One-Bit Over-the-Air Aggregation for Communication-Efficient Federated Edge Learning: Design and Convergence Analysis

Guangxu Zhu, Yuqing Du, Deniz Gündüz, and Kaibin Huang

***Abstract*—*Federated edge learning* (FEEL) is a popular framework for model training at an edge server using data distributed at edge devices (e.g., smart-phones and sensors) without compromising their privacy. In the FEEL framework, edge devices periodically transmit high-dimensional stochastic gradients to the edge server, where these gradients are aggregated and used to update a global model. When the edge devices share the same communication medium, the multiple access channel (MAC) from the devices to the edge server induces a communication bottleneck. To overcome this bottleneck, an efficient broadband analog transmission scheme has been recently proposed, featuring the aggregation of analog modulated gradients (or local models) via the waveform-superposition property of the wireless medium. However, the assumed linear analog modulation makes it difficult to deploy this technique in modern wireless systems that exclusively use digital modulation. To address this issue, we propose in this work a novel digital version of broadband over-the-air aggregation, called *one-bit broadband digital aggregation* (OBDA). The new scheme features one-bit gradient quantization followed by digital *quadrature amplitude modulation* (QAM) at edge devices and over-the-air majority-voting based decoding at edge server. We provide a comprehensive analysis of the effects of wireless channel hostilities (channel noise, fading, and channel estimation errors) on the convergence rate of the proposed FEEL scheme. The analysis shows that the hostilities slow down the convergence of the learning process by introducing a scaling factor and a bias term into the gradient norm. However, we show that all the negative effects vanish as the number of participating devices grows, but at a different rate for each type of channel hostility.**

***Index Terms*—Over-the-air computation, Federated learning, Multiple access channels, Quantization, Digital modulation**

I. INTRODUCTION

Edge learning refers to the deployment of learning algorithms at the network edge so as to have rapid access to massive mobile data for training *artificial intelligence* (AI) models [1]–[3]. A popular framework, called *federated edge learning* (FEEL), distributes the task of model training over many edge devices [4], [5]. Thereby, FEEL exploits distributed data without compromising their privacy, and furthermore leverages the computation resources of edge devices. Essentially, the framework is a distributed implementation of *stochastic gradient descent* (SGD) in a wireless network. In FEEL, edge devices periodically upload locally trained models to an edge server, which are then aggregated and used to update a global model. When the edge devices participating in the learning process share the same wireless medium to convey local updates to the edge server, the limited radio resources can cause severe congestion over the air interface, resulting in a communication bottleneck for FEEL. One promising solution is *over-the-air aggregation* also called *over-the-air computation* (AirComp) that exploits the waveform superposition property of the wireless medium to support simultaneous transmission by all the devices [6], [7]. Compared with conventional orthogonal access schemes, this can dramatically reduce the required resources, particularly many devices participate in FEEL. However, the uncoded linear analog modulation scheme used for over-the-air aggregation in [6], [7] may cause difficulty in their deployment in existing systems, which typically use digital modulation (e.g., 3GPP). In this work, we propose a new design, called *one-bit broadband digital aggregation* (OBDA), that extends the conventional analog design by featuring digital modulation of gradients to facilitate efficient implementation of FEEL in a practical wireless system. We present a comprehensive convergence analysis of OBDA to quantify the effects of channel hostilities, including channel noise, fading and estimation errors, on its convergence rate. The results provide guidelines on coping with the channel hostilities in FEEL systems with OBDA.

A. FEEL over a Multiple Access Channel (MAC)

A typical FEEL algorithm involves iterations between two steps as shown in Fig. 1. In the first step, the edge server broadcasts the current version of the global model to the participating edge devices. Each edge device employs SGD using only locally available data. In the next step, edge devices convey their local updates (gradient estimates or model updates) to the edge server. Each iteration of these two steps is called one *communication round*. The iteration continues until the global model converges.

The communication bottleneck has already been acknowledged as a major challenge in the federated learning literature,

G. Zhu is with the Shenzhen Research Institute of Big Data, Shenzhen, China (Email: gxzhu@sribd.cn). Y. Du and K. Huang are with the Dept. of Electrical and Electronic Engineering at The University of Hong Kong, Hong Kong (Email: {yqdu, huangkb}@eee.hku.hk). Deniz Gündüz is with the Dept. of Electrical and Electronics Engineering at the Imperial College of London, London, UK (Email: d.gunduz@imperial.ac.uk). Corresponding author: K. Huang. Part of this paper has been presented at the IEEE Global Communications Conference (GLOBECOM), Taipei, Taiwan, Dec. 2020.

The work of G. Zhu was supported in part by National Key R&D Program of China under grant 2018YFB1800800, in part by National Natural Science Foundation of China under grant 62001310, in part by Guangdong Province Key Area R&D Program under grant 2018B030338001, and in part by Shenzhen Peacock Plan under grant KQTD2015033114415450. The work of K. Huang is supported by the Hong Kong Research Grants Council under the Grants 17208319 and 17209917, Innovation and Technology Fund under Grant GHP/016/18GD, and Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515130003. The work of D. Gündüz received funding from the European Research Council (ERC) under Starting Grant BEACON (agreement no. 677854).

and several strategies have been proposed to reduce the communication overhead. We can identify three main approaches. The first is to discard the updates from slow-responding edge devices (stragglers) for fast update synchronization [8], [9]. Another approach is to employ update significance rather than the computation speed to schedule devices [10], [11]. Update significance is measured by either the model variance [10], or the gradient divergence [11] corresponding to model-averaging [4] and gradient-averaging [5] implementation methods, respectively. The last approach focuses on update compression by exploiting the sparsity of gradient updates [12], [13] and low-resolution gradient/model-parameter quantization [14]–[16]. However, all these approaches assume reliable links between the devices and the server and ignore the wireless nature of the communication medium.

The envisioned implementation of FEEL in practical wireless networks requires taking into account wireless channel hostilities and the scarcity of radio resources. The first works in the literature that studied FEEL taking into account the physical layer resource constraints focus on over-the-air aggregation [6], [7], [17]–[19]. Specifically, a broadband over-the-air aggregation system based on analog modulation called *broadband analog aggregation* (BAA) is designed in [6], where the gradients/models transmitted by devices are averaged over frequency sub-channels. For such a system, several communication-learning tradeoffs are derived to guide the designs of broadband power control and device scheduling. Over-the-air aggregation is also designed for narrow-band channels with an additional feature of gradient dimension reduction exploiting gradient sparsity in [7], [17]. Subsequently, *channel-state information* (CSI) requirement for over-the-air aggregation is relaxed by exploiting multiple antennas at the edge server in [18]. The joint design of device scheduling and beamforming for over-the-air aggregation is investigated in [19] to accelerate FEEL in a multi-antenna system. Few other recent works focus on radio resource management [20]–[22]. In [20], a hierarchical FEEL framework is introduced in a cellular network. In [21], a novel bandwidth allocation strategy is proposed for minimizing the model training loss of FEEL via convergence analysis accounting for packet transmission errors. Different classic scheduling schemes, such as proportional fair scheduling, are applied in a FEEL system and their effects on the convergence rate are studied in [22].

The performance of FEEL is typically measured in terms of the convergence rate, which quantifies how fast the global model converges over communication rounds. The current work is the first to present a comprehensive framework for convergence analysis targeting FEEL with AirComp.

B. Over-the-Air Aggregation

With over-the-air aggregation, the edge server receives an approximate (noisy version) of the desired functional value, efficiently exploiting the available bandwidth by simultaneous transmission from all the devices, as opposed to orthogonalized massive access. The idea of over-the-air aggregation has previously been studied in the context of data aggregation in sensor networks, known also as AirComp. In [23], function computation over a MAC is studied from a fundamental information theoretic point of view, assuming *independent and identically distributed* (i.i.d.) source sequences, and focusing on the asymptotic computation rate. This is extended in [24] to wireless MACs and a more general class of non-linear functions. A practical implementation is presented in [25]. From an implementation perspective, techniques for distributed power control and robust design against channel estimation errors are proposed in [26] and [27], respectively. More recently, AirComp techniques are designed for *multiple-input-multiple-output* (MIMO) systems for enabling spatially multiplexed vector-valued function computation in [28]–[30]. To this end, receive beamforming and an enabling scheme for CSI feedback are designed in [28]. The framework is extended to wirelessly-powered AirComp [29] and massive MIMO AirComp systems [30].

Existing works on over-the-air aggregation, as well as its application in the context of FEEL [6], [7], [17]–[19] consider analog modulation assuming that the transmitter can modulate the carrier waveform as desired, freely choosing the I/Q coefficients as arbitrary real number. However, most existing wireless devices come with embedded digital modulation chips, and they may not be capable of employing an arbitrary modulation scheme. In particular, modern cellular systems are based on *orthogonal frequency division multiple access* (OFDMA) using *quadrature amplitude modulation* (QAM). Therefore, the goal of the paper is to extend the over-the-air aggregation framework to FEEL considering transmitters that are limited to QAM.

C. Contributions and Organization

In this paper, we consider the implementation of over-the-air aggregation for FEEL over a practical wireless system with digital modulation. Building on the signSGD proposed in [15] and *orthogonal frequency division multiplexing* (OFDM) transceivers, we design an elaborate FEEL scheme, called OBDA, which features one-bit gradient quantization and QAM modulation at devices, and over-the-air majority-vote based gradient-decoding at the edge server. This novel design will allow implementing AirComp across devices that are endowed with digital modulation chips, without requiring significant changes in the hardware or the communication architecture.

Moreover, while existing works on over-the-air analog aggregation mostly rely on numerical experiments for performance evaluation [6], [7], one of the main contributions of this work is an analytical study of the convergence rate of the OBDA scheme. The considered (model) convergence rate is defined as the rate at which the expected value of average gradient norm, denoted by $\bar{G}$, diminishes as the number of rounds, denoted by N, and the number of devices, denoted by K, grow. For ideal FEEL with single-bit gradient quantization transmitted over noiseless channels, the convergence rate is shown to be [15]

$$\bar{G} \leq \frac{1}{\sqrt{N}} \left(c + \frac{c'}{\sqrt{K}} \right), \tag{1}$$

where c and c' denote some constants related to the landscape of the training loss function.

The convergence rates of OBDA in the presence of channel hostilities are derived for three scenarios: 1) Gaussian MAC, 2) fading MAC with perfect (transmit) CSI, and 3) fading MAC with imperfect (transmit) CSI. For all three scenarios, the derived convergence rates share the following form:

$$\bar{G} \leq \frac{a}{\sqrt{N}} \left(c + \frac{c'}{\sqrt{K}} + b \right), \tag{2}$$

where the channel hostilities are translated into a scaling factor $a \in (1, \infty)$ and a bias term $b \in (0, \infty)$. It is clear that the wireless channel imperfections slow down the model convergence with respect to the noiseless counterpart in (1). The two terms a and b are independent of N, but are functions of K and the received *signal-to-noise ratio* (SNR) as shown explicitly in the sequel. Particularly, as K or the received SNR grows, a and b are shown to converge to their noiseless limits of 1 and 0, respectively. The convergence is shown to follow different scaling laws described as follows.

- **Gaussian MAC**: We consider a MAC with unit channel gain and additive complex Gaussian noise. For this case, the scaling factor a and the bias term b in (2) scale as $1 + O\left(\frac{1}{K\sqrt{\text{SNR}}}\right)$ and $O\left(\frac{1}{K\sqrt{\text{SNR}}}\right)$, respectively.
- **MAC with fading and perfect CSI**: We consider a fading MAC and OBDA with perfect CSI, which is needed for transmit power control. Let α denote the expected ratio of gradient coefficients truncated due to the adopted truncated channel inversion power control under a power constraint. Then, the terms a and b in (2) scale as $1 + O\left(\frac{1}{\alpha K\sqrt{\text{SNR}}}\right)$ and $O\left(\frac{1}{\alpha K\sqrt{\text{SNR}}}\right)$, respectively. In other words, the variation in channel quality due to fading is translated into an effective reduction on the number of devices by a factor of α in each round. This leads to a decreased convergence rate compared to the case of Gaussian channels.
- **MAC with fading and imperfect CSI**: In practice, CSI errors may exist due to inaccurate channel estimation. Using imperfect CSI for power control results in additional perturbation to the received gradients. As a result, the terms a and b in (2) scale as $1 + O\left(\frac{1}{\alpha\sqrt{K}}\right)$ and $O\left(\frac{1}{\alpha\sqrt{K}}\right)$, respectively, and are asymptotically independent of the receive SNR. Compared with the preceding two cases, the much slower rates and their insensitivity to increasing SNR shows that CSI errors can incur severe degradation of the OBDA performance.

Finally, it is worth emphasizing that the current work differs from the closely related work [15], [16] in that our focus is on designing customized wireless communication techniques for implementing FEEL over a wireless network rather than developing new federated learning algorithms. To this end, we particularly take into account the characteristics of wireless channels in the design of communication techniques for FEEL, and characterize their effects on the resultant convergence rate. The key findings are summarized above, which are originally presented in this work and represent the novelty *with respect to* (w.r.t.) [15], [16].

Organization: The remainder of the paper is organized as follows. Section II introduces the learning and communication models. Section III presents the proposed OBDA scheme. The convergence analysis under the AWGN channel case is presented in Section IV, and further extended to the fading channel case in Section V accounting for both the perfect CSI and imperfect CSI scenarios. Section VI presents the experimental results using real datasets followed by concluding remarks in Section VII.

II. Learning and Communication Models

We consider a FEEL system comprising a single edge server coordinating the learning process across K edge devices as shown in Fig. 1. Device k, $k = 1, \ldots, K$, has its own local dataset, denoted by $\mathcal{D}_k$, consisting of labeled data samples $\{(\mathbf{x}_i, y_j)\} \in \mathcal{D}_k$, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the unlabeled data and $y_j \in \mathbb{R}$ the associated label. A common model (e.g., a classifier), represented by the parameter vector $\mathbf{w} \in \mathbb{R}^q$, is trained collaboratively across the edge devices, orchestrated by the edge server. Here q denotes the model size.

A. Learning Model

The loss function measuring the model error is defined as follows. The *local loss function* of the model $\mathbf{w}$ on $\mathcal{D}_k$ is

$$F_k(\mathbf{w}) = \frac{1}{|\mathcal{D}_k|} \sum_{(\mathbf{x}_j, y_j) \in \mathcal{D}_k} f(\mathbf{w}, \mathbf{x}_j, y_j), \tag{3}$$

where $f(\mathbf{w}, \mathbf{x}_j, y_j)$ denotes the sample loss quantifying the prediction error of the model $\mathbf{w}$ on the training sample $\mathbf{x}_j$ w.r.t. its true label y_j. For convenience, we rewrite $f(\mathbf{w}, \mathbf{x}_j, y_j)$ as $f_j(\mathbf{w})$, and assume uniform sizes for local datasets; that is, $|\mathcal{D}_k| = D$, $\forall k$. Then, the *global loss function* on all the distributed datasets can be written as

$$F(\mathbf{w}) \triangleq \frac{\sum_{k=1}^{K} \sum_{j \in \mathcal{D}_k} f_j(\mathbf{w})}{\sum_{k=1}^{K} |\mathcal{D}_k|} = \frac{1}{K} \sum_{k=1}^{K} F_k(\mathbf{w}). \tag{4}$$

The goal of the learning process is thus to minimize the global loss function $F(\mathbf{w})$.[1] One way to do this is to upload all the local datasets to the edge server, and solve the problem in a centralized manner. However, this is typically undesirable due to either privacy concerns, or the sheer size of the datasets. Alternatively, the FEEL framework can be employed to minimize $F(\mathbf{w})$ in a distributed manner. We focus on the gradient-averaging implementation of FEEL in the current work as illustrated in Fig. 1 with the detailed procedure eloborated in the sequel.

In each communication round of FEEL, say the n-th round, device k computes a local estimate of the gradient of the loss function in (4) using its local dataset $\mathcal{D}_k$ and the current parameter-vector $\mathbf{w}^{(n)}$. Let $\mathbf{g}_k^{(n)} \in \mathbb{R}^q$ denote the local gradient estimate at device k in the n-th round, where we

[1]The problem formulation follows the standard in the existing federated learning literature (see e.g., [3]-[22]). Note that the formulated problem for model training is a stochastic problem as we construct the loss function using a randomly sampled subset of data, and the minimization of the loss function is solved by *stochastic gradient decent* (SGD). The statistical distribution of the stochastic gradient estimate is assumed to satisfy Assumption 3, which is a widely-adopted assumption for tractable convergence analysis for non-convex loss functions. Note that the SGD approach can also handle online training scenarios where the dataset is collected sequentially in real-time.

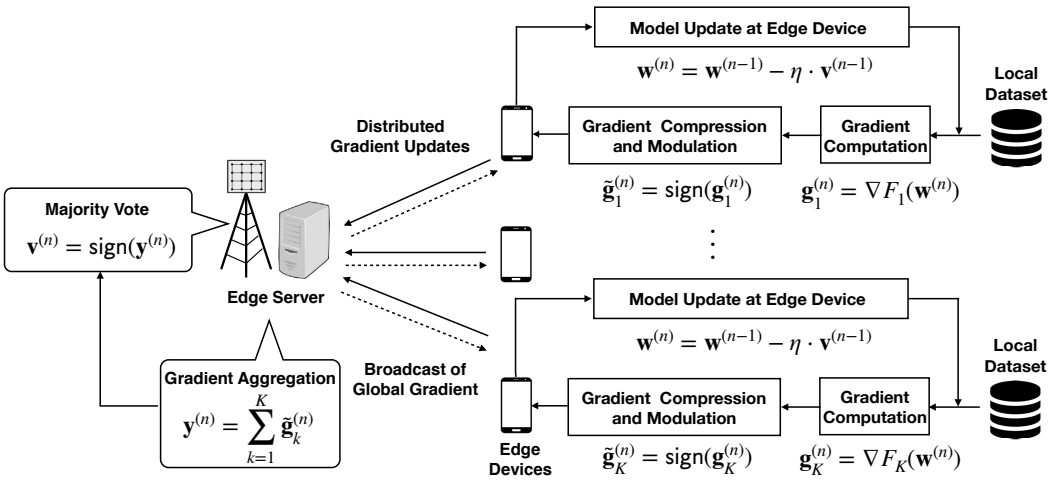


Figure 1. FEEL via OBDA from distributed data.

remove the dependence on the parameter vector $\mathbf{w}^{(n)}$ for simplicity. We then have:

$$\mathbf{g}_k^{(n)} = \frac{1}{n_b} \sum_{j \in \tilde{\mathcal{D}}_k} \nabla f_j(\mathbf{w}^{(n)}), \tag{5}$$

where ∇ represents the gradient operator. $\tilde{\mathcal{D}}_k \subset \mathcal{D}_k$ is the selected data batch from the local dataset for computing the local gradient estimate, and n_b is the batch size. Accordingly, $n_b = |\mathcal{D}_k|$ means that all the local dataset is used for gradient estimation. If the local gradient estimates can be reliably conveyed to the edge server, the global estimate of the gradient of the loss function in (4) would be computed as follows:[2]

$$\hat{\mathbf{g}}^{(n)} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{g}_k^{(n)}. \tag{6}$$

Then, the global gradient estimate is broadcast back to each device, which then uses it to update the current model via gradient descent based on the following equation:

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} - \eta \cdot \hat{\mathbf{g}}^{(n)}, \tag{7}$$

where η denotes the learning rate. The steps in (5), (6), and (7) are iterated until a convergence condition is met.

As observed from (6), it is only the aggregated gradient, i.e., $\sum_{k=1}^{K} \mathbf{g}_k$, not the individual gradient estimates $\{\mathbf{g}_k\}$, needed at the edge server for computing the global gradient estimate. This motivates the communication efficient aggregation scheme presented in Section III.

B. Communication Model

Local gradient estimates of edge devices are transmitted to the edge server over a broadband MAC. To cope with the frequency selective fading and inter-symbol interference, OFDM modulation is adopted to divide the available bandwidth B into M orthogonal sub-channels. We assume that a fixed digital constellation is employed by all the devices to transmit over each sub-channel. Thus, each device needs to transmit its local gradient estimate using a finite number of digital symbols. This requires quantization of the local gradient estimates, and mapping each quantized gradient element to one digital symbol to facilitate the proposed OBDA. Let $\tilde{\mathbf{g}}_k = [\tilde{g}_{k,1}, \ldots, \tilde{g}_{k,q}]^T$ denote the channel input vector of the k-th device, where q is the size of the gradient vector, and $\tilde{\mathbf{g}}_{k,j} \in \mathcal{Q}$ for some finite digital input constellation $\mathcal{Q}$.

During the gradient-uploading phase, all the devices transmit simultaneously over the whole available bandwidth. At each communication round, gradient-uploading duration consists of $N_s = \frac{q}{M}$ OFDM symbols for complete uploading of all the gradient parameters. We assume symbol-level synchronization among the transmitted devices through a synchronization channel (e.g., "timing advance" in LTE systems [31]).[3] Accordingly, the i-th aggregated gradient parameter, denoted by $\tilde{g}_i$, with $i = (t-1)M + m$, received at the m-th sub-carrier and t-th OFDM symbol is given by

$$\tilde{g}_i = \sum_{k=1}^{K} h_k[t,m] p_k[t,m] \tilde{g}_{k,i} + z[t,m], \qquad \forall i, \tag{8}$$

where $\{h_k[t,m]\}$ are the channel coefficients with identical Rayleigh distribution, i.e., $h_k[t,m] \sim \mathcal{CN}(0,1)$;[4] [5] and

[2]For the case with heterogenous dataset sizes at different devices, the global gradient estimate is a weighted average of the local ones, i.e., $\mathbf{g}^{(n)} = \frac{\sum_{k=1}^{K} |\mathcal{D}_k| \mathbf{g}_k^{(n)}}{\sum_{k=1}^{K} |\mathcal{D}_k|}$. The desired weighted aggregation of the local gradient estimate can also be attained by the proposed over-the-air aggregation with an additional pre-processing $\varphi_k(\cdot)$ on each of the transmitted signals, x_k, via $\varphi_k(x_k) = \frac{|\mathcal{D}_k|}{\sum_{k=1}^{K} |\mathcal{D}_k|} x_k$ similarly as in [19].

[3]The accuracy of synchronization is proportional to the bandwidth dedicated for the synchronization channel. Particularly, the current state-of-the-art phase-locked loop can achieve a synchronization offset of $0.1 B_s^{-1}$, where B_s is the amount of bandwidth used for synchronization. In existing LTE systems, the typical value of B_s is 1MHz. Thus, a sufficiently small synchronization offset of $0.1\mu sec$ can be achieved. Note that in a broadband OFDM system, as long as the synchronization offset is smaller than the length of cyclic prefix (the typical value is $5\mu sec$ in the LTE systems), the offset simply introduces a phase shift to the received symbol. The phase shift can be easily compensated by channel equalization, incurring no performance loss [32].

[4]It should be emphasized that the current analysis does not require the assumption of independent channel realizations over different time slots. In particular, the same analytical results hold as long as the channel coefficients at different time slots follow identical Rayleigh (complex Gaussian) distribution, regardless of whether there is correlation over time or not.

[5]We also assume identical path-losses for different devices to simplify exposition. Note that the difference in path-losses between devices can be equalized by the channel inversion power control applied in the design as specified in the sequel. If there exists a bottleneck device with severe path-loss that does not allow channel inversion within the given power budget, it can be excluded in the scheduling phase.

$p_k[t,m]$ is the associated power control policy to be specified in the sequel. Finally, $z[t,m]$ models the zero-mean i.i.d. *additive white Gaussian noise* (AWGN) with variance σ_z^2. For ease of notation, we skip the index of OFDM symbol t in the subsequent exposition whenever no confusion is incurred.

The power allocation over sub-channels, $\{p_k[m]\}$, will be adapted to the corresponding channel coefficients, $\{h_k[m]\}$, for implementing gradient aggregation via AirComp as presented in the sequel. The transmission of each device is subject to a long-term transmission power constraint:

$$\mathbb{E}\left[\sum_{m=1}^{M}|p_k[m]|^2\right] \leq P_0, \qquad \forall k, \tag{9}$$

where the expectation is taken over the distribution of random channel coefficients, and we assume $\mathbb{E}\left[\tilde{g}_{k,i}\tilde{g}_{k,i}^*\right]=1$ without loss of generality. Since channel coefficients are identically distributed over different sub-channels, the above power constraint reduces to

$$\mathbb{E}\left[|p_k[m]|^2\right] \leq \frac{P_0}{M}, \qquad \forall k. \tag{10}$$

III. One-Bit Broadband Digital Aggregation (OBDA): System Design

As discussed, the essential idea of OBDA is to integrate signSGD and AirComp so as to support communication-efficient FEEL using digital modulation. The implementation of the idea requires an elaborate system design, which is explained in detail in this section.

A. Transmitter Design

The transmitter design for edge devices is shown in Fig. 2(a). The design builds on a conventional OFDM transmitter with *truncated channel-inversion power control*. However, unlike in conventional communication systems, where coded data bits are passed to the OFDM encoder, here we feed raw quantized bits without any coding. Inspired by signSGD [15], we apply *one-bit quantization* of local gradient estimates, which simply corresponds to taking the signs of the local gradient parameters element-wise:

$$\text{(One-bit quantization)} \qquad \tilde{g}_{k,i} = \mathsf{sign}(g_{k,i}), \quad \forall k,i. \tag{11}$$

Each of the binary gradient parameters is modulated into one *binary phase shift keying* (BPSK) symbol. We emphasize that, even though we use BPSK modulation in our presentation and the convergence analysis for simplicity, the extension of OBDA to 4-QAM configuration is straightforward by simply viewing each 4-QAM symbol as two orthogonal BPSK symbols. Indeed, we employ 4-QAM modulation for the numerical experiments in Section VI. The long symbol sequence is then divided into blocks, and each block of M symbols is transmitted as a single OFDM symbol with one symbol/parameter over each frequency sub-channel.

Assuming perfect CSI at the transmitter, sub-channels are inverted by power control so that gradient parameters transmitted by different devices are received with identical amplitudes, achieving amplitude alignment at the receiver as required for OBDA. Nevertheless, a brute-force approach is inefficient if not impossible under a power constraint since some sub-channels are likely to encounter deep fades. To overcome the issue, we adopt the more practical *truncated-channel-inversion* scheme [6]. To be specific, a sub-channel is inverted only if its gain exceeds a so called *power-cutoff threshold*, denoted by g_{th}, or otherwise allocated zero power. Then the transmission power of the k-th device on the m-th sub-channel, $p_k[m]$, is

$$p_k[m] = \begin{cases} \dfrac{\sqrt{\rho_0}}{|h_k[m]|}\dfrac{h_k[m]^\dagger}{|h_k[m]|}, & |h_k[m]|^2 \geq g_{\mathsf{th}} \\ 0, & |h_k[m]|^2 < g_{\mathsf{th}}, \end{cases} \tag{12}$$

where ρ_0 is a scaling factor set to satisfy the average-transmit-power constraint in (10), and determines the receive power of the gradient update from each device as observed from (8). The exact value of ρ_0 can be computed via

$$\rho_0 = \frac{P_0}{M\mathsf{Ei}(g_{\mathsf{th}})}, \tag{13}$$

where $\mathsf{Ei}(x)$ is the exponential integral function defined as $\mathsf{Ei}(x) = \int_x^\infty \frac{1}{t}\exp(-t)dt$. The result follows from the fact that the channel coefficient is Rayleigh distributed $h_k[m] \sim CN(0,1)$, and thus the channel gain $g_k = |h_k[m]|^2$ follows an exponential distribution with unit mean. Thus the power constraint in (10) is explicitly given by $\rho_0 \int_{g_{\mathsf{th}}}^\infty \frac{1}{g}\exp(-g)dg = \frac{P_0}{M}$. The desired result follows by solving the integral.

We remark that the policy can cause the loss of those gradient parameters that are mapped to the truncated sub-channels. To measure the loss, we define the *non-truncation probability* of a parameter, denoted by α, as the probability that the associated channel gain is above the power-cutoff threshold:

$$\alpha = \mathsf{Pr}(|h_k|^2 \geq g_{\mathsf{th}}) = \exp(-g_{\mathsf{th}}). \tag{14}$$

The result immediately follows from the exponential distribution of the channel gain. The value of α affects the learning convergence rate as shown in the sequel.

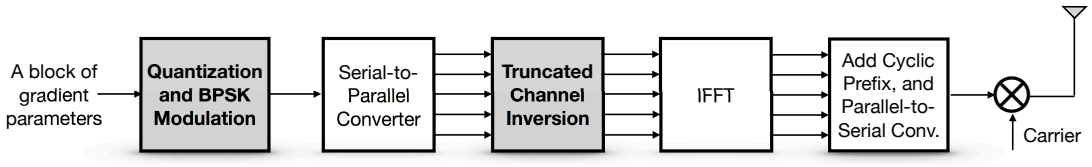
B. Receiver Design

Fig. 2(b) shows the receiver design for the edge server. It has the same architecture as a conventional OFDM receiver except that the digital detector is replaced with a *majority-vote based decoder* for estimating the global gradient-update from the received signal as elaborated in the following.

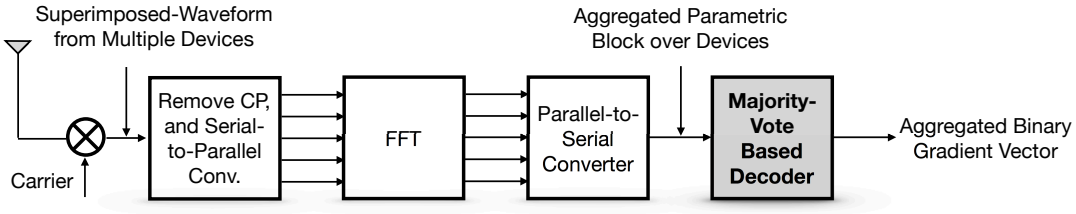
Consider an arbitrary communication round. Given the simultaneous transmission of all participating devices, the server receives superimposed waveforms. By substituting the truncated-channel-inversion policy in (12) into (8), the server obtains the aggregated local-gradient block, denoted by a $M \times 1$ vector $\tilde{\mathbf{g}}[t]$, at the parallel-to-serial converter output [see Fig. 2(b)] as:

$$\text{(Over-the-air aggregation)} \ \tilde{\mathbf{g}}[t] = \sum_{k=1}^{K}\sqrt{\rho_0}\tilde{\mathbf{g}}_k^{(\mathsf{Tr})}[t] + \mathbf{z}[t], \tag{15}$$

where t is the index of the local-gradient block (OFDM symbol) as defined in (8), $\tilde{\mathbf{g}}_k^{(\mathsf{Tr})}[t]$ is the truncated version of $\tilde{\mathbf{g}}_k[t] = [\tilde{g}_{k,(t-1)M+1},\ldots,\tilde{g}_{k,tM}]^T$, with the truncated elements determined by the channel realizations according to (12) and set to zero. Next, cascading all the N_s blocks recovers



(a) Transmitter design for edge devices



(b) Receiver design for edge server

Figure 2. Transceiver design for OBDA.

the full-dimension aggregated one-bit quantized local gradient estimates:

$$\tilde{\mathbf{g}} = \left[\tilde{\mathbf{g}}[1]^T, \tilde{\mathbf{g}}[2]^T, \ldots, \tilde{\mathbf{g}}[N_s]^T\right]^T. \tag{16}$$

Finally, to attain a global gradient estimate from $\tilde{\mathbf{g}}$ for model updating, a majority-vote based decoder is adopted and enforced by simply taking the element-wise sign of $\tilde{\mathbf{g}}$:

$$\text{(Majority-vote based decoder)} \qquad \mathbf{v} = \text{sign}(\tilde{\mathbf{g}}). \tag{17}$$

The operation essentially estimates the global gradient-update by over-the-air element-wise majority vote based on the one-bit quantized local gradient estimates attained at different devices.

Remark 1 (Why majority vote?)**.** Instead of simply taking an average of the one-bit local gradient estimates, majority vote that involves additional sign-taking operation brings additional benefit from the following two aspects: 1) it is more hardware-friendly and robust (against noise) to detect the sign of the aggregated gradient estimate (majority vote) than recovering its actual value in full precision at the edge server; 2) it allows communication-efficient broadcasting of the aggregated gradient estimates back to the edge devices thanks to the one-bit quantized elements after the majority vote.

Then, the server initiates the next communication round by broadcasting the global gradient estimate to all the devices for model updating via

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} - \eta \mathbf{v}^{(n)}, \tag{18}$$

or completes the learning process if the convergence criterion (e.g., target number of communication rounds) is met. We assume that the global gradient parameters can be sent to the devices perfectly, due to the high transmit power available at the edge server and the use of the whole downlink bandwidth for broadcasting.

IV. Convergence Analysis for OBDA over AWGN Channels

In this section, we formally characterize the learning performance of a FEEL system deploying the proposed OBDA scheme in Section III over static AWGN channels; that is, we assume $h_k[t,m] = 1$, $\forall k, t, m$, in this section. Particularly, we focus on understanding how the channel noise affects the convergence rate of the proposed scheme.

A. Basic Assumptions

To facilitate the convergence analysis, several standard assumptions are made on the loss function and computed gradient estimates. In order to allow the developed theory to be applicable to the popular *deep neural networks* (DNNs), we do not assume a convex loss function, but require a lower bounded one as formally stated below, which is the minimal assumption needed for ensuring convergence to a stationary point [33].

Assumption 1 (Bounded Loss Function)**.** For any parameter vector $\mathbf{w}$, the associated loss function is lower bounded by some value F^*, *i.e.,* $F(\mathbf{w}) \geq F^*$, $\forall \mathbf{w}$.

Assumptions 2 and 3 below, on the Lipschitz smoothness and bounded variance, respectively, are standard in the stochastic optimization literature [33].

Assumption 2 (Smoothness)**.** Let $\mathbf{g}$ denote the gradient of the loss function $F(\mathbf{w})$ in (4) evaluated at point $\mathbf{w} = [w_1, \cdots, w_q]$ with q denoting the number of model parameters. We assume that there exists a vector of non-negative constants $\mathbf{L} = [L_1, \cdots, L_q]$, for any $\mathbf{w}, \mathbf{w}'$, that satisfy

$$|F(\mathbf{w}') - [F(\mathbf{w}) + \mathbf{g}^T(\mathbf{w}' - \mathbf{w})]| \leq \frac{1}{2}\sum_{i=1}^{q} L_i (w_i' - w_i)^2. \tag{19}$$

Assumption 3 (Variance Bound)**.** It is assumed that the stochastic gradient estimates $\{\mathbf{g}_j\}$ defined in (5) are independent and unbiased estimates of the batch gradient $\mathbf{g} = \nabla F(\mathbf{w})$ with coordinate bounded variance, i.e.,

$$\mathbb{E}[\mathbf{g}_j] = \mathbf{g}, \qquad \forall j, \tag{20}$$

$$\mathbb{E}[(g_{j,i} - g_i)^2] \leq \sigma_i^2 \qquad \forall j, i, \tag{21}$$

where $g_{j,i}$ and g_i denote the i-th element of $\mathbf{g}_j(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$, respectively, and $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_q]$ is a vector of non-negative constants.

We further assume that the data-stochasticity induced gradient noise, which causes the discrepancy between $\mathbf{g}_j$ and $\mathbf{g}$,

is unimodal and symmetric, as verified by experiments in [15] and formally stated below.

Assumption 4 (Unimodal, Symmetric Gradient Noise)**.** For any given parameter vector $\mathbf{w}$, each element of the stochastic gradient vector $\mathbf{g}_j(\mathbf{w})$, $\forall j$, has a unimodal distribution that is also symmetric around its mean (the ground-truth full-batch gradient elements).

Clearly, Gaussian noise is a special case. Note that even for a moderate mini-batch size, we expect the central limit theorem to take effect and render typical gradient noise distributions close to Gaussian.

B. Convergence Analysis

The above assumptions allow tractable convergence analysis as follows. Given AWGN channels, the gradient aggregation is a direct consequence of the MAC output and the power control in (12) is not needed in this case. Specifically, without truncation due to fading, the full-dimension aggregated local gradient defined in (16) is given by

$$\tilde{\mathbf{g}} = \sum_{k=1}^{K} \sqrt{\rho_0}\tilde{\mathbf{g}}_k + \mathbf{z}, \tag{22}$$

where we have $\rho_0 = \frac{P_0}{M}$ according to the power constraint in (10).

The resulting convergence rate of the proposed OBDA scheme is derived as follows. Throughout the paper, we set the learning rate to $\eta = \frac{1}{\sqrt{\|\mathbf{L}\|_1 n_b}}$ and the mini-batch size to $n_b = \frac{1}{\gamma}N$, with an arbitrary $\gamma > 0$ and N denoting the number of communication rounds.

Theorem 1. Consider a FEEL system deploying OBDA over AWGN channels, the convergence rate is given by

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right] \leq \frac{a_{\mathsf{AWGN}}}{\sqrt{N}}\left(\sqrt{\|\mathbf{L}\|_1}(F^{(0)} - F^* + \frac{\gamma}{2}) + \frac{2\gamma}{\sqrt{K}}\|\boldsymbol{\sigma}\|_1 + b_{\mathsf{AWGN}}\right), \tag{23}$$

where the scaling factor a_{AWGN} and the bias term b_{AWGN} are given by

$$a_{\mathsf{AWGN}} = \frac{1}{1 - \frac{1}{K\sqrt{\rho}}}, \qquad b_{\mathsf{AWGN}} = \frac{2\gamma}{K\sqrt{\rho}}\|\boldsymbol{\sigma}\|_1, \tag{24}$$

and $\rho \triangleq \frac{\rho_0}{\sigma_z^2} = \frac{P_0}{M\sigma_z^2}$ denotes the *receive SNR*.

Proof: See Appendix A. ∎

For comparison, we reproduce below the convergence rate over noiseless channels derived as Theorem 2 in [15].

Lemma 1. The convergence rate with OBDA over error-free channels is given by

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right] \leq \frac{1}{\sqrt{N}}\left(\sqrt{\|\mathbf{L}\|_1}(F^{(0)} - F^* + \frac{\gamma}{2}) + \frac{2\gamma}{\sqrt{K}}\|\boldsymbol{\sigma}\|_1\right). \tag{25}$$

Remark 2 (Effect of Channel Noise)**.** A comparison between the results in Theorem 1 and Lemma 1 reveals that the existence of channel noise slows down the convergence rate by adding a scaling factor and a positive bias term, i.e., a_{AWGN} and b_{AWGN}, respectively, to the upper bound on the time-averaged gradient norm. Due to the increased bound, more communication rounds will be needed for convergence. Nevertheless, the negative effect of channel noise vanishes at a scaling rate of $\frac{1}{K}$ as the number of participating devices grows. We can also see that we recover the convergence rate in Lemma 1 when $\rho \to \infty$.

V. Convergence Analysis for OBDA over Fading Channels

In this section, we extend the convergence result for AWGN channels to the more general case of fading channels. For this case, transmit CSI is needed for power control. We consider both the cases of perfect CSI and imperfect CSI in the analysis. The same set of assumptions as in Section IV-A are made here.

A. Convergence Rate with Perfect CSI

With perfect CSI at each device, the truncated channel inversion power control can be accurately performed. The resultant convergence rate of the OBDA scheme is derived as follows.

Theorem 2. Consider a FEEL system deploying OBDA over fading channels with truncated channel inversion power control using perfect CSI, the convergence rate is given by

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right] \leq \frac{a_{\mathsf{FAD}}}{\sqrt{N}}\left(\sqrt{\|\mathbf{L}\|_1}(F^{(0)} - F^* + \frac{\gamma}{2}) + \frac{2\gamma}{\sqrt{K}}\|\boldsymbol{\sigma}\|_1 + b_{\mathsf{FAD}}\right), \tag{26}$$

where the scaling factor a_{FAD} and the bias term b_{FAD} are

$$a_{\mathsf{FAD}} = \frac{1}{1 - (1-\alpha)^K - \frac{2}{\alpha K\sqrt{\rho}}}, \; b_{\mathsf{FAD}} = \frac{4\gamma}{\alpha K\sqrt{\rho}}\|\boldsymbol{\sigma}\|_1, \tag{27}$$

and $\rho \triangleq \frac{\rho_0}{\sigma_z^2} = \frac{P_0}{M\mathsf{Ei}(g_{\mathsf{th}})\sigma_z^2}$ denotes the average *receive SNR*.

Proof: See Appendix C ∎

Remark 3 (Effect of Channel Fading)**.** A comparison between Theorems 1 and 2 reveals that the existence of channel fading further slows down the convergence rate of the OBDA by introducing a larger scaling factor and a bias term: $a_{\mathsf{FAD}} > a_{\mathsf{AWGN}}$ and $b_{\mathsf{FAD}} > b_{\mathsf{AWGN}}$. This negative effect of channel fading vanishes at a scaling rate of $\frac{1}{\alpha K}$ as the number of participating devices grows. Compared with the AWGN counterpart, the rate is slowed down by a factor of α. The degradation is due to the gradient truncation induced by truncated channel inversion power control to cope with fading.

B. Convergence Rate with Imperfect CSI

In practice, there may exist channel estimation errors that lead to imperfect channel inversion and, as a result, reduces the convergence rate. To facilitate the analysis, we adopt the bounded channel estimation error model (see e.g., [34]), where

the estimated CSI is a perturbed version of the ground-true one and the additive perturbation, denoted as Δ, is assumed to be bounded:

$$\hat{h}_k[m] = h_k[m] + \Delta, \qquad \forall k, m, \tag{28}$$

where we assume the absolute value of the perturbation is bounded by $|\Delta| \leq \Delta_{\max} \ll \sqrt{g_{\mathsf{th}}}$[6] with a zero mean $\mathbb{E}(\Delta) = 0$, and a variance of $\mathsf{Var}(\Delta) = \sigma_\Delta^2$.

Based on the above CSI model, the model convergence rate can be derived below.

Theorem 3. Consider a FEEL system deploying OBDA over fading channels with truncated channel inversion power control using imperfect CSI, the convergence rate is given by

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right] \leq \frac{a_{\mathsf{CERR}}}{\sqrt{N}}\left(\sqrt{\|\mathbf{L}\|_1}(F^{(0)} - F^* + \frac{\gamma}{2}) + \frac{2\gamma}{\sqrt{K}}\|\boldsymbol{\sigma}\|_1 + b_{\mathsf{CERR}}\right), \tag{29}$$

where the scaling factor a_{CERR} and the bias term b_{CERR} are given by

$$a_{\mathsf{CERR}} = \frac{1}{1-(1-\alpha)^K - \frac{2}{\alpha K\sqrt{\rho}} - \frac{2\sqrt{6}\sigma_\Delta}{\sqrt{\alpha K}\sqrt{\sqrt{g_{\mathsf{th}}}-\Delta_{\max}}}},$$
$$b_{\mathsf{CERR}} = \left(\frac{4}{\alpha K\sqrt{\rho}} + \frac{4\sqrt{6}\sigma_\Delta}{\sqrt{\alpha K}\sqrt{\sqrt{g_{\mathsf{th}}}-\Delta_{\max}}}\right)\gamma\|\boldsymbol{\sigma}\|_1, \tag{30}$$

and $\rho \triangleq \frac{\rho_0}{\sigma_z^2} = \frac{P_0}{M\mathsf{Ei}(g_{\mathsf{th}})\sigma_z^2}$ denotes the average *receive SNR*.

Proof: See Appendix D ∎

Remark 4 (Effect of Imperfect CSI)**.** Comparing Theorem 3 and Theorem 2, one can observe that the imperfect CSI reduces the convergence rate for OBDA even further as reflected by larger scaling factor and bias terms: $a_{\mathsf{CERR}} > a_{\mathsf{FAD}}$ and $b_{\mathsf{CERR}} > b_{\mathsf{FAD}}$. Particularly, with imperfect CSI, the negative effect of channel fading vanishes at a slower scaling law of $\frac{1}{\sqrt{\alpha K}}$ as the number of participating devices increases. In contrast, it is a $\frac{1}{\alpha K}$ for the perfect CSI case, which is much faster. The results in (30) also quantify the effect of the level of CSI accuracy (represented by $\Delta_{\max}$) on the convergence rate for the proposed scheme.

VI. Simulation results

For numerical experiments we consider a FEEL system with one edge server and $K = 100$ edge devices. The simulation settings are given as follows unless specified otherwise. The number of sub-channels is $M = 1000$, and the average receive SNR, defined as $\rho = \frac{P_0}{M\sigma_z^2}$, is set to be 10 dB. We consider the learning task of image classification using the well-known MNIST and CIFAR10 datasets, where the former consists of 10 classes of black-and-white digits ranging from "0" to "9" and the latter comprises 10 categories of colorful objectives such as airplanes, cars, etc.. In particular, for the MNIST dataset, as illustrated in Fig. 3, the classifier model is implemented using a 6-layer *convolutional neural network* (CNN) that consists of two 5×5 convolution layers with ReLU activation (the first with 32 channels, the second with 64), each followed with a 2×2 max pooling; a fully connected layer with 512 units, ReLU activation; and a final softmax output layer (i.e., $q = 582,026$). While for the CIFAR10 dataset, the well-known classifier model, ResNet18 with batch normalization proposed in [35], is applied. We adopt the 4-QAM for the quantized gradient element modulation, where the odd and even gradient coefficients are mapped to the real and imaginary parts of 4-QAM symbols, respectively.

A. Performance Evaluation of OBDA

For both MNIST and CIFAR10 datasets, the effectiveness of the OBDA is evaluated in the three considered scenarios, namely over an AWGN MAC, and fading MACs with and without perfect CSI, which represent three levels of wireless hostilities. Test accuracy is plotted as a function of the number of communication rounds in Fig. 4. The proposed OBDA scheme converges in all three scenarios, but at different rates depending on the level of wireless hostility the scheme suffers from. In the presence of channel fading the convergence is slower compared with its counterpart over an AWGN channel. This is because part of the gradient signs corresponding to subchannels experiencing deep fade are truncated, rendering a smaller number of effective participating devices for each gradient entry. The imperfect CSI further slows down the convergence of the proposed OBDA. This is due to inaccurate aggregation, which results in a deviated gradient for model updating. These observations are aligned with our analysis presented in Theorems 1-3.

B. Effect of Device Population

The effect of the device population on the convergence behaviour is illustrated in Fig. 5, where we set the number of communication-rounds as 150 for the MNIST dataset and 200 for the CIFAR10 dataset, and plot the curves of test accuracy w.r.t. the total number of edge-devices K for the three considered scenarios. It is observed that, in all scenarios, the test accuracy grows as K increases. This is because a larger K suppresses the noise variance inherent in stochastic gradients as well as the negative effects due to wireless hostilities. This phenomenon is coined as *majority-vote gain*. In particular, the majority-vote gain is observed to be the most prominent in the gentle wireless condition (i.e., AWGN), and weakened in the hostile one (i.e., fading with imperfect CSI). This behaviour is aligned with analytical results in Theorems 1-3, which showed that the negative effects introduced by different wireless hostilities vanish, at different rates, with the growth of the device population.

C. Performance Comparison: OFDMA, BAA and OBDA

A performance comparison between digital transmission using OFDMA, BAA developed in [6], and the proposed OBDA is presented in Fig. 6. In this figure, the test accuracy and communication latency are plotted for these three schemes over a

[6]When there exist channel estimation errors, it is desirable to set a relative high channel cutoff threshold g_{th} to ensure that $g_{\mathsf{th}} \gg \Delta_{\max}^2$ for avoiding the channel truncation decision misled by the estimation perturbation Δ.

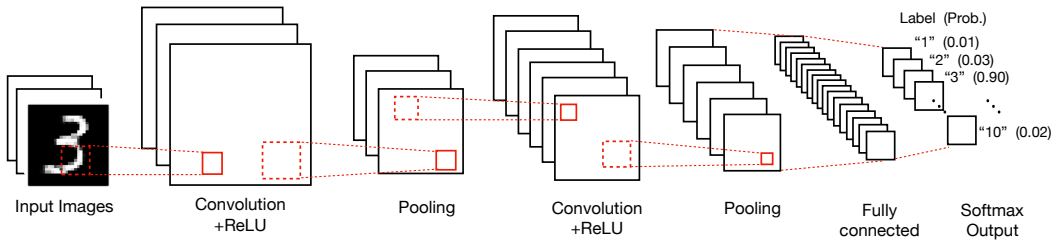
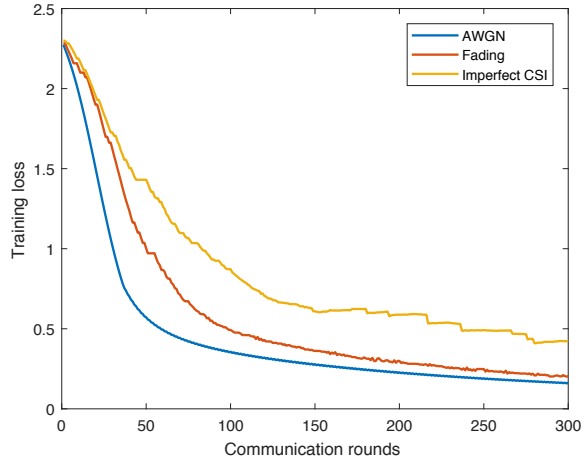
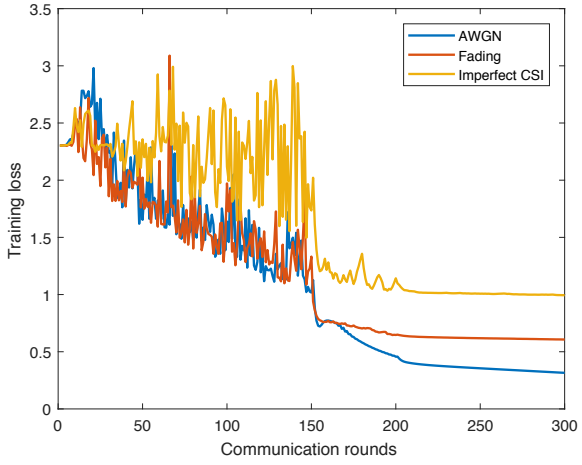


Figure 3. Architecture of the CNN used in our experiments.

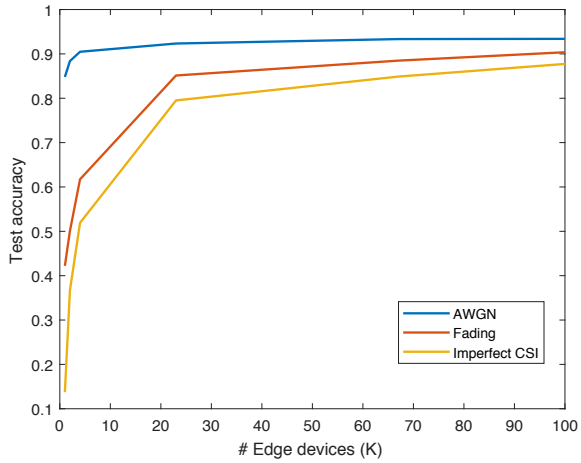


(a) Dataset: MNIST; Classifier: CNN

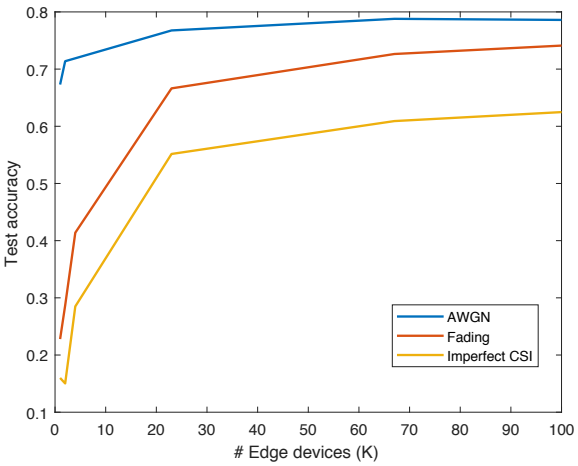


(b) Dataset: CIFAR10; Classifier: ResNet18

Figure 4. Convergence performance of FEEL using OBDA.



(a) Dataset: MNIST; Classifier: CNN



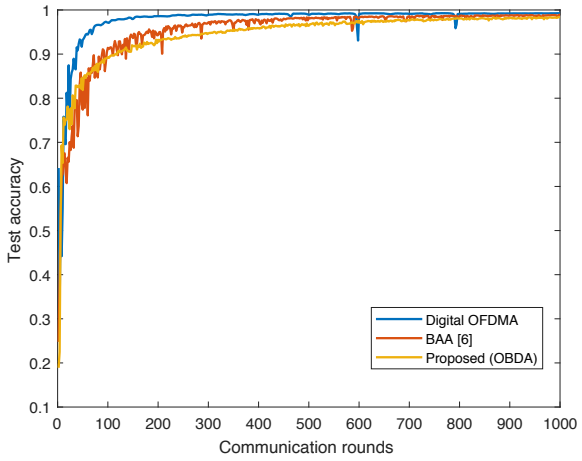
(b) Dataset: CIFAR10; Classifier: ResNet18

Figure 5. Effect of device population on the convergence.

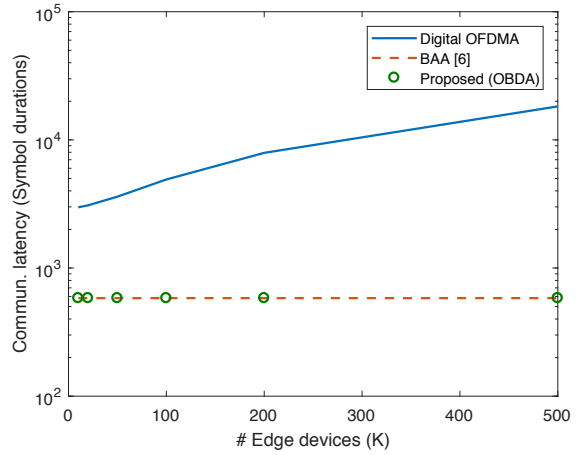
fading MAC with perfect CSI. For the digital OFDMA, sub-channels are evenly allocated to the edge devices; gradient-update parameters are quantized into bit sequences with 16-bit per coefficient; and adaptive MQAM modulation is used to maximize the spectrum efficiency under a target bit error rate of 10^{-3}. It can be observed from Fig. 6(a) that the convergence speed of digital OFDMA, BAA and OBDA are in descending order while all three schemes achieve comparable accuracies after sufficient number of communication rounds. The reason behind the faster convergence of digital OFDMA w.r.t. BAA is that the direct exposure of the analog modulated signals in BAA to the channel hostilities results in a boosted expected gradient norm as mentioned in Remarks 1-3. The performance gap between the BAA and OBDA in terms of convergence speed arises from the quantization loss introduced by the latter. However, we observe that the gap between the two is small, which shows that over-the-air gradient aggregation can be employed with devices employing simple 4-QAM modulation without significant performance loss w.r.t. analog aggregation. Moreover, Fig. 6(b) shows that, without compromising the learning accuracies, the *per-round communication latencies* for both OBDA and BAA are independent of the number of devices, while that of the digital OFDMA scales up as the device population grows.

VII. Concluding Remarks

In the context of FEEL, we have proposed a novel digital over-the-air gradient aggregation scheme, called OBDA, for



(a) Test accuracy



(b) Communication latency

Figure 6. Performance comparison among digital OFDMA, BAA and OBDA.

communication-efficient distributed learning across wireless devices. To understand its performance, a comprehensive convergence analysis framework for OBDA subject to wireless channel hostilities is developed. This work represents the first attempt to develop digital AirComp, which is more practical, compared to its analog counterpart, in terms of the compatibility with the modern digital communication infrastructure. For future work, we will consider the generalization of the current work to multi-cell FEEL, where the effect of inter-cell interference should also be taken into account. As another interesting direction, the proposed design assuming a single learning task can be extended to the multi-task learning scenario, where an additional task scheduler at the server needs to be designed in an effort to reducing the frequency of the gradient updates, further accelerating the learning process.

APPENDIX

A. Proof of Theorem 1

The proof is conducted following the widely-adopted strategy of relating the norm of the gradient to the expected improvement made in a single algorithmic step, and comparing this with the total possible improvement under Assumption 1. A key technical challenge we overcome is in showing how to directly deal with a *biased gradient estimate* $\mathbf{v}$ of the full-batch gradient $\mathbf{g}$ as specified in the sequel. Note that the current subsection also serves as another purpose of presenting the general framework for convergence analysis of OBDA, which also applies to the later extension to the more complicated scenarios.

To start with, we first bound the improvement of the objective during a single step of the algorithm for one instantiation of the data-stochasticity induced noise based on Assumption 2. To this end, substituting the step in (18) to (19) and decomposing the improvement to expose the (data-and-channel) stochasticity-induced error we have:

$$\begin{aligned}
&F^{(n+1)}-F^{(n)}\\
\leq&(\mathbf{g}^{(n)})^T(\mathbf{w}^{(n+1)}-\mathbf{w}^{(n)})+\frac{1}{2}\sum_{i=1}^{q}L_i\left(w_i^{(n+1)}-w_i^{(n)}\right)^2,\\
=&-\eta(\mathbf{g}^{(n)})^T\mathsf{sign}(\tilde{\mathbf{g}}^{(n)})+\eta^2\sum_{i=1}^{q}\frac{L_i}{2},\\
=&-\eta\|\mathbf{g}^{(n)}\|_1+\frac{\eta^2}{2}\|\mathbf{L}\|_1+\\
&\underbrace{2\eta\sum_{i=1}^{q}|g_i^{(n)}|\mathbb{I}[\mathsf{sign}(\tilde{g}_i^{(n)})\neq\mathsf{sign}(g_i^{(n)})]}_{\mathsf{Stochasticity-induced\ \ error}},
\end{aligned}\tag{31}$$

where $\tilde{g}_i^{(n)}$ and $g_i^{(n)}$ denote the i-th element of $\tilde{\mathbf{g}}^{(n)}$ and $\mathbf{g}^{(n)}$, respectively.

Next we find the expected improvement at time $n+1$ conditioned on the previous iterate.

$$\begin{aligned}
\mathbb{E}[F^{(n+1)}-F^{(n)}|\mathbf{w}^{(n)}]\leq&-\eta\|\mathbf{g}^{(n)}\|_1+\frac{\eta^2}{2}\|\mathbf{L}\|_1+\\
&\underbrace{2\eta\sum_{i=1}^{q}|g_i^{(n)}|\mathbb{P}[\mathsf{sign}(\tilde{g}_i^{(n)})\neq\mathsf{sign}(g_i^{(n)})]}_{\mathsf{Stochasticity-induced\ \ error}},
\end{aligned}\tag{32}$$

where $\mathbf{g}^{(n)}$ is a constant vector due to the conditioning. It can be noted from (32) that the expected improvement on the objective crucially depends on the *(decoding) bit error probability*, i.e.,

$$P_i^{\mathsf{err}}=\mathbb{P}[\mathsf{sign}(\tilde{g}_i^{(n)})\neq\mathsf{sign}(g_i^{(n)})],\tag{33}$$

which is intuitively determined by the level of the noise introduced by the data-stochasticity and the wireless channel. To formalize the intuition, we have the following lemma.

Lemma 2. The bit error probability in the AWGN channel is bounded by

$$P_i^{\mathsf{err}}\leq\frac{1}{\sqrt{K}S_i}+\frac{\sigma_z}{KS_i\sqrt{\rho_0}}+\frac{\sigma_z}{2K\sqrt{\rho_0}},\tag{34}$$

where we have defined $S_i=\sqrt{n_b}\frac{|g_i^{(n)}|}{\sigma_i}$ as the *gradient-signal-to-data-noise ratio*. The coefficient $\sqrt{n_b}$ is because each of the local gradient estimate $\tilde{g}_{k,i}$ is computed over a mini-batch of size n_b, thus the resultant gradient variance reduces from σ_i^2 to $\frac{\sigma_i^2}{n_b}$ according to Assumption 3 and Equation (5).

Proof: See Appendix B ■

Then, substituting Lemma 2 into (32) we have

$$\begin{aligned}
&\mathbb{E}[F^{(n+1)}-F^{(n)}|\mathbf{w}^{(n)}]\\
\leq&-\eta\|\mathbf{g}^{(n)}\|_1+\frac{\eta^2}{2}\|\mathbf{L}\|_1+\\
&\qquad 2\frac{\eta}{\sqrt{n_b}}\left(\frac{1}{\sqrt{K}}+\frac{\sigma_z}{K\sqrt{\rho_0}}\right)\|\boldsymbol{\sigma}\|_1+\frac{\eta\sigma_z}{K\sqrt{\rho_0}}\|\mathbf{g}^{(n)}\|_1,\\
=&\eta\left(\frac{\sigma_z}{K\sqrt{\rho_0}}-1\right)\|\mathbf{g}^{(n)}\|_1+\frac{\eta^2}{2}\|\mathbf{L}\|_1+\\
&\qquad 2\frac{\eta}{\sqrt{n_b}}\left(\frac{1}{\sqrt{K}}+\frac{\sigma_z}{K\sqrt{\rho_0}}\right)\|\boldsymbol{\sigma}\|_1. \quad (35)
\end{aligned}$$

Further plug in the learning rate and mini-batch settings $\eta=\frac{1}{\sqrt{\|\mathbf{L}\|_1 n_b}}$ and $n_b=\frac{1}{\gamma}N$, we have

$$\begin{aligned}
\mathbb{E}[F^{(n+1)}-F^{(n)}|\mathbf{w}^{(n)}]\leq&\frac{1}{\sqrt{\|\mathbf{L}\|_1 N}}\left(\frac{\sigma_z}{K\sqrt{\rho_0}}-1\right)\|\mathbf{g}^{(n)}\|_1+\\
&\frac{\gamma}{2N}+\frac{2\gamma}{\sqrt{\|\mathbf{L}\|_1 N}}\left(\frac{1}{\sqrt{K}}+\frac{\sigma_z}{K\sqrt{\rho_0}}\right)\|\boldsymbol{\sigma}\|_1.
\end{aligned}$$

Now we take the expectation over $\mathbf{w}^{(n)}$ to average out the randomness in the optimization trajectory and perform a telescoping sum over the iterates:

$$\begin{aligned}
&F^{(0)}-F^*\\
\geq&F^{(0)}-\mathbb{E}[F^{(n)}]=\mathbb{E}\left[\sum_{n=0}^{N-1}F^{(n)}-F^{(n+1)}\right],\\
\geq&\mathbb{E}\left[\sum_{n=0}^{N-1}\frac{1}{\sqrt{\|\mathbf{L}\|_1 N}}\left(1-\frac{\sigma_z}{K\sqrt{\rho_0}}\right)\|\mathbf{g}^{(n)}\|_1-\right.\\
&\left.\frac{\gamma}{2\sqrt{\|\mathbf{L}\|_1 N}}\left((\frac{4}{\sqrt{K}}+\frac{4\sigma_z}{K\sqrt{\rho_0}})\|\boldsymbol{\sigma}\|_1+\sqrt{\|\mathbf{L}\|_1}\right)\right],\\
=&\sqrt{\frac{N}{\|\mathbf{L}\|_1}}\left(1-\frac{\sigma_z}{K\sqrt{\rho_0}}\right)\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\mathbf{g}^{(n)}\|_1\right]-\\
&\frac{\gamma}{2\sqrt{\|\mathbf{L}\|_1}}\left((\frac{4}{\sqrt{K}}+\frac{4\sigma_z}{K\sqrt{\rho_0}})\|\boldsymbol{\sigma}\|_1+\sqrt{\|\mathbf{L}\|_1}\right). \quad (36)
\end{aligned}$$

In the end, the desired result in Theorem 1 can be easily obtained by rearranging the terms in (36), which completes the proof.

B. Proof of Lemma 2

The key idea of the proof is to establish an equivalent mathematical event of $\mathsf{sign}(\tilde{g}_i^{(n)})=\mathsf{sign}(g_i^{(n)})$ described by well-defined random variables with known distributions. To this end, let X_i denote the number of edge devices with correct sign bit at the i-th element of the gradient vector, namely, that with $\mathsf{sign}(g_{k,i}^{(n)})=\mathsf{sign}(g_i^{(n)})$, and $\tilde{X}_i=X_i+\tilde{z}_i$ denote the noisy version of X_i corrupted by the effective channel noise $\tilde{z}_i=\frac{z_i}{2\sqrt{\rho_0}}\sim\mathcal{N}(0,\frac{\sigma_z^2}{4\rho_0})$. Note that X_i is the sum of K independent Bernoulli trials, and thus binomial with success probability and failure probability denoted by

$$p_i=\mathbb{P}[\mathsf{sign}(g_{k,i}^{(n)})=\mathsf{sign}(g_i^{(n)})], \quad (37)$$

$$q_i=\mathbb{P}[\mathsf{sign}(g_{k,i}^{(n)})\neq\mathsf{sign}(g_i^{(n)})], \quad (38)$$

respectively. Then we derive the mean and variance of $\tilde{X}_i$ as follows:

$$\mathbb{E}[\tilde{X}_i]=Kp_i=K\left(\epsilon_i+\frac{1}{2}\right), \quad (39)$$

$$\mathsf{Var}(\tilde{X}_i)=Kp_iq_i+\frac{\sigma_z^2}{4\rho_0}=K\left(\frac{1}{4}-\epsilon_i^2\right)+\frac{\sigma_z^2}{4\rho_0}, \quad (40)$$

where we have defined $\epsilon_i=p_i-1/2=1/2-q_i$ for ease of subsequent derivation.

According to (22), to ensure that

$$\mathsf{sign}(\tilde{g}_i^{(n)})=\mathsf{sign}\left(\sum_{k=1}^{K}\sqrt{\rho_0}\tilde{g}_{k,i}^{(n)}+z_i\right)=\mathsf{sign}(g_i^{(n)}), \quad (41)$$

$\tilde{X}_i$ must be larger than $\frac{K}{2}$. Therefore we have

$$P_i^{\mathsf{err}}=\mathbb{P}\left(\tilde{X}_i\leq\frac{K}{2}\right)=\mathbb{P}\left[\mathbb{E}[\tilde{X}_i]-\tilde{X}_i\geq\mathbb{E}[\tilde{X}_i]-\frac{K}{2}\right]. \quad (42)$$

Applying the known Cantellis' inequality, $\mathbb{P}(X-\mathbb{E}[X]\geq\lambda)\leq\frac{\mathsf{var}(X)}{\mathsf{var}(X)+\lambda^2}$, $\lambda>0$, on (42) yields

$$P_i^{\mathsf{err}}\leq\frac{1}{1+\frac{(\mathbb{E}[\tilde{X}_i]-K/2)^2}{\mathsf{var}(\tilde{X}_i)}}\leq\frac{1}{2}\sqrt{\frac{\mathsf{var}(\tilde{X}_i)}{\mathbb{E}[\tilde{X}_i]-\frac{K}{2}}}, \quad (43)$$

where the second inequality is due to the fact that $1+a^2\geq 2a$. Next, we plug in the statistics of $\tilde{X}$ in (39) and (40), to obtain

$$\begin{aligned}
P_i^{\mathsf{err}}&\leq\frac{1}{2}\sqrt{\frac{1}{K}\left(\frac{1}{4\epsilon_i^2}-1\right)+\frac{\sigma_z^2}{K^2\epsilon_i^2 4\rho_0}},\\
&\leq\frac{1}{2}\sqrt{\frac{1}{K}\left(\frac{1}{4\epsilon_i^2}-1\right)}+\frac{1}{2}\frac{\sigma_z}{\epsilon_i K\sqrt{4\rho_0}}, \quad (44)
\end{aligned}$$

where the second inequality is due to the fact that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$.

To proceed with, we need a bound on $\epsilon_i=1/2-q_i$ that can relate it to the gradient-signal-to-data-noise ratio S_i defined in Lemma 2. The bound can be attained from the following bound on q_i, the failure probability for the sign bit of a single device, under the unimodal symmetric gradient noise assumption stated in Assumption 4.

Lemma 3. Under the assumption of unimodal symmetric gradient noise in Assumption 4, the failure probability for the sign bit of a single device can be bounded by

$$\begin{aligned}
q_i&=\mathbb{P}\left[\mathsf{sign}(\tilde{g}_{k,i}^{(n)})\neq\mathsf{sign}(g_i^{(n)})\right],\\
&\leq\begin{cases}\frac{2}{9}\frac{1}{S_i^2}, & \text{if } S_i>\frac{2}{\sqrt{3}}\\ \frac{1}{2}-\frac{S_i}{2\sqrt{3}}, & \text{otherwise.}\end{cases}\qquad\forall i. \quad (45)
\end{aligned}$$

Proof: The result follows from the known Gauss' inequality recalled below. For a unimodal symmetric random variable X with mean μ and variance σ^2, the below inequality holds:

$$\mathbb{P}[|X-\mu|>x]\leq\begin{cases}\frac{4}{9}\frac{\sigma^2}{x^2}, & \text{if } \frac{x}{\sigma}>\frac{2}{\sqrt{3}}\\ 1-\frac{x}{\sqrt{3}\sigma}, & \text{otherwise.}\end{cases} \quad (46)$$

Without loss of generality, assume that $g_i^{(n)}$ is negative. Then applying the Gauss' inequality, the bound on the failure probability for the sign bit can be derived as follows:

$$\begin{aligned} q_i &= \mathbb{P}\left[\mathsf{sign}(\tilde{g}_{k,i}^{(n)}) \neq \mathsf{sign}(g_i^{(n)})\right] = \mathbb{P}\left[\tilde{g}_{k,i}^{(n)} - g_i^{(n)} \geq |g_i^{(n)}|\right], \\ &= \frac{1}{2}\mathbb{P}\left[|\tilde{g}_{k,i}^{(n)} - g_i^{(n)}| \geq |g_i^{(n)}|\right], \\ &\leq \begin{cases} \frac{2}{9}\frac{\sigma_i^2}{n_b|g_i^{(n)}|^2}, & \text{if } \frac{|g_i^{(n)}|}{\sigma_i/\sqrt{n_b}} > \frac{2}{\sqrt{3}} \\ \frac{1}{2} - \frac{|g_i^{(n)}|}{2\sqrt{3}\sigma_i/\sqrt{n_b}}, & \text{otherwise,} \end{cases} \end{aligned} \tag{47}$$

where the term $\sqrt{n_b}$ is due to that each of the local gradient estimate $\tilde{g}_{k,i}$ is computed over a mini-batch of size n_b. Finally, the desired result is obtained by noting $S_i = \sqrt{n_b}\frac{|g_i^{(n)}|}{\sigma_i}$. ■

Next, combining the equation $\epsilon_i = 1/2 - q_i$ and Lemma 3, we can obtain:

$$\frac{1}{4\epsilon_i^2} - 1 \leq \frac{4}{S_i^2}, \tag{48}$$

whose proof involves only simple algebraic manipulations, e.g., checking the monotonicity of the piece-wise function on the right hand side of (47), and is skipped here for brevity. Note that (48) further implies that

$$\frac{1}{\epsilon_i} \leq 2\sqrt{\frac{4}{S_i^2} + 1} \leq \frac{4}{S_i} + 2, \tag{49}$$

which follows from the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Finally, by substituting (48) and (49) into (44), the error probability of received sign vector can be further bounded by

$$P_i^{\mathsf{err}} \leq \frac{1}{\sqrt{K}S_i} + \left(\frac{2}{S_i} + 1\right)\frac{\sigma_z}{K\sqrt{4\rho_0}}. \tag{50}$$

This gives the desired result by simply rearranging the terms.

C. Proof of Theorem 2

The convergence analysis for the fading channel case also follows the general strategy of relating the norm of the gradient to the expected improvement in objective as presented in Appendix A. Specifically, the expression for calculating the single-step expected improvement in (32) still applies for the fading channel case, while a new expression for bit error probability P_i^{err} defined in (33) need be derived to account for the additional randomness introduced by channel fading. This is the key challenge tackled in the following proof.

Due to the adoption of the truncated channel inversion power control in (12), the random channel fading makes the devices sending the i-th gradient element a random set denoted by $\mathcal{K}_i$, with its size denoted by K_i. We can rewrite the i-th aggregated gradient element at the channel output in (8) as follows:

$$\tilde{g}_i^{(n)} = \sum_{k\in\mathcal{K}_i} \sqrt{\rho_0}\tilde{g}_{k,i}^{(n)} + z_i^{(n)}, \tag{51}$$

Conditioned on the size of the transmitting set K_i, the error probability of received sign bit is first derived. Based on this, the unconditional error probability will be derived next.

Lemma 4. Consider the fading channel with truncated channel inversion power control with perfect CSI. The bit error probability conditioned on the size of the transmitting set K_i is given by

$$\begin{aligned} P_i^{\mathsf{err}}(K_i) &= \mathbb{P}[\mathsf{sign}(\tilde{g}_i^{(n)}) \neq \mathsf{sign}(g_i^{(n)})|K_i], \\ &\leq \begin{cases} \frac{1}{\sqrt{K_i}S_i} + \frac{1}{K_i}\frac{\sigma_z}{\sqrt{\rho_0}}\left(\frac{1}{S_i} + \frac{1}{2}\right), & K_i \geq 1 \\ \frac{1}{2}, & K_i = 0. \end{cases} \end{aligned} \tag{52}$$

Proof: For the case with non-empty transmitting set, i.e., $K_i \geq 1$, the result directly follows Lemma 2 by noting the similarity between the channel model in (51) and that in (22). While for the case $K_i = 0$, no device is transmitting, and thus only channel noise is received at the edge server, thereby the decoding process reduces to a random guess, with an error probability of $\frac{1}{2}$. ■

We note that the size of the transmitting set, K_i follows a binomial distribution $K_i \sim B(K, \alpha)$ with α denoting the non-truncation probability derived in (14). This immediately leads to the following results:

$$\mathbb{P}(K_i = 0) = (1-\alpha)^K, \tag{53}$$

$$\mathbb{P}(K_i \geq 1) = \sum_{k=1}^{K}\binom{K}{k}\alpha^k(1-\alpha)^{K-k}. \tag{54}$$

By the law of total probability, the unconditioned error probability P_i^{err} can be computed via

$$P_i^{\mathsf{err}} = P_i^{\mathsf{err}}(K_i = 0)\mathbb{P}(K_i = 0) + P_i^{\mathsf{err}}(K_i \geq 1)\mathbb{P}(K_i \geq 1). \tag{55}$$

Then, by substituting Lemma 4 and the results in (53) into (55), we can compute a bound on the unconditional error probability as follows:

$$\begin{aligned} P_i^{\mathsf{err}} \leq \frac{1}{2}(1-\alpha)^K + \sum_{k=1}^{K}\binom{K}{k}\alpha^k(1-\alpha)^{K-k}\times \\ \left[\frac{1}{\sqrt{k}S_i} + \frac{1}{k}\frac{\sigma_z}{\sqrt{\rho_0}}\left(\frac{1}{S_i} + \frac{1}{2}\right)\right]. \end{aligned} \tag{56}$$

To simplify (56), we find it useful to establish the following two important inequalities.

Lemma 5. The following two inequalities hold:

$$f(K,\alpha) \equiv \sum_{k=1}^{K}\frac{1}{k}\binom{K}{k}\alpha^k(1-\alpha)^{K-k} \leq \frac{2}{K\alpha}, \tag{57}$$

$$g(K,\alpha) \equiv \sum_{k=1}^{K}\frac{1}{\sqrt{k}}\binom{K}{k}\alpha^k(1-\alpha)^{K-k} \leq \frac{\sqrt{6}}{\sqrt{K\alpha}}. \tag{58}$$

Proof: We start with the first inequality. Function $f(K,\alpha)$

can be rewritten as follows:

$$\begin{aligned} &f(K,\alpha)\\ =&K\alpha\sum_{k=1}^{K}\frac{1}{k^2}\binom{K-1}{k-1}\alpha^{k-1}(1-\alpha)^{K-k},\\ =&K\alpha\sum_{k=0}^{K-1}\frac{1}{(k+1)^2}\binom{K-1}{k}\alpha^{k}(1-\alpha)^{K-k-1},\\ =&K\alpha\sum_{k=0}^{K-1}\frac{k+2}{k+1}\frac{1}{(k+1)(k+2)}\binom{K-1}{k}\alpha^{k}(1-\alpha)^{K-k-1}. \end{aligned}$$

Note that since $\frac{k+2}{k+1}\leq 2$, the function $f(K,\alpha)$ is bounded by (59) as shown on the top of next page.

Next, we move to the proof of the second inequality. Similarly, we have

$$\begin{aligned} g(K,\alpha)=&K\alpha\sum_{k=1}^{K}\frac{1}{k\sqrt{k}}\binom{K-1}{k-1}\alpha^{k-1}(1-\alpha)^{K-k}, \qquad (60)\\ =&K\alpha\sum_{k=0}^{K-1}\left(\frac{1}{k+1}\right)^{\frac{3}{2}}\binom{K-1}{k}\alpha^{k}(1-\alpha)^{K-k-1}. \end{aligned}$$

We remark that the undesired square root operation on $\frac{1}{k+1}$ makes it harder to derive a bound on $g(K,\alpha)$ as the tricks used for deriving (57) cannot be applied here. Nevertheless the challenge can be overcome by rewritting (60) as

$$g(K,\alpha)=K\alpha\mathbb{E}\left[\left(\frac{1}{k+1}\right)^{\frac{3}{2}}\right], \qquad (61)$$

where the expectation is taken over a random variable k that follows the binomial distribution $B(K-1,\alpha)$. Then by Jensen's inequality, we have

$$\mathbb{E}\left[\left(\frac{1}{k+1}\right)^{\frac{3}{2}}\right]\leq\sqrt{\mathbb{E}\left(\frac{1}{k+1}\right)^{3}}. \qquad (62)$$

This suggests that we can obtain a bound on $g(K,\alpha)$ by bounding $\mathbb{E}\left[\left(\frac{1}{k+1}\right)^{3}\right]$, which is derived below. We have

$$\begin{aligned} \mathbb{E}\left[\left(\frac{1}{k+1}\right)^{3}\right]=&K\alpha\sum_{k=0}^{K-1}\frac{k+3}{k+1}\frac{k+2}{k+1}\times\\ &\frac{1}{(k+1)(k+2)(k+3)}\binom{K-1}{k}\alpha^{k}(1-\alpha)^{K-k-1}. \qquad (63) \end{aligned}$$

Note that since $\frac{k+3}{k+1}\frac{k+2}{k+1}\leq 6$, the expectation can be further bounded by (64) as shown on the top of next page. Finally, plugging (64) and (62) into (61) we can attain the desired second inequality. ∎

Then applying Lemma 5 on (56) gives the unconditional bit error probability as follows.

Lemma 6. Consider the fading channel with truncated channel inversion power control with perfect CSI. The unconditional bit error probability is bounded by

$$P_i^{\mathsf{err}}\leq\frac{1}{2}(1-\alpha)^K+\frac{\sqrt{6}}{\sqrt{\alpha K}S_i}+\frac{2}{\alpha K}\frac{\sigma_z}{\sqrt{\rho_0}}\left(\frac{1}{S_i}+\frac{1}{2}\right). \qquad (65)$$

With Lemma 6 at hand, the desired result in Theorem 2 can be easily derived by substituting Lemma 6 into (32) and following the same machinery presented in (35)-(36).

D. Proof of Theorem 3

Again, the same analysis framework presented in Appendix A applies, while the remaining work is to derive a new bit error probability P_i^{err} (defined in (33)) to account for the additional error introduced by the imperfect CSI as presented in the following.

First, we define $\mathcal{K}_i=\{k\mid|\hat{h}_k|^2\geq g_{\mathsf{th}}\}$ as the set of devices transmitting the i-th gradient element, whose estimated channel gain is larger than the cutoff threshold. Then, according to the imperfect CSI model in (28), we can rewrite the i-th aggregated gradient element at the channel output in (8) as follows:

$$\tilde{g}_i^{(n)}=\sum_{k\in\mathcal{K}_i}h_k\frac{\sqrt{\rho_0}}{\hat{h}_k}\tilde{g}_{k,i}^{(n)}+z_i=\sum_{k\in\mathcal{K}_i}\frac{\sqrt{\rho_0}}{1+\frac{\Delta}{h_k}}\tilde{g}_{k,i}^{(n)}+z_i. \qquad (66)$$

Next, according to the assumption $|\Delta|\leq\Delta_{\max}\ll\sqrt{g_{\mathsf{th}}}$, and the fact that $|\hat{h}_k|\geq\sqrt{g_{\mathsf{th}}}$, for $k\in\mathcal{K}_i$, we can show that

$$\frac{|\Delta|}{|h_k|}\ll 1, \qquad \text{for } k\in\mathcal{K}_i. \qquad (67)$$

The proof is as follows: the condition $|\hat{h}_k|=|h_k+\Delta|\geq\sqrt{g_{\mathsf{th}}}$, for $k\in\mathcal{K}_i$ suggests that $|h_k|+|\Delta|\geq\sqrt{g_{\mathsf{th}}}$, for $k\in\mathcal{K}_i$. Then we have $\frac{|h_k|}{|\Delta|}\geq\frac{\sqrt{g_{\mathsf{th}}}}{|\Delta|}-1\geq\frac{\sqrt{g_{\mathsf{th}}}}{|\Delta_{\max}|}-1\gg 1$ as desired.

From (67), by Taylor expansion we have $\frac{1}{1+\frac{\Delta}{h_k}}=1-\frac{\Delta}{h_k}+O\left(\left(\frac{\Delta}{h_k}\right)^2\right)$. By ignoring the high order term, $\tilde{g}_i^{(n)}$ in (66) can be approximated as

$$\tilde{g}_i^{(n)}\approx\sum_{k\in\mathcal{K}_i}\sqrt{\rho_0}\tilde{g}_{k,i}^{(n)}+z_i-I_i, \qquad (69)$$

where $I_i=\sum_{k\in\mathcal{K}_i}\frac{\Delta}{h_k}\sqrt{\rho_0}\tilde{g}_{k,i}^{(n)}$ captures the error introduced by the imperfect CSI. Note that $\tilde{g}_{k,i}^{(n)}$ takes only the binary values of $+1$ and -1, and $|h_k|\geq\sqrt{g_{\mathsf{th}}}-\Delta_{\max}$, for $k\in\mathcal{K}_i$. Conditioned on K_i, we can bound the variance of I_i by

$$\mathsf{Var}(I_i)\leq\frac{\rho_0K_i\sigma_\Delta^2}{\sqrt{g_{\mathsf{th}}}-\Delta_{\max}}. \qquad (70)$$

By noting the similarity between (69) and (22), and taking I_i as an additional noise introduced by the imperfect CSI, we can apply the same machinery presented in Appendix B and the argument in the proof of Lemma 4 to derive the conditional bit error probability for the imperfect CSI case as follows. The detailed proof is skipped due to space constraint.

Lemma 7. Consider the fading channel with truncated channel inversion using imperfect CSI. The bit error probability conditioned on the size of the transmitting set K_i is given by (68) as shown on the top of the current page.

Then by the law of total probability and using the results in (53), the unconditional error probability $P_i^{\mathsf{err}}=P_i^{\mathsf{err}}(K_i=$

$$f(K,\alpha) \leq \frac{2K\alpha}{K(K+1)\alpha^2}\sum_{k=0}^{K-1}\binom{K+1}{k+2}\alpha^{k+2}(1-\alpha)^{K-k-1}, \tag{59}$$
$$= \frac{2}{(K+1)\alpha}\left[\underbrace{\overbrace{\sum_{k=-2}^{K-1}\binom{K+1}{k+2}\alpha^{k+2}(1-\alpha)^{K-k-1}}^{<1} - (1-\alpha)^{K+1} - (K+1)\alpha(1-\alpha)^K}_{=1}\right] \leq \frac{2}{(K+1)\alpha} \leq \frac{2}{K\alpha}.$$

$$\mathbb{E}\left[\left(\frac{1}{k+1}\right)^3\right] \leq \frac{6}{K(K+1)(K+2)\alpha^3}\sum_{k=0}^{K-1}\binom{K+2}{k+3}\alpha^{k+3}(1-\alpha)^{K-k-1} \tag{64}$$
$$= \frac{6\left[1-(1-\alpha)^{K+2} - (K+2)\alpha(1-\alpha)^{K+1} - \frac{(K+2)(K+1)}{2}\alpha^2(1-\alpha)^K\right]}{K(K+1)(K+2)\alpha^3} \leq \frac{6}{K(K+1)(K+2)\alpha^3} \leq \frac{6}{K^3\alpha^3}.$$

$$P_i^{\mathsf{err}}(K_i) \leq \begin{cases} \dfrac{1}{\sqrt{K_i}S_i} + \left(\dfrac{\sigma_z}{K_i\sqrt{\rho_0}} + \dfrac{2\sigma_\Delta}{\sqrt{K_i}\sqrt{\sqrt{g_{\mathsf{th}}}-\Delta_{\max}}}\right)\left(\dfrac{1}{S_i}+\dfrac{1}{2}\right), & K_i \geq 1, \\ \dfrac{1}{2}, & K_i = 0. \end{cases} \tag{68}$$

$0)\cdot\mathbb{P}(K_i=0) + P_i^{\mathsf{err}}(K_i\geq 1)\cdot\mathbb{P}(K_i\geq 1)$ is bounded by

$$P_i^{\mathsf{err}} \leq \frac{1}{2}(1-\alpha)^K + \sum_{k=1}^{K}\binom{K}{k}\alpha^k(1-\alpha)^{K-k}\times$$
$$\left[\frac{1}{\sqrt{k}S_i} + \left(\frac{\sigma_z}{k\sqrt{\rho_0}} + \frac{2\sigma_\Delta}{\sqrt{k}\sqrt{\sqrt{g_{\mathsf{th}}}-\Delta_{\max}}}\right)\left(\frac{1}{S_i}+\frac{1}{2}\right)\right]. \tag{71}$$

The above expression can be further simplified by applying Lemma 5 as presented below.

Lemma 8. Consider the fading channel with truncated channel inversion power control with imperfect CSI. The unconditional bit error probability is given by

$$P_i^{\mathsf{err}} = \frac{1}{2}(1-\alpha)^K + \frac{\sqrt{6}}{\sqrt{\alpha K}}\times$$
$$\left[\frac{1}{S_i} + \left(\frac{2}{S_i}+1\right)\frac{\sigma_\Delta}{\sqrt{\sqrt{g_{\mathsf{th}}}-\Delta_{\max}}}\right] + \frac{2}{\alpha K}\frac{\sigma_z}{\sqrt{\rho_0}}\left(\frac{1}{S_i}+\frac{1}{2}\right).$$

With Lemma 8 at hand, the desired result in Theorem 3 can be easily derived by substituting Lemma 8 into (32) and following the same machinery presented in (35)-(36).

REFERENCES

[1] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, pp. 19–25, Jan. 2020.

[2] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[3] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *IEEE Intl. Conf. on Computer Commun. (INFOCOM)*, (Honolulu, USA), pp. 63–71, 2018.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intell. and Statistics*, pp. 1273–1282, 2017.

[5] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *[Online]. Available: https://arxiv.org/abs/1610.05492*, 2017.

[6] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, pp. 491–506, Jan. 2020.

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

[8] J. Chen, X. Pan, R. Monga, S. Bengio, and R. Jozefowicz, "Revisiting distributed synchronous SGD," in *Intl. Conf. Learning Repr. (ICLR)*, (San Juan, Puerto Rico), May 2016.

[9] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient coding: Avoiding stragglers in distributed learning," in *Intl. Conf. Mach. Learning (ICML)*, (Sydney, Australia), Aug. 2017.

[10] M. Kamp, L. Adilova, J. Sicking, F. Hüger, P. Schlicht, T. Wirtz, and S. Wrobel, "Efficient decentralized deep learning by dynamic model averaging," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 393–409, Springer, 2018.

[11] T. Chen, G. B. Giannakis, T. Sun, and W. Yin, "Lag: Lazily aggregated gradient for communication-efficient distributed learning," in *Conf. Neural Info. Proc. Syst. (NIPS)*, (Montreal, CANADA), Dec. 2018.

[12] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Intl. conf. learning repr. (ICLR)*, (Vancouver, Canada), May 2018.

[13] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," in *IEEE Intl. Joint Conf. Neural Netw. (IJCNN)*, (Budapest, Hungary), Jul. 2019.

[14] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Conf. Neural Info. Proc. Syst. (NIPS)*, (Long Beach, USA), Dec. 2017.

[15] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Intl. Conf. Mach. Learning (ICML)*, vol. 80, (Stockholm, Sweden), pp. 560–569, PMLR, 2018.

[16] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Intl. conf. Artif. Intel. Stat. (AISTATS)*, (Online), Aug 2020.

[17] M. M. Amiri and D. Gunduz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, pp. 3546 – 3557, May 2020.

[18] M. M. Amiri, T. M. Duman, and D. Gunduz, "Collaborative machine

learning at the wireless edge with blind transmitters," in *IEEE Global Conf. Sig. Info. Proc. (GlobalSIP)*, (Ottawa, Canada), Nov. 2019.
[19] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans Wireless Commun.*, vol. 19, pp. 2022 – 2035, Mar. 2020.
[20] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," in *Proc. IEEE intl. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, (Barcelona, Spain), 2020.
[21] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *to appear in IEEE Trans. Wireless Commun.*, 2020.
[22] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, pp. 317 – 333, Jan. 2020.
[23] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, 2007.
[24] M. Goldenbaum and S. Stanczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Trans. Commun.*, vol. 61, pp. 3863–3877, Sep. 2013.
[25] O. Abari, H. Rahul, and D. Katabi, "Over-the-air function computation in sensor networks," *[Online]. Available: https://arxiv.org/pdf/1612.02307.pdf.*, 2016.
[26] J.-J. Xiao, S. Cui, Z.-Q. Luo, and A. J. Goldsmith, "Linear coherent decentralized estimation," *IEEE Trans. Sig. Process.*, vol. 56, no. 2, pp. 757–770, 2008.
[27] M. Goldenbaum and S. Stanczak, "On the channel estimation effort for analog computation over wireless multiple-access channels," *IEEE Wireless Commun. Lett.*, vol. 3, no. 3, pp. 261–264, 2014.
[28] G. Zhu and K. Huang, "MIMO over-the-air computation for high-mobility multi-modal sensing," *IEEE IoT Journal*, vol. 6, pp. 6089–6103, Aug. 2018.
[29] X. Li, G. Zhu, Y. Gong, and K. Huang, "Wirelessly powered data aggregation for IoT via over-the-air function computation: Beamforming and power control," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 3437–3452, Jul. 2019.
[30] D. Wen, G. Zhu, and K. Huang, "Reduced-dimension design of MIMO over-the-air computing for data aggregation in clustered IoT networks," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 5255–5268, Nov. 2019.
[31] "Timing advance (TA) in LTE." 3GPP specifications, [Online]. Available: http://4g5gworld.com/blog/timing-advance-ta-lte.
[32] G. Arunabha, J. Zhang, J. G. Andrews, and R. Muhamed, "Fundamentals of LTE," *The Prentice Hall communications engineering and emerging technologies series*, 2010.
[33] Z. A. Zhu, "Natasha2: Faster non-convex optimization than SGD," *[Online]. Available: https://arxiv.org/abs/1708.08694/*, 2017.
[34] M. Hong and Z.-Q. Luo, "Signal processing and optimal resource allocation for the interference channel," in *Academic Press Library in Signal Processing*, vol. 2, pp. 409–469, Elsevier, 2014.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conf. Comput. Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Guangxu Zhu received the B.S. and M.S. degrees from Zhejiang University and the Ph.D. degree from the University of Hong Kong, all in electronic and electrical engineering. He is now a research scientist with the Shenzhen Research Institute of Big Data. His research interests include edge intelligence, distributed machine learning, 5G technologies such as massive MIMO, mmWave communication, and wirelessly powered communication. He is a recipient of the Hong Kong Postgraduate Fellowship (HKPF), Outstanding Ph.D. Thesis Award from HKU, and a Best Paper Award from WCSP 2013. He was invited to be a co-chair for the "MAC and cross-layer design" track in IEEE PIMRC 2021.

Yuqing Du received the B.Eng. degree in Electrical Engineering from Harbin Engineering University, Harbin, China, in 2016, and the Ph.D. degree in Electronics and Electrical Engineering, The University of Hong Kong, Hong Kong, in 2020. His research interests include edge intelligence and distributed machine learning.

Deniz Gündüz (Senior Member, IEEE) received M.S. and Ph.D. degrees in electrical engineering from NYU Tandon School of Engineering (formerly Polytechnic University) in 2004 and 2007, respectively. After his PhD, he served as a postdoctoral research associate at Princeton University, and as a consulting assistant professor at Stanford University. From Sep. 2009 until Sep. 2012 he served as a research associate at CTTC in Barcelona, Spain. In Sep. 2012, he joined the Electrical and Electronic Engineering Department of Imperial College London, UK, where he is currently a Professor in Information Processing, serves as the deputy head of the Intelligent Systems and Networks Group, and leads the Information Processing and Communications Laboratory (IPC-Lab). He is also a part-time faculty member at the University of Modena and Reggio Emilia, and has held visiting positions at University of Padova (2018-2020) and Princeton University (2009-2012).

His research interests lie in the areas of communications and information theory, machine learning, and privacy. Dr. Gündüz is an Area Editor for the IEEE Transactions on Communications and the IEEE Journal on Selected Areas in Communications (JSAC) - Special Series on Machine Learning in Communications and Networks. He also serves as an Editor of the IEEE Transactions on Wireless Communications. Previously, he served as an Editor for the IEEE Transactions on Green Communications and Networking (2016-2020), and the Transactions on Communications (2013-18). He is a Distinguished Lecturer for the IEEE Information Theory Society (2020-22). He is the recipient of the IEEE Communications Society - Communication Theory Technical Committee (CTTC) Early Achievement Award in 2017, a Starting Grant of the European Research Council (ERC) in 2016, and the IEEE Communications Society Best Young Researcher Award for the Europe, Middle East, and Africa Region in 2014. He coauthored papers that received best paper awards at 2019 IEEE GlobalSIP and 2016 IEEE WCNC, and Best Student Paper Awards at 2018 IEEE WCNC and 2007 IEEE ISIT.

Kaibin Huang (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the National University of Singapore, and the Ph.D. degree from The University of Texas at Austin, all in electrical engineering. Presently, he is an associate professor in the Dept. of Electrical and Electronic Engineering at The University of Hong Kong. He received the IEEE Communication Societys 2019 Best Tutorial Paper Award, 2015 Asia Pacific Best Paper Award, and 2019 Asia Pacific Outstanding Paper Award as well as Best Paper Awards at IEEE GLOBECOM 2006 and IEEE/CIC ICCC 2018. Moreover, he received an Outstanding Teaching Award from Yonsei University in S. Korea in 2011. He has served as the lead chairs for the Wireless Comm. Symp. of IEEE Globecom 2017 and the Comm. Theory Symp. of IEEE GLOBECOM 2014 and the TPC Co-chairs for IEEE PIMRC 2017 and IEEE CTW 2013. He is an Associate Editor for IEEE Transactions on Wireless Communications and Journal on Selected Areas in Communications (JSAC), and an Area Editor for IEEE Transactions on Green Communications and Networking. Previously, he has also served on the editorial board of IEEE Wireless Communications Letters. Moreover, he has guest edited special issues for IEEE JSAC, IEEE Journal on Selected Topics on Signal Processing, and IEEE Communications Magazine. He is an IEEE Distinguished Lecturer and an ISI Highly Cited Researcher.