting the multiple antennas at the BS, we can reduce the communication error without increasing the communication resources. For example, it is shown that "MD-AirComp (6, 20, 15), $M = 8$" achieves nearly the same accuracy as that of the PA (6, 20). From Fig. 4(b), it can be seen that the MD-AirComp scheme is more communication efficient than ever the OBDA scheme, that is, to achieve the same test accuracy it requires the

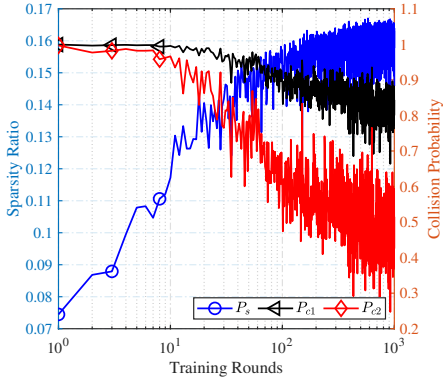


(a)

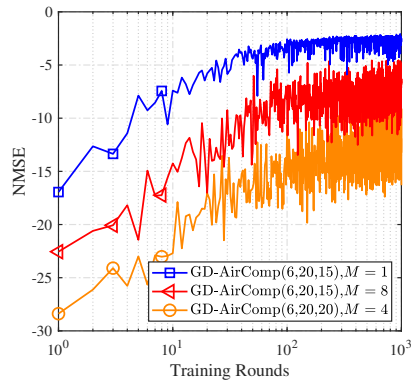


(b)

Fig. 4. FEEL performance comparison of the proposed MD-AirComp scheme with the benchmarks: (a) Test accuracy vs training rounds; (b) Test accuracy vs the number of transmitted symbols.



(a)



(b)

Fig. 5. Wireless transmission performance of the proposed MD-AirComp scheme: (a) Sparsity ratio and device collision probability vs training rounds; (b) Detection NMSE vs training rounds.

transmission of fewer symbols over the channel. In addition, although the MD-AirComp (6, 20, 20) achieves slightly better test accuracy as shown in Fig. 4(a), it is less communication efficient than "MD-AirComp (6, 20, 15), $M = 8$". This reveals that we can reduce the code-length L, and hence, increase the bandwidth efficiency of MD-AirComp if more antennas are available at the receiver end.

D. Results of Wireless Transmission

In Fig. 5(a), we investigate the average sparsity ratio $P_s \triangleq \frac{\sum_{d=1}^{D} \|\mathbf{x}_d\|_0}{ND}$, and collision probability of MD-AirComp (6, 20, 20) versus the global training rounds[10]. We denote the probability of device collision occurred in D blocks by $P_{c1} \triangleq 1 - P\{d \in [D] : \|\mathbf{x}_d\|_0 = K_a\}$. In addition, the probability that more than two devices collide is denoted by $P_{c2} \triangleq P_{c1} - P\{d \in [D] : \|\mathbf{x}_d\|_0 = K_a - 1\}$. To enhance the clarity of the presentation, we employ logarithmic scaling on the x-axis. It can be seen that with fewer device collisions, the sparsity ratio increases along training rounds. The reason is that the variance of model updates gradually decrease as the training progresses [27]. Hence, with the same number of quantization levels, the values of different devices' model updates become more distinguishable, resulting in a decrease in the collision probability.

[10]We observed that MD-AirComp (6, 20, 20) and MD-AirComp (6, 20, 15) have similar average sparsity ratio and collision probability during the training. For simplicity, we only present the related results of MD-AirComp (6, 20, 20).

For given number of observations L and signal dimension N, the reconstruction performance of CS algorithms gets worse if the sparsity ratio of the signal increases, i.e., as the signal becomes less sparse. The metric to evaluate the signal reconstruction performance is normalized MSE (NMSE), which is expressed as

$$\text{NMSE} = 10\log_{10}\left(\frac{\sum_{d=1}^{D} \|[\mathbf{X}_d]_{:,1} - [\widehat{\mathbf{X}}_d]_{:,1}\|_2^2}{\sum_{d=1}^{D} \|[\mathbf{X}_d]_{:,1}\|_2^2}\right). \quad (44)$$

From Fig. 5(b), we can observe that NMSE decreases due to the increase in the sparsity ratio. In addition, NMSE of "MD-AirComp (6, 20, 15), $M = 8$" is better than that of "MD-AirComp (6, 20, 1), $M = 1$", which verifies the performance improvement shown in Fig. 4(a). Furthermore, NMSE of "MD-AirComp (6, 20, 20), $M = 4$" is the best due to more observations. Although NMSE of "MD-AirComp (6, 20, 15), $M = 8$" is smaller than -10 dB when the training rounds is more than 200, its test accuracy is as good as the PA scheme due to the error correction property of FL.

Fig. 6 illustrates the K_a estimation error. We consider the probability mass function (PMF) of the majority voting (MV) method and the mean estimation (ME) method for different MD-AirComp schemes. Note that the PMF results are calculated based on one realization of the scheme and on 1000 samples (training rounds). It is obvious that the MV method estimates K_a much better than the ME scheme. Furthermore, due to the worse NMSE performance, the K_a estimation error of "MD-AirComp (6, 20, 15), $M = 1$" is much worse than that

TABLE II. Performance comparison under various SNRs

SNR	0 dB	5 dB	20 dB
Test Accuracy ($t = 1000$)	0.8402	0.8495	0.8513
NMSE ($t = 1000$)	-2.75 dB	-5.33 dB	-16.30 dB
NMSE ($t = 10$)	-3.67 dB	-7.05 dB	-22.7 dB

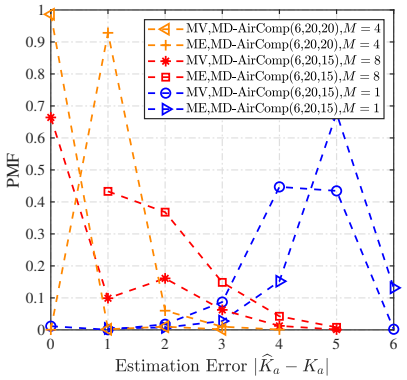


Fig. 6. The PMF of the K_a estimation error over 1000 training rounds.

of "MD-AirComp (6, 20, 15), $M = 8$". Furthermore, Table II compares the performance of "MD-AirComp (6, 20, 20), $M = 4$" scheme under various SNRs. It can be observed that the proposed MD-AirComp scheme still works well in low SNR, i.e., SNR $= 0$ dB or 5 dB, with a negligible loss in the test accuracy.

Fig. 7 investigates the performance of the proposed MD-AirComp in a larger cell with $K = 100$ devices. This only changes the local training dataset setting, other parameters are the same unless otherwise specified. Specifically, each one of the 100 devices is randomly assigned 100 samples from the dataset, while the remaining 40,000 samples are sorted by their labels and grouped into 100 shards of size 400. We first evaluate the average number of non-zero elements, i.e., the average $\|\mathbf{x}_d\|_0$, denoted as $\overline{\|\mathbf{x}_d\|_0} \triangleq \frac{\sum_{d=1}^{D}\|\mathbf{x}_d\|_0}{D}$. We consider different device activity ratios, i.e., 10%, 30%, 50%, and run the PA (6, 20) scheme to calculate the average $\|\mathbf{x}_d\|_0$. It can be seen that $\overline{\|\mathbf{x}_d\|_0}$ can be much smaller than K_a, indicating a high collision rate among active devices (similar conclusion can be seen in Fig. 5(a)). We reiterate that the collision helps to reduce the communication overhead of the proposed MD-AirComp scheme, and L can be smaller than K_a due to the collisions. As can be verified in Fig. 7, considering $K_a = 30$, "MD-AirComp (6, 20, 25), $M = 4$", i.e., $L = 25 < K_a$, approaches the PA(6, 20) benchmark very well.

Furthermore, in Fig. 8, we study effects of the various terms in the convergence analysis and Theorem 1 through simulations. According to Corollary 1, with the proper choice of the local and global learning rates, the convergence rate for MD-AirComp-based FEEL is $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ and independent of the communication errors and quantization effects. This is confirmed in Fig. 8 as we see a roughly similar slope in convergence of IFed, PA, and MD-AirComp schemes. Please note that, IFed and PA are communication error-free, i.e., the third term Ψ in (42) is zero. IFed is also free of the quantization losses, i.e., $q = 0$, indicating a smaller C_1 and a smaller second term Ξ than that of PA. MD-AirComp has the third term that relate to both communication error σ_n^2 and

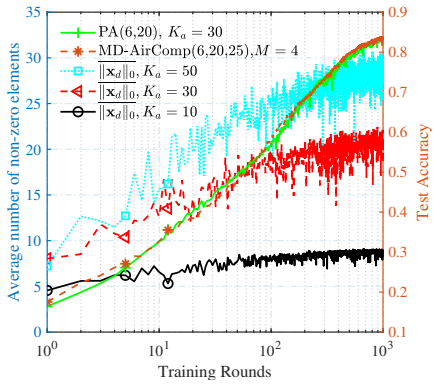


Fig. 7. Performance evaluation when $K = 100$. Left y-axis: Average number of non-zero elements $\|\mathbf{x}_d\|_0$ under different K_a; Right y-axis: Test accuracy of "PA(6, 20)" and "MD-AirComp(6, 20, 25)", where $K_a = 30$.

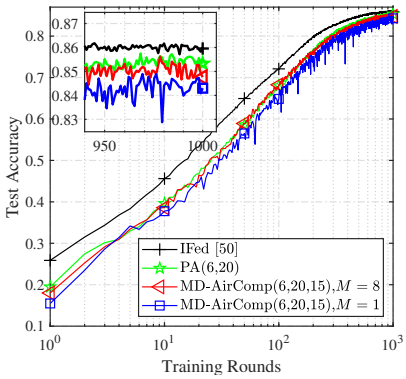


Fig. 8. Test accuracy versus the logarithmic training rounds.

quantization error C_0. From Fig. 8, the quantization effects seem to be more destructive. On the other hand, our proposed MD-AirComp technique seems to tackle the communication errors efficiently thereby achieving a performance very close to the PA benchmark.

E. Robustness to Imperfections

For simplicity, we focus on the received signal at the first antenna of the BS (pre-equalization applied at the device).

1) Imperfect Channel Estimation: Here, we use $\mathbf{y}_d \in \mathbb{C}^L$ and $\mathbf{z}_d \in \mathbb{C}^L$ to denote the received signal and the noise of the d-th, $\forall d \in [D]$, block at the BS, respectively. In addition, denote h_k and $\widehat{h}_k$ as the actual channel and the estimated channel of the k-th device, respectively. Then, we can have the received signal model as

$$\begin{aligned}\mathbf{y}_d &= \sum\nolimits_{k\in\mathcal{S}_a} \widetilde{h}_k \mathbf{P}\mathbf{x}_k^d + \mathbf{z}_d \\ &= \mathbf{P}\mathbf{x}_d + \sum\nolimits_{k\in\mathcal{S}_a} (\widetilde{h}_k - 1)\mathbf{P}\mathbf{x}_k^d + \mathbf{z}_d, \\ &= \mathbf{P}\mathbf{x}_d + \mathbf{z}_e + \mathbf{z}_d,\end{aligned} \tag{45}$$

where $\widetilde{h}_k = \frac{h_k}{\widehat{h}_k} \in \mathbb{C}$ denotes the residual after pre-equalization, $\mathbf{z}_e$ is the error term caused by imperfect downlink channel estimation. If $\widehat{h}_k = h_k$, $\forall k \in \mathcal{S}_a$, $\mathbf{z}_e = 0$ and (45) simplifies to (7). According to (45), even though there is channel estimation error, by seeing $\mathbf{z}_e + \mathbf{z}_d$ as a whole, we can still employ **Algorithm 2** to estimate $\mathbf{x}_d$ and then aggregate the

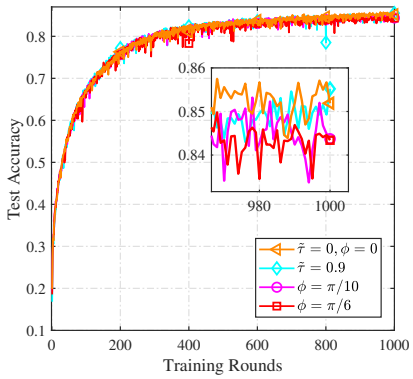
model. Note that we assume the noise obeys complex Gaussian distribution, and we estimate the mean and variance of such distribution iteratively by using EM algorithm. Similar to [17], [19], [20], we set the amplitude of $\widetilde{h}_k$ as 1, $\forall k$, and focus on the phase error ϕ_k. Specifically, we consider $\widetilde{h}_k = e^{\phi_k}$ and $\phi_k, \forall k \in [K]$, is uniformly distributed in the set (0, ϕ), where ϕ is the maximum phase offset.

2) Imperfect Time Synchronization: To describe the effect of imperfect time synchronization, we follow the time-domain signal model of [17]. The received signal $r(\widetilde{t})$ at the BS can be written as

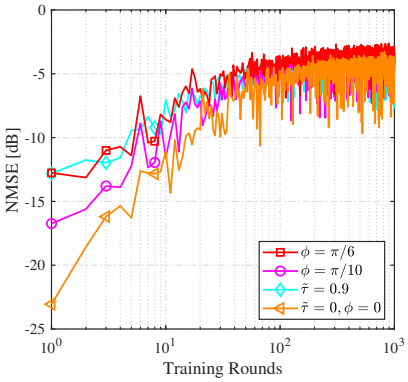
$$r(\widetilde{t}) = \sum_{k\in\mathcal{S}_a} \sum_{l=1}^{L} \left[\mathbf{P}\mathbf{x}_k^d\right]_l \text{rect}(\widetilde{t} - \widetilde{\tau}_k - l\widetilde{T}) + z(\widetilde{t}), \qquad (46)$$

where $\text{rect}(\widetilde{t}) = \frac{1}{2}[\text{sgn}(\widetilde{t} + \widetilde{T}) - \text{sgn}(\widetilde{t})]$ is a rectangular pulse of duration $\widetilde{T}$, $\text{sgn}(\cdot)$ denotes the sign function, $\widetilde{\tau}_k$ denotes the time offset of the k-th device, and $z(\widetilde{t})$ is the noise. For simplicity, perfect channel gain pre-equalization is considered in (46). Without loss of generality, we assume $\widetilde{T} = 1$ and the time offset $\widetilde{\tau}_k$, $\forall k$, is less than the symbol duration $\widetilde{T}$. We fix a maximum time offset $\widetilde{\tau}$; then, we generate the time offset $\widetilde{\tau}_k$, $\forall k$, uniformly in the set (0,$\widetilde{\tau}$). Similar to [17], [19], [20], by applying whitened matched filtering and sampling scheme, we get oversampled, but independent, samples from the overlapped signals (46). Then, the *aligned-sample estimator* (refer to Definition 3 of [17]) is applied in our simulation to estimate $\widehat{\mathbf{y}}_d \triangleq \widehat{\sum_{k\in\mathcal{S}_a} \mathbf{P}\mathbf{x}_k^d}$. Then, $\widehat{\mathbf{y}}_d$ is used as input to conduct our proposed **Algorithm 2**.

In Fig. 9, we investigate the performance of the proposed "MD-AirComp (6, 20, 20), $M = 1$" scheme under various imperfections, where "$\phi = \pi/6$", "$\phi = \pi/10$", "$\widetilde{\tau} = 0.9$", and "$\widetilde{\tau} = 0, \phi = 0$" are considered. From Fig. 9(a), it can be seen that the performance still converges to the same level of test accuracy at negligible loss of convergence speed. From Fig. 9(b), it is shown that both increasing the time offset and increasing the phase offset result in an SNR penalty, leading to degradation in NMSE performance of the proposed AMP-DA algorithm. Fortunately, the degraded estimation accuracy of AMP-DA algorithm is still acceptable for the FEEL task. In addition, as has been concluded from Fig. 5(b), it is also possible to increase L and M to further improve the NMSE performance of the AMP-DA algorithm. To sum up, our proposed MD-AirComp scheme is robust to imperfect channel estimation or imperfect time synchronization.



(a)



(b)

Fig. 9. Performance of the proposed MD-AirComp scheme under various imperfections: (a) Test accuracy vs training rounds; (b) Detection NMSE vs training rounds.

VII. Conclusions

We have proposed a novel communication-efficient MD-AirComp scheme which can allow using over-the-air computation over legacy digital communication systems. The main idea of the proposed solution relies on the fact that, in AirComp, the receiver is only interested in the summation of the signals from multiple devices, but not the identities of the transmitting devices or their individual signals. The individual values to be summed are first quantized based on a common quantization codebook, then each quantized element is modulated into a transmit sequence selected from the common non-orthogonal modulation codebook. These transmit sequences from different devices overlap at the BS. Assuming that the number of transmitting devices is much less than the codebook size, the proposed sparsity promoting AMP-DA algorithm can efficiently compute the summation results by estimating which sequences have been transmitted, and by how many devices.

Furthermore, we applied the proposed technique for efficient model aggregation in FEEL, which is one of the most exciting parallel and distributed optimization applications. Here, the large-dimension model update is divided into blocks, and MD-AirComp is applied separately on each block. Our analysis proved a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ for the proposed MD-AirComp-based FEEL scheme, where T denotes the number of global training rounds. Simulation results verified that the proposed MD-AirComp scheme significantly outperforms the state-of-the-art schemes in test accuracy, and approaches the performance of ideal federated learning, despite using limited communication resources. Simulation results also verify the robustness of our proposed MD-AirComp scheme to imperfect channel estimation, imperfect time synchronization, and low SNR scenarios. There are various promising directions to explore, such as the development of adaptive quantization strategies (possibly layer-wise) for MD-AirComp, pursuing end-to-end optimization using deep learning, and investigating the applicability of MD-AirComp for privacy, among several other possibilities.

APPENDIX

A. Proof of Proposition 1

For simplicity, we omit the subscriptions n and m in the derivations. By substituting (17), (18), and (24) into (16), we obtain

$$\begin{aligned} & p\left(x|r\right) \\ &= \frac{1}{p\left(r\right)}\mathcal{CN}(r;x,\varphi)\left[(1-a)\delta\left(x\right)+a\mathcal{CN}\left(x;\mu_0,\tau_0\right)\right], \\ &= \frac{1}{p\left(r\right)}\Bigg[(1-a)\mathcal{CN}(0;r,\varphi)\delta\left(x\right)+ \\ &\qquad a\mathcal{CN}(0;r-\mu_0,\varphi+\tau_0)\mathcal{CN}(x;\frac{\mu_0\varphi+r\tau_0}{\varphi+\tau_0},\frac{\varphi\tau_0}{\varphi+\tau_0})\Bigg]. \end{aligned} \tag{47}$$

Then, substituting $p(r) = (1-a)\mathcal{CN}(0;r,\varphi) + a\mathcal{CN}(x;r-\mu_0,\varphi+\tau_0)$ into (47), we can obtain

$$p\left(x|r\right) = (1-\pi)\delta\left(x\right) + \pi\mathcal{CN}(x;\frac{\mu_0\varphi+r\tau_0}{\varphi+\tau_0},\frac{\varphi\tau_0}{\varphi+\tau_0}), \tag{48}$$

where $\pi = \dfrac{a}{a+(1-a)\frac{\mathcal{CN}(0;r,\varphi)}{\mathcal{CN}(0;r-\mu_0,\varphi+\tau_0)}}$. Hence, we can easily obtain (28)-(31). Hence, proposition 1 is proven.

B. Proof of Theorem 1

According to (2)-(4) and (11), we have $\mathbf{s}_k^t = \mathcal{C}(\Delta_k^t + \mathbf{e}_k^t)$, $\mathbf{e}_k^{t+1} = \Delta_k^t + \mathbf{e}_k^t - \mathbf{s}_k^t$, and $\widehat{\mathbf{w}}^{t+1} - \widehat{\mathbf{w}}^t = \eta\widehat{\mathbf{s}}^t$. To facilitate the proof, we introduce a new vector $\mathbf{z}^{t+1} \triangleq \widehat{\mathbf{w}}^{t+1} + \eta\mathbf{e}^{t+1}$, $\forall t \in [T]$, where $\mathbf{e}^t = \frac{1}{K}\sum_{k=1}^K \mathbf{e}_k^t$. We have $\mathbf{z}^{t+1} - \mathbf{z}^t = \eta(\Delta^t + \mathbf{m}^t)$, where $\Delta^t \triangleq \frac{1}{K}\sum_{k=1}^K \Delta_k^t$ and the aggregation error is $\mathbf{m}^t \triangleq \widehat{\mathbf{s}}^t - \mathbf{s}^t$. Note that the aggregation error $\mathbf{m}^t$ results from the wireless transmission in FEEL.

According to Assumption 1, we have

$$\begin{aligned} & \mathbb{E}[f(\mathbf{z}^{t+1})] - f(\mathbf{z}^t) \\ &\le \mathbb{E}[\langle\nabla f(\mathbf{z}^t), \mathbf{z}^{t+1} - \mathbf{z}^t\rangle] + \frac{\overline{L}}{2}\mathbb{E}[\|\mathbf{z}^{t+1} - \mathbf{z}^t\|^2] \\ &\le \mathbb{E}\left[\left\langle\nabla f(\mathbf{z}^t), \eta\Delta^t\right\rangle\right] + \frac{\eta^2\overline{L}}{2}\mathbb{E}\left[\left\|\Delta^t + \mathbf{m}^t\right\|^2\right] \\ &= \underbrace{\mathbb{E}\left[\left\langle\nabla f(\mathbf{w}^t), \eta\Delta^t\right\rangle\right]}_{R_1} + \underbrace{\frac{\eta^2\overline{L}}{2}\mathbb{E}\left[\left\|\Delta^t + \mathbf{m}^t\right\|^2\right]}_{R_2} \\ &+ \underbrace{\mathbb{E}\left[\left\langle\nabla f(\mathbf{z}^t) - \nabla f(\mathbf{w}^t), \eta\Delta^t\right\rangle\right]}_{R_3}. \end{aligned} \tag{49}$$

Bounding R_1: We have

$$\begin{aligned} R_1 &= \mathbb{E}\left[\left\langle\nabla f(\mathbf{w}^t), \eta\Delta^t\right\rangle\right] \\ &= \eta\mathbb{E}\left[\left\langle\nabla f(\mathbf{w}^t), \Delta^t + \eta_l T_l\nabla f(\mathbf{w}^t) - \eta_l T_l\nabla f(\mathbf{w}^t)\right\rangle\right] \\ &= -\eta\eta_l T_l\mathbb{E}\left[\left\|\nabla f(\mathbf{w}^t)\right\|^2\right] + \eta\mathbb{E}\left[\left\langle\nabla f(\mathbf{w}^t), \Delta^t + \eta_l T_l\nabla f(\mathbf{w}^t)\right\rangle\right] \\ &= -\eta\eta_l T_l\mathbb{E}\left[\left\|\nabla f(\mathbf{w}^t)\right\|^2\right] + \\ &\underbrace{\eta\left\langle\nabla f(\mathbf{w}^t), \mathbb{E}\left[-\frac{\eta_l}{K}\sum_{k=1}^K\sum_{t_l=0}^{T_l-1}\mathbf{g}_{k,t_l}^t + \frac{\eta_l T_l}{K}\sum_{k=1}^K\nabla F_k(\mathbf{w}^t)\right]\right\rangle}_{\widetilde{R_1}}. \end{aligned} \tag{50}$$

Note that $f(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^K F_k(\mathbf{w})$, and $\mathbf{g}_{k,t_l}^t = \nabla F_k(\mathbf{w}_{k,t_l}^t)$ is the local gradient. For the last term in (50), by applying $\langle\mathbf{x},\mathbf{y}\rangle = \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2)$ and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \widetilde{R_1} &= \eta\Bigg\langle\sqrt{\eta_l T_l}\nabla f(\mathbf{w}^t), \\ &- \frac{\sqrt{\eta_l T_l}}{T_l K}\mathbb{E}\left[\sum_{k=1}^K\sum_{t_l=0}^{T_l-1}(\nabla F_k(\mathbf{w}_{k,t_l}^t) - \nabla F_k(\mathbf{w}^t))\right]\Bigg\rangle \\ &\le \frac{\eta\eta_l T_l}{2}\left\|\nabla f(\mathbf{w}^t)\right\|^2 - \frac{\eta\eta_l}{2T_l K^2}\mathbb{E}\left[\left\|\sum_{k=1}^K\sum_{t_l=0}^{T_l-1}\nabla F_i(\mathbf{w}_{k,t_l}^t)\right\|^2\right] \\ &+ \frac{\eta\eta_l}{2K}\sum_{k=1}^K\sum_{t_l=0}^{T_l-1}\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{k,t_l}^t) - \nabla F_i(\mathbf{w}^t)\right\|^2\right]. \end{aligned} \tag{51}$$

Then, by applying Lemma 3 of [63] and using Assumption 1, under the constraint of local learning rate $\eta_l \le \frac{1}{8T_l\overline{L}}$, we have

$$\begin{aligned} & \frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\nabla F_i(\mathbf{w}_{k,t_l}^t) - \nabla F_i(\mathbf{w}^t)\right\|^2\right] \\ &\le \overline{L}^2\mathbb{E}\left[\left\|\mathbf{w}_{k,t_l}^t - \mathbf{w}^t\right\|^2\right] \\ &\le 5T_l\eta_l^2\overline{L}^2(\sigma_l^2 + 6T_l\sigma_g^2) + 30{T_l}^2\eta_l^2\overline{L}^2\mathbb{E}[\|\nabla f(\mathbf{w}^t)\|^2]. \end{aligned} \tag{52}$$

Hence, by merging (50)-(52) together, we obtain the bound

$$\begin{aligned} R_1 \le & -\frac{\eta\eta_l T_l}{4}\mathbb{E}\left[\left\|\nabla f(\mathbf{w}^t)\right\|^2\right] + \frac{5\eta\eta_l^3{T_l}^2\overline{L}^2}{2}(\sigma_l^2 + 6T_l\sigma_g^2) \\ & - \frac{\eta\eta_l}{2T_l K^2}\mathbb{E}\left[\left\|\sum_{k=1}^K\sum_{t_l=0}^{T_l-1}\nabla F_k(\mathbf{w}_{k,t_l}^t)\right\|^2\right]. \end{aligned} \tag{53}$$

Bounding R_2: It can be bounded as follows:

$$\begin{aligned} R_2 &= \frac{\eta^2\overline{L}}{2}\mathbb{E}\left[\left\|\Delta^t + \mathbf{m}^t\right\|^2\right] \\ &\le \eta^2\overline{L}\mathbb{E}\left[\left\|\Delta^t\right\|^2\right] + \eta^2\overline{L}\mathbb{E}\left[\left\|\mathbf{m}^t\right\|^2\right], \end{aligned} \tag{54}$$

where the first inequality follows from Cauchy-Schwarz inequality. Referring to Section III-B, we can obtain

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{m}^t\right\|^2\right] &= \mathbb{E}\left[\left\|\widehat{\mathbf{s}}^t - \mathbf{s}^t\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\sum_{d=1}^D\left(\frac{1}{\widehat{K}^t}\mathbf{U}^t\widehat{\mathbf{x}}_d^t - \frac{1}{K^t}\mathbf{U}^t\mathbf{x}_d^t\right)\right\|^2\right] \\ &\overset{(a)}{\le} \frac{\|\mathbf{U}^t\|^2}{(K^t)^2}\sum_{d=1}^D\left\|\widehat{\mathbf{x}}_d^t - \mathbf{x}_d^t\right\|^2 \overset{(b)}{=} \frac{\|\mathbf{U}^t\|^2 DN\sigma_n^2}{(K)^2}, \end{aligned} \tag{55}$$

where (a) and (b) hold by Assumption 5. Note that (b) also assumes $K^t = K$, i.e., the number of participating devices does not change along training rounds.

According to (5), for a sufficiently large VQ codebook, the average power of $\mathbf{U}^t$ will be the average power of $\overline{\mathbf{s}}_k^t$, $\forall k \in [K]$. Hence, we have

$$\frac{\|\mathbf{U}^t\|^2 DN\sigma_n^2}{(K^t)^2} \approx \frac{QN^2\|\overline{\mathbf{s}}_k^t\|^2 D\sigma_n^2}{W(K^t)^2} \le \frac{C_0\eta_l^2 T_l^2 G^2 N^2\sigma_n^2}{(K^t)^2}, \tag{56}$$

where the inequality holds by Assumptions 2 and 4, and $C_0 \triangleq \frac{2(1+q^2)^2}{(1-q^2)^2}$.

Bounding R_3:

$$\begin{aligned} R_3 &= \mathbb{E}\left[\left\langle \nabla f(\mathbf{z}^t) - \nabla f(\mathbf{w}^t), \eta\Delta^t \right\rangle\right] \\ &\le \mathbb{E}\left[\left\|\nabla f(\mathbf{z}^t) - \nabla f(\mathbf{w}^t)\right\| \left\|\eta\Delta^t\right\|\right] \\ &\le \overline{L}\mathbb{E}\left[\left\|\mathbf{z}^t - \mathbf{w}^t\right\| \left\|\eta\Delta^t\right\|\right] \\ &\le \frac{\eta^2\overline{L}}{2}\mathbb{E}\left[\left\|\Delta^t\right\|^2\right] + \frac{\eta^2\overline{L}}{2}\mathbb{E}\left[\left\|\mathbf{e}^t\right\|^2\right], \end{aligned} \tag{57}$$

where the first inequality follows from the fact that $\langle \mathbf{a}, \boldsymbol{b}\rangle \le \|\mathbf{a}\|\|\boldsymbol{b}\|$; the second one from Assumption 1; and the third from the definition of $\mathbf{z}^t$ and the fact that $\|\mathbf{a}\|\|\boldsymbol{b}\| \le \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$. Then, summing R_3 over $t = 1, \ldots, T$, we have

$$\sum_{t=1}^{T} R_3 \le \frac{\eta^2\overline{L}}{2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\Delta^t\right\|^2\right] + \frac{\eta^2\overline{L}}{2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\mathbf{e}^t\right\|^2\right]. \tag{58}$$

Then, by referring to Lemma C.7 of [64], the R_3 term is bounded by

$$\begin{aligned} \sum_{t=1}^{T} R_3 \le{}& \frac{\eta^2\overline{L}}{2}\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2] + C_1\eta^2\overline{L}\sigma_l^2\frac{TT_l\eta_l^2}{K} \\ &+ C_1\eta^2\overline{L}\frac{\eta_l^2}{K^2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\sum_{t_l=0}^{T_l-1}\nabla F_i(\mathbf{w}_{k,t_l}^t)\right\|^2\right], \end{aligned} \tag{59}$$

where $C_1 = \frac{2(q+\gamma)^2}{(1-q^2)^2}$.

Merging pieces together: Substituting (53) and (54) into (49), summing over from $t = 1$ to T, adding (59), and then applying Lemma C.5 of [64], we have

$$\begin{aligned} &\mathbb{E}[f(\mathbf{z}^{T+1})] - f(\mathbf{z}^1) = \sum_{t=1}^{T}[R_1 + R_2 + R_3] \\ &\le -\frac{\eta\eta_l T_l}{4}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{w}^t)\right\|^2\right] + \frac{5\eta\eta_l^3 {T_l}^2\overline{L}^2 T}{2}(\sigma_l^2 + 6T_l\sigma_g^2) \\ &\quad + (\frac{3}{2} + C_1)\eta^2\overline{L}\sigma_l^2\frac{TT_l\eta_l^2}{K} + \eta^2\overline{L}\sum_{t=1}^{T}\frac{\|\mathbf{U}^t\|^2 DN\sigma_n^2}{(K)^2} + \\ &\left(\frac{3\eta^2\overline{L}\eta_l^2}{2K^2} - \frac{\eta\eta_l}{2T_lK^2} + \frac{C_1\eta^2\overline{L}\eta_l^2}{K^2}\right)\sum_{t=1}^{T}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\sum_{t_l=0}^{T_l-1}\nabla F_i(\mathbf{w}_{k,t_l}^t)\right\|^2\right] \\ &\overset{(c)}{\le} -\frac{\eta\eta_l T_l}{4}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{w}^t)\right\|^2\right] + \frac{5\eta\eta_l^3 {T_l}^2\overline{L}^2 T}{2}(\sigma_l^2 + 6T_l\sigma_g^2) \\ &+ \left(\frac{3}{2} + C_1\right)\eta^2\overline{L}\sigma_l^2\frac{TT_l\eta_l^2}{K} + \frac{T\eta^2\overline{L}C_0\eta_l^2 T_l^2 G^2N^2\sigma_n^2}{K^2}, \end{aligned} \tag{60}$$

where the inequality (c) holds since $\eta_l \le \frac{1}{\eta LT_l(3+2C_1)}$. Hence, we have

$$\begin{aligned} &\frac{\eta\eta_l T_l}{4T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{w}^t)\|^2] \\ &\le \frac{f(\mathbf{z}^0) - \mathbb{E}[f(\mathbf{z}^T)]}{T} + \frac{5\eta\eta_l^3 {T_l}^2\overline{L}^2}{2}(\sigma_l^2 + 6T_l\sigma_g^2) \\ &+ (\frac{3}{2} + C_1)\eta^2\overline{L}\sigma_l^2\frac{T_l\eta_l^2}{K} + \frac{\eta^2\eta_l^2\overline{L}C_0 T_l^2 G^2N^2\sigma_n^2}{K^2}, \end{aligned} \tag{61}$$

which implies

$$\min_{t\in[T]}\mathbb{E}[\|\nabla f(\mathbf{w}^t)\|^2] \le 4\left[\frac{f_0 - f_*}{\eta\eta_l T_l T} + \Xi + \Psi\right], \tag{62}$$

where $\Xi \triangleq (\frac{3}{2} + C_1)\eta\overline{L}\sigma_l^2\frac{\eta_l}{K} + \frac{5\eta_l^2 T_l\overline{L}^2}{2}(\sigma_l^2 + 6T_l\sigma_g^2)$, $\Psi \triangleq \frac{\eta\eta_l\overline{L}C_0T_lG^2N^2\sigma_n^2}{K^2}$.

REFERENCES

[1] L. Qiao, Z. Gao, Z. Li, and D. Gündüz, "Unsourced massive access-based digital over-the-air computation for efficient federated edge learning," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Taipei, Taiwan, 2023, pp. 2003-2008.

[2] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Select. Areas Commun.,* vol. 39, no. 3, pp. 615-637, Mar. 2021.

[3] S.-C. Lin, T.-H. Chang, E. A. Jorswieck, and P.-H. Lin, *Information theory, mathematical optimization, and their crossroads in 6G system design*. 1st ed. Springer Series in Wireless Technology, 2023.

[4] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gunduz, and H. V. Poor, "Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication," *IEEE Wirel. Commun.*, vol. 30, no. 6, pp. 127-135, Dec. 2023.

[5] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated learning for Internet of Things: A comprehensive survey," *IEEE Commun. Surv. Tut.*, vol. 23, no. 3, pp. 1622-1658, thirdquarter 2021.

[6] Y.-F. Liu *et al.*, "A survey of advances in optimization methods for wireless communication system design." to appear in *IEEE J. Sel. Areas Commun.*, arXiv preprint arXiv:2401.12025, 2024.

[7] G. Scutari, F. Facchinei, L. Lampariello, S. Sardellitti, and P. Song, "Parallel and distributed methods for constrained nonconvex optimization-Part II: Applications in communications and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1945-1960, Apr. 2017.

[8] D. Gündüz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to learn at the edge," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 14-19, Dec. 2020.

[9] M. B. Mashhadi, M. Mahdavimoghadam, R. Tafazolli, and W. Saad, "Collaborative learning with a drone orchestrator," *IEEE Trans. Veh. Technol.*, doi: 10.1109/TVT.2023.3303630.

[10] 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.

[11] 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.

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

[13] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Time-correlated sparsification for efficient over-the-air model aggregation in wireless federated learning," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Seoul, Republic of Korea, 2022, pp. 3388-3393.

[14] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation", *IEEE Trans. Wireless Commun.,* vol. 19, no. 3, pp. 2022-2035, Mar. 2020.

[15] Y. -S. Jeon, M. M. Amiri, J. Li, and H. V. Poor, "A compressive sensing approach for federated learning over massive MIMO communication systems," *IEEE Trans. Wireless Commun.,* vol. 20, no. 3, pp. 1990-2004, Mar. 2021.

[16] Y. Shao, D. Gündüz, and S. C. Liew, "Bayesian over-the-air computation," *IEEE J. Select. Areas Commun.,* vol. 41, no. 3, pp. 589-606, Mar. 2023.

[17] Y. Shao, D. Gündüz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," *IEEE Trans. Wireless Commun.,* vol. 21, no. 6, pp. 3951-3964, Jun. 2022.

[18] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Sel. Areas Commun.,* vol. 40, no. 1, pp. 227-242, Jan. 2022.

[19] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.,* vol. 39, no. 12, pp. 3579-3605, Dec. 2021.

[20] A. Şahin and R. Yang, "A survey on over-the-air computation," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 3, pp. 1877-1908, thirdquarter 2023.

[21] A. S. Berahas, R. Bollapragada, N. S. Keskar, and E. Wei, "Balancing communication and computation in distributed optimization," *IEEE Trans. Automat. Contr.,* vol. 64, no. 8, pp. 3141-3155, Aug. 2019.

[22] S. F. Yilmaz, B. Hasırcıoğlu, and D. Gündüz, "Over-the-air ensemble inference with model privacy", in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2022.

[23] Z. Liu, Q. Lan, A. E. Kalør, P. Popovski, and K. Huang, "Over-the-air multi-view pooling for distributed sensing". ArXiv preprint arXiv:2302.09771.

[24] 3GPP, "5G NR overall description Stage-2," 3rd Generation Partnership Project (3GPP), TS 38.300 V15.3.1, Oct. 2018.

[25] A. Reisizadeh *et al.*, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," *Int. Conf. Artif. Intell. Stat.*, 2020, pp. 2021–2031.

[26] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient federated learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, 2021.

[27] L. Qu, S. Song, and C. -Y. Tsui, "FedDQ: Communication-efficient federated learning with descending quantization," in *Proc. Global Commun. Conf. (Globecom)*, Rio de Janeiro, Brazil, 2022, pp. 281-286.

[28] Y. Wang, Y. Xu, Q. Shi, and T.-H. Chang, "Quantized federated learning under transmission delay and outage constraints," *IEEE J. Select. Areas Commun.,* vol. 40, no. 1, pp. 323-341, Jan. 2022.

[29] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500-514, 2021.

[30] Y. Oh, Y. -S. Jeon, M. Chen and W. Saad, "FedVQCS: Federated learning via vector quantized compressed sensing," *IEEE Trans. Wireless Commun.,* doi: 10.1109/TWC.2023.3291877.

[31] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.,* vol. 20, no. 3, pp. 2120-2135, Mar. 2021.

[32] A. Şahin, "Distributed learning over a wireless network with non-coherent majority vote computation," *IEEE Trans. Wireless Commun.,*, vol. 22, no. 11, pp. 8020-8034, Nov. 2023.

[33] A. Şahin, "Over-the-air computation based on balanced number systems for federated edge learning," *IEEE Trans. Wireless Commun.,* doi: 10.1109/TWC.2023.3320116.

[34] S. Razavikia, J. M. B. Da Silva, and C. Fischione, "ChannelComp: A general method for computation by communications," *IEEE Trans. Commun.,* doi: 10.1109/TCOMM.2023.3324999.

[35] ——, "SumComp coding: Digital over-the-air computation via the ring of integers," *arXiv preprint arXiv:2310.20504*, 2023.

[36] ——, "Blind federated learning via over-the-air q-QAM," *arXiv preprint arXiv:2311.04253*, 2023.

[37] Z. Gao *et al.*, "Compressive sensing-based grant-free massive access for 6G massive communication," *IEEE Internet Things J.*, doi: 10.1109/JIOT.2023.3334878.

[38] Z. Chen, F. Sohrabi, Y.-F. Liu, and W. Yu, "Phase transition analysis for covariance based massive random access with massive MIMO," *IEEE Trans. Inf. Theory,* vol. 68, no. 3, pp. 1696-1715, Mar. 2022.

[39] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933-2946, Jun. 2018.

[40] Y. Mei *et al.*, "Compressive sensing-based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.,* vol. 21, no. 3, pp. 1851-1869, Mar. 2022.

[41] Y. Mei *et al.*, "Massive access in extra large-scale MIMO With mixed-ADC over near-field channels," *IEEE Trans. Veh. Technol.,* vol. 72, no. 9, pp. 12373-12378, Sept. 2023.

[42] L. Qiao *et al.*, "Joint activity and blind information detection for UAV-assisted massive IoT access," *IEEE J. Select. Areas Commun.,* vol. 40, no. 5, pp. 1489-1508, May 2022.

[43] K. Ying, Z. Gao, S. Chen, M. Zhou, D. Zheng, S. Chatzinotas, B. Ottersten, and H. V. Poor, "Quasi-synchronous random access for massive MIMO-based LEO satellite constellations," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1702-1722, Jun. 2023.

[44] Z. Wang, Y.-F. Liu, and L. Liu, "Covariance-based joint device activity and delay detection in asynchronous mMTC," *IEEE Signal Process. Lett.,* vol. 29, pp. 538–542, Jan. 2022.

[45] L. Qiao *et al.*, "Sensing user's activity, channel, and location with near-field extra-large-scale MIMO," *IEEE Trans. Commun.*, vol. 72, no. 2, pp. 890-906, Feb. 2024.

[46] Y. Polyanskiy, "A perspective on massive random-access", in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pp. 1-5, Jun. 2017.

[47] T. Li *et al.*, "Joint device detection, channel estimation, and data decoding with collision resolution for MIMO massive unsourced random access," *IEEE J. Select. Areas Commun.,* vol. 40, no. 5, pp. 1535-1555, May 2022.

[48] K. -H. Ngo, G. Durisi, A. G. i. Amat, P. Popovski, A. E. Kalør, and B. Soret, "Unsourced multiple access with common alarm messages: Network slicing for massive and critical IoT," *IEEE Trans. Wireless Commun.,*, vol. 72, no. 2, pp. 907-923, Feb. 2024.

[49] X. Shao, X. Chen, D. W. K. Ng, C. Zhong, and Z. Zhang, "Cooperative activity detection: Sourced and unsourced massive random access paradigms," *IEEE Trans. Signal Process.,* vol. 68, pp. 6578-6593, 2020.

[50] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Stat.*, 2017, pp. 1273-1282.

[51] J. Hamer, M. Mohri, and A. T. Suresh, "Fedboost: A communication-efficient algorithm for federated learning", in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 3973-3983.

[52] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, 2020, pp. 429-450.

[53] J. Wang, Q. Liu, H. Liang, G. Joshi, and V. H. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 7611-7623.

[54] S. U. Stich, J. B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 4452-4463, Dec. 2018.

[55] R. J. Baxley, B. T. Walkenhorst, and G. Acosta-Marum, "Complex gaussian ratio distribution with applications for error rate calculation in fading channels with imperfect CSI," in *Proc. Global Commun. Conf. (Globecom)*, Miami, FL, USA, 2010, pp. 1-5.

[56] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, vol. 1. Englwood Cliffs, NJ, USA: Prentice-Hall, 1993.

[57] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914-18919, 2009.

[58] X. Meng, S. Wu, L. Kuang, and J. Lu, "An expectation propagation perspective on approximate message passing," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1194-1197, Aug. 2015.

[59] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, threshold achieving matrices," *J. Stat. Mech., Theor. Exp.*, vol. 2012, no. 8, p. P08009, Aug. 2012.

[60] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[61] L. Liu, S. Huang, and B. M. Kurkoski, "Memory AMP," *IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 8015-8039, Dec. 2022.

[62] Q. Hu, Y. Cai, G. Zhang *et al.*, "Deep-unfolding for next-generation transceivers," *arXiv preprint arXiv:2305.08303*, 2023.

[63] S. Reddi *et al.*, "Adaptive federated optimization," arXiv preprint arXiv:2003.00295, 2020.

[64] Y. Wang, L. Lin, and J. Chen, "Communication-efficient adaptive federated learning", in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 22802-22838.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[66] Y. Zhao *et al.*, "Federated learning with non-iid data", arXiv preprint arXiv:1806.00582, 2018.

[67] D. Arthur and S. Vassilvitskii, "K-means++ the advantages of careful seeding," *Proc. 18th Ann. ACM-SIAM Symp. Discrete Algorithms*, pp. 1027-1035, 2007.

[68] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, arXiv:1608.03983.



Li Qiao (Graduate Student Member, IEEE) received his Bachelor of Science degree at the Beijing Institute of Technology in 2019. He is currently pursuing a Ph.D. at the School of Information and Electronics, Beijing Institute of Technology, China. He is also a collaborative Ph.D. student with the 5G/6G Innovation Centre (5G/6GIC) at the Institute for Communication Systems (ICS), University of Surrey, UK. Previously, he was a visiting student at the Information Processing and Communications Lab (IPC Lab) in the Department of Electrical and Electronic Engineering, Imperial College London, UK. His research interests include wireless communications, with a focus on massive communication, federated edge intelligence, and generative semantic communications.



Zhen Gao (Member, IEEE) received the B.S. degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in communication and signal processing from the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China, in 2016. He is currently an Associate Professor with the Beijing Institute of Technology. His research interests include wireless communications, with a focus on multi-carrier modulations, multiple antenna systems, and sparse signal processing. He was a recipient of the IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (Best Paper), an Exemplary Reviewer of IEEE COMMUNICATION LETTERS in 2016, IET Electronics Letters Premium Award (Best Paper) in 2016, and the Young Elite Scientists Sponsorship Program (from 2018 to 2021) from the China Association for Science and Technology.



Mahdi Boloursaz Mashhadi [S'14-M'18-SM'23] is a Lecturer at the 5G/6G Innovation Centre (5G/6GIC) at the Institute for Communication Systems (ICS), University of Surrey (UoS), UK. Prior to joining ICS, he was a postdoctoral research associate at the Intelligent Systems and Networks (ISN) Research Group, Imperial College London, 2019-2021. He received B.S., M.S., and Ph.D. degrees in mobile telecommunications from the Sharif University of Technology (SUT), Tehran, Iran, in 2011, 2013, and 2018, respectively. He was a visiting research associate with the University of Central Florida, Orlando, USA, in 2018, and Queen's University, Ontario, Canada, in 2017. He has more than 40 peer-reviewed publications and patents in the areas of wireless communications, machine learning, and signal processing. He received the Best Paper Award from the IEEE EWDTS 2012 conference, and the Exemplary Reviewer Award from the IEEE ComSoc in 2021 and 2022. He has served as a panel judge for the International Telecommunication Union (ITU) to evaluate innovative submissions on applications of AI/ML in 5G and beyond wireless networks since 2021. He is an associate editor for the Springer Nature Wireless Personal Communications Journal.



Deniz Gündüz [S'03-M'08-SM'13-F'22] received the B.S. degree in electrical and electronics engineering from METU, Turkey in 2002, and the M.S. and Ph.D. degrees in electrical engineering from NYU Tandon School of Engineering (formerly Polytechnic University) in 2004 and 2007, respectively. Currently, he is a Professor of Information Processing in the Electrical and Electronic Engineering Department at Imperial College London, UK, where he also serves as the deputy head of the Intelligent Systems and Networks Group. In the past, he held various positions at the University of Modena and Reggio Emilia (part-time faculty member, 2019-22), University of Padova (visiting professor, 2018, 2020), Centre Tecnologic de Telecomunicacions de Catalunya (CTTC) (research associate, 2009-12), Princeton University (postdoctoral researcher, 2007-09, visiting researcher, 2009-11) and Stanford University (research assistant professor, 2007-09). His research interests lie in the areas of communications and information theory, machine learning, and privacy.

Dr. Gündüz is a Fellow of the IEEE. He is an elected member of the IEEE Signal Processing Society Signal Processing for Communications and Networking (SPCOM) and Machine Learning for Signal Processing (MLSP) Technical Committees. He serves as an Area Editor for the IEEE Transactions on Information Theory and IEEE Transactions on Communications. He is the recipient of the IEEE Communications Society - Communication Theory Technical Committee (CTTC) Early Achievement Award in 2017, Starting (2016) and Consolidator (2022) and Proof-of-Concept (2023) Grants of the European Research Council (ERC), and has co-authored several award-winning papers, including the IEEE Communications Society - Young Author Best Paper Award (2022), and IEEE International Conference on Communications Best Paper Award (2023). He received the Imperial College London - President's Award for Excellence in Research Supervision in 2023.