Blind Federated Learning at the Wireless Edge With Low-Resolution ADC and DAC

Busra Tegin, *Graduate Student Member, IEEE*, and Tolga M. Duman, *Fellow, IEEE*

***Abstract*— We study collaborative machine learning systems where a massive dataset is distributed across independent workers which compute their local gradient estimates based on their own datasets. Workers send their estimates through a multipath fading multiple access channel with orthogonal frequency division multiplexing to mitigate the frequency selectivity of the channel. We assume that there is no channel state information (CSI) at the workers, and the parameter server (PS) employs multiple antennas to align the received signals. To reduce the power consumption and the hardware costs, we employ complex-valued low-resolution digital-to-analog converters (DACs) and analog-to-digital converters (ADCs), at the transmitter and the receiver sides, respectively, and study the effects of practical low-cost DACs and ADCs on the learning performance. Our theoretical analysis shows that the impairments caused by low-resolution DACs and ADCs, including those of one-bit DACs and ADCs, do not prevent the convergence of the federated learning algorithms, and the multipath channel effects vanish when a sufficient number of antennas are used at the PS. We also validate our theoretical results via simulations, and demonstrate that using low-resolution, even one-bit, DACs and ADCs causes only a slight decrease in the learning accuracy.**

***Index Terms*— Distributed machine learning, federated learning, stochastic gradient descent, wireless channels, OFDM, low-resolution DAC and ADC, one-bit DAC and ADC.**

I. Introduction

THE rapid growth of data sensing and collection capabilities of computation devices facilitates the use of massive datasets enabling machine learning (ML) systems to make more intelligent decisions than ever. However, this growth makes the processing of all the data in a central processor troublesome due to increased energy consumption and privacy concerns. As an alternative to using a central processor, performing the ML task in a distributed manner, called federated learning, has recently drawn significant attention [1], [2]. In federated learning, each device connected to the central processor performs the required gradient computation based on its local dataset, and sends it to the central processor. The global parameter update is performed at the central processor using the local computations of the connected devices.

While federated learning can be considered as a combination of two broadly studied areas: statistical learning and communications, it also opens up new research avenues. With this motivation, different problems related to federated learning are studied in the recent literature. These include studies on the effects of energy constraints, resource allocation, privacy, compression of local computations, convergence analysis of the learning algorithms, and performance over different channel models. In particular, in [3], digital and analog distributed stochastic gradient descent (D-DSGD and A-DSGD) algorithms over a Gaussian multiple-access channel (MAC) are proposed. The authors use the superposition property of the MAC to recover the mean of the local gradients computed at remote workers. In D-DSGD, workers digitally compress their locally computed gradients into a finite number of bits, while in A-DSGD, workers use an analog compression similar to what is done in compressed sensing (CS) to obey the bandwidth limitations. In [4] and [5], the channel between the parameter server (PS) and the workers is modeled as a fading MAC. Ref. [4] performs power allocation among the gradients to schedule workers according to their channel state information (CSI). The authors show that the latency reduction of the proposed method scales linearly with the device population. Ref. [5] proposes a gradient sparsification method which is followed by a CS algorithm to reduce the dimensions of a large parameter vector. By reducing the dimensionality of the gradients and designing a power allocation scheme, the authors obtain significant performance improvements compared to the existing benchmarks.

In addition to the studies that decrease the communication load, Ref. [6] considers transmission energy, and formulates an optimization problem for the joint learning and communication process. The goal is to minimize the total energy consumption for local computations and wireless transmission under latency constraints. In [7], the authors focus on the minimization of the convergence time of a federated learning system by jointly considering user selection and resource allocation. The aim of the PS is to include as many workers as possible in the learning process for convergence to the global model with limited resources. There are also several studies on data exchange rate reduction via quantization [8]–[11]. Specifically, in [11], the authors introduce a lossy federated learning (LFL) system, which directly quantizes both the global and the local model parameters to reduce the communication loss.

Manuscript received September 25, 2020; revised March 12, 2021; accepted May 17, 2021. Date of publication June 15, 2021; date of current version December 10, 2021. Part of the material in this article is submitted for presentation in the 2021 IEEE Global Communication Conference (GLOBECOM). The associate editor coordinating the review of this article and approving it for publication was D. Li. *(Corresponding author: Tolga M. Duman.)*

Busra Tegin is with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey, and also with the Turkey R&D Center, Huawei Technologies Company Ltd., 34768 Istanbul, Turkey (e-mail: btegin@ee.bilkent.edu.tr).

Tolga M. Duman is with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: duman@ee.bilkent.edu.tr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3087594.

Digital Object Identifier 10.1109/TWC.2021.3087594

They show that the convergence of the learning algorithm is guaranteed despite the quantization process. When the training data is randomly split among the workers, LFL with a small number of quantization levels performs as well as a system with unquantized parameters. In another line of research, [12] considers a federated learning system for which there is no CSI at the workers; hence the PS employs multiple antennas to align the received signals. In [13], this study is extended further, and a convergence analysis for the blind federated learning with both perfect and imperfect CSI is performed.

While different aspects of federated learning, such as gradient compression, resource allocation, latency constraints, and fading channel effects are studied in the recent literature, the existing studies do not consider very realistic transmission models or channels. To make the use of federated learning practical, one should also consider these extensions and low-cost implementations with hardware-induced distortion for a complete system design, which is the subject of our study.

In this paper, our main objective is to study federated learning over wireless channels in realistic settings by considering practical implementation issues as well as the wireless channel effects. We model the communication link as a frequency selective fading channel, and transmit the local gradients using orthogonal frequency division multiplexing (OFDM). We consider the blind transmitter scenario, i.e., there is no CSI at the transmitters, hence multiple (even a massive number of) antennas are employed at the receiver side. Furthermore, to reduce the hardware complexity and power consumption, we employ low-resolution digital-to-analog converters (DACs) at the transmitter side (at each worker), and analog-to-digital converters (ADCs) at the receiver side. In fact, this is nothing but the over-the-air machine learning, except that here we are taking into account the effects of the wireless medium as well as the use of low-resolution DACs and ADCs. Note that while OFDM transmission with low-resolution ADCs and DACs has extensively been studied from a communication theory perspective in the literature (see, e.g., [14]–[21]), this is the first paper on their use for federated learning over wireless channels.

The main contributions of the paper can be summarized as follows:

- Different from previous works regarding federated learning reviewed above ([3]–[5], [8]–[13]), we consider a realistic wireless channel model where the channel between the workers and PS is modeled as a multipath fading MAC.
- To cope with the realistic channel impairments, we transmit the local gradients using OFDM with a cyclic prefix (CP) to mitigate the ISI caused by the multipath. Thus, different from [11], we consider the transmission and reception of actual OFDM signals as would be necessitated in a practical implementation.
- Since one of our main concerns is a practical implementation of federated learning, we also employ low-resolution DACs and ADCs separately at the workers and the PS side, respectively. Also, we extend our studies to the case of a system which utilizes both low-resolution DACs and ADCs.

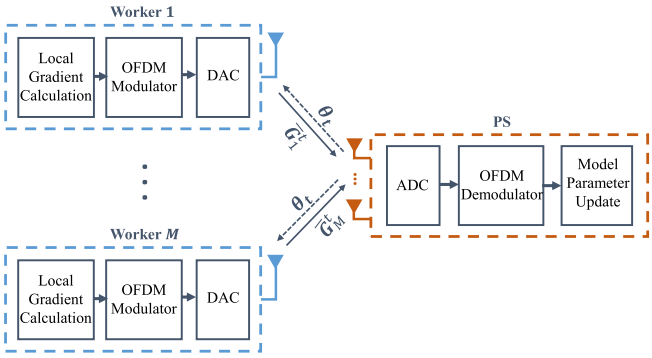


Fig. 1. System model for distributed machine learning at the wireless edge.

- Via both theoretical analysis and extensive simulations, we find that the effects of imperfections due to finite resolution DACs and/or ADCs can be alleviated using a sufficient number of receive antennas at the PS, and the convergence of the distributed learning algorithm is guaranteed even if we employ low-cost (even one-bit) DACs and/or ADCs.

The paper is organized as follows. Section II introduces the system model and preliminaries. DSGD with low-resolution DACs is analyzed in Section III, and the effect of low-resolution ADCs at the receiver side is studied in Section IV, respectively. Joint utilization of low-resolution DACs and ADCs are considered in Section V. Performance of blind federated learning with realistic channel effects and hardware limitations is studied via simulations in Section VI, and the paper is concluded in Section VII.

Notation: Throughout this paper, the real and imaginary parts of $x \in \mathbb{C}$ are represented by x^R and x^I, respectively. We use the notation $[a\ b]$ to indicate the integer set $\{a, \ldots, b\}$ where $a \leq b$, a and b are positive integers, and $[b] = [1\ b]$. We denote l_2 norm of a vector $\mathbf{x}$ by $||\mathbf{x}||_2$. The entry in the i-th row and j-th column of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}[i, j]$. N-point Discrete Fourier Transform (DFT) of vector $\mathbf{x} \in \mathbb{C}^N$ is defined as

$$\mathbf{X}[u] = \sum_{n=1}^{N} \mathbf{x}[n] e^{-j2\pi nu/N}, \tag{1}$$

while the N-point inverse discrete Fourier Transform (IDFT) of vector $\mathbf{X} \in \mathbb{C}^N$ is given by

$$\mathbf{x}[n] = \frac{1}{N} \sum_{u=1}^{N} \mathbf{X}[u] e^{j2\pi nu/N}. \tag{2}$$

II. System Model

We consider a distributed ML system where each worker calculates its gradient estimate and sends it to a central PS through a multipath fading MAC using OFDM as illustrated in Fig. 1. At the receiver side, OFDM demodulation, signal combining and global model parameter update are performed. The global parameter is broadcast to the workers over an error-free link. We assume that there is no transmit side CSI, and that the PS employs multiple antennas to recover the

average of the workers' gradients. With the use of a higher number of workers and many antennas, a significant amount of power at the transmitter and receiver is consumed by the DACs and ADCs [22]. The power consumption of DACs and ADCs increases linearly, and their hardware cost increases exponentially with the number of quantization bits [23]. In order to keep the implementation cost and power consumption low, we consider a distributed learning system where the transmitters and receivers are equipped with low-resolution, even one-bit, DACs and ADCs, respectively.

We jointly train a learning model by using iterative stochastic gradient descent (SGD) to minimize a loss function $f(\cdot)$. During the t-th iteration, worker $m \in [M]$ calculates the gradient estimate $\mathbf{g}_m^t \in \mathbb{R}^d$ by processing its local dataset $\mathcal{B}_m$ according to $\frac{1}{|\mathcal{B}_m|}\sum_{u\in\mathcal{B}_m} \nabla f(\boldsymbol{\theta}_t, u)$ where $\boldsymbol{\theta}_t \in \mathbb{R}^d$ is the vector of model parameters, d is the number of model parameters, and $g_m^t[n]$ represents the n-th entry of the gradient estimate. We form the baseband frequency domain signal of the local gradient vector as

$$\hat{\mathbf{g}}_m^t = \big[g_m^t[1] + jg_m^t[s+1], g_m^t[2] + jg_m^t[s+2], \cdots, g_m^t[s] + jg_m^t[2s]\big], \quad (3)$$

where $s = \lceil d/2 \rceil$, $\hat{\mathbf{g}}_m^t \in \mathbb{R}^s$, and $g_m^t[2s]$ is assigned as zero if $d \equiv 1 \pmod 2$. Then, the first step is to form the OFDM signal by taking an N-point IDFT of the gradient vector as

$$G_m^t[u] = \frac{1}{N}\sum_{n=1}^{N} \hat{g}_m^t[n] e^{j2\pi nu/N}, \quad (4)$$

for $u \in [N]$. If $s < N$, $\hat{g}_m^t[n] = 0$ for $n > s$, i.e., $\hat{\mathbf{g}}_m^t$ is zero padded.

The channel between the m-th worker and the k-th antenna of the PS is modeled as a (wireless) multipath MAC. We assume that the channel does not change during the transmission of one OFDM word, while it may be different for different OFDM words. The impulse response of the channel is

$$h_{mk}^t[n] = \sum_{l=1}^{L} h_{mkl}^t \delta[n - \tau_{mkl}], \quad (5)$$

where $n \in [N+N_{cp}]$, L is the number of channel taps, τ_{mkl} is the time delay and $h_{mkl}^t \in \mathbb{C}$ is the gain of the l-th channel tap from the m-th worker to the k-th antenna of the PS. Note that this is nothing but the machine learning over-the-air framework of [12]. We assume that h_{mkl}^t are zero-mean (circularly symmetric) complex Gaussian with $\mathbb{E}\left[(h_{mkl}^t)\cdot(h_{m'k'l'}^t)^*\right] = 0$ for $(m,k,l) \neq (m',k',l')$, and $\mathbb{E}\left[|h_{mkl}^t|^2\right] = \sigma_{h,l}^2$, i.e., all the channel taps experience Rayleigh fading.

To mitigate the ISI caused by the multipath channel, CP addition is performed, i.e.,

$$\bar{\mathbf{G}}_m^t = \big[G_m^t[N - N_{cp} + 1] \ldots G_m^t[N]\ G_m^t[1] \ldots G_m^t[N]\big], \quad (6)$$

where $\bar{\mathbf{G}}_m^t \in \mathbb{C}^{N+N_{cp}}$ is the OFDM word to be transmitted by the m-th worker. The CP length N_{cp} is chosen to be greater than the delay spread of all the channels. The resulting (depending on the setup – quantized or full resolution) OFDM words are transmitted to the PS which are equipped with K receive antennas. The PS uses the received signal to update the model and sends it back to all the receivers over an error-free link.

At the k-th receive chain, after removing the CP, the n-th entry of the received vector at the input of the k-th receive antenna during iteration t is written as

$$Y_k^t[n] = \sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}^t G_m^t[n - \tau_{mkl}] + z_k^t[n], \quad (7)$$

where the additive noise terms $z_k^t[n]$ are independent and identically distributed (i.i.d.) circularly symmetric zero mean complex Gaussian random variables, i.e., $z_k^t[n] \sim \mathcal{CN}(0, \sigma_z^2)$ for $k \in [K]$.

Ideally, the PS updates the model parameter according to $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mu_t \frac{1}{M}\sum_{m=1}^{M} \mathbf{g}_m^t$, and it is shared with the workers. However, in our setup, the local gradients are not available at the PS, instead the PS uses noisy and distorted version (by low-resolution DACs and/or ADCs) of the local gradients to recover the estimate of the gradient vector as will become apparent in the subsequent sections. In the following, we drop the subscripts referring to iteration index t for ease of exposition.

III. DSGD With Low-Resolution DACs at the Workers

In this section, we study the effects of employing low-resolution DACs at the workers on the distributed learning process in an effort to reduce the hardware complexity and power consumption.

After constructing the OFDM word corresponding to the gradient vectors, a complex-valued low-resolution DAC is employed to generate the transmitted signal at each worker. A b-bit complex-valued DAC consists of two parallel real-valued DACs with quantization function $Q_b(\cdot)$. The real and imaginary parts are separately quantized into $\beta = 2^b$ reconstruction levels. The reconstruction levels are denoted by $\hat{\mathbf{a}} = [\hat{a}_1\ \hat{a}_2 \cdots \hat{a}_\beta] \in \mathbb{R}^\beta$ while the boundaries of the quantization regions are denoted by $\hat{\mathbf{x}} = [\hat{x}_1\ \hat{x}_2 \cdots \hat{x}_{\beta+1}] \in \mathbb{R}^{\beta+1}$ where $\hat{x}_1 = -\infty$ and $\hat{x}_{\beta+1} = +\infty$ for convenience. Also, we have, $\hat{a}_i < \hat{a}_j$, if $1 \le i < j \le \beta$, $\hat{x}_i < \hat{x}_j$ if $1 \le i < j \le \beta+1$, and $\hat{x}_i \le \hat{a}_j < \hat{x}_k$ if $1 \le i \le j < k \le \beta+1$. The corresponding real valued quantizer is $Q_b(z) = \hat{a}_i$ for $\hat{x}_i \le z < \hat{x}_{i+1}$, $i \in [\beta]$, $z \in \mathbb{R}$. The complex-valued DAC operation can be expressed as $Q_b(x) = Q_b(x^R) + jQ_b(x^I)$. We assume that the quantizer output is chosen such that $Q_b(x) = \mathbb{E}[X|Q_b(X)]$, i.e., the reconstruction level is selected to minimize the mean squared error for each quantization region. The corresponding signal to quantization noise ratio (SQNR) of the input vector $\mathbf{x}$ is calculated as

$$\text{SQNR} = \frac{\mathbb{E}\left[|X|^2\right]}{\mathbb{E}\left[|Q_b(X) - X|^2\right]}. \quad (8)$$

We model the OFDM words as wide-sense stationary (WSS) Gaussian processes based on an argument similar to the one made in [24]. That is, if the input data which forms the OFDM word is i.i.d. and bounded, the convex envelope of the OFDM word weakly converges to a Gaussian random process as the number of subcarriers goes to infinity through an application

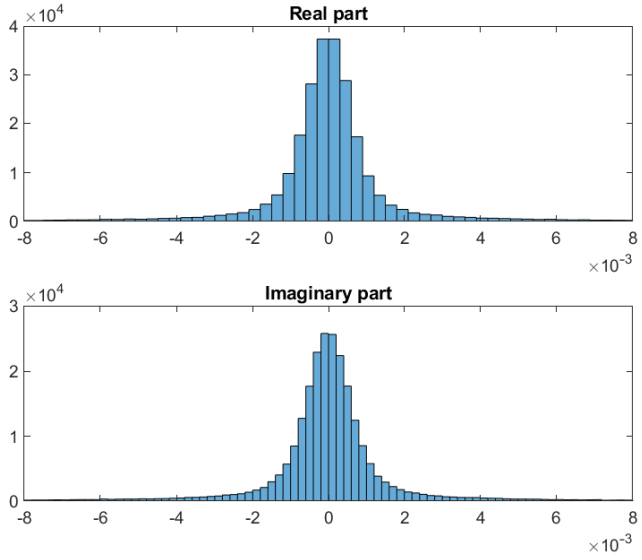


Fig. 2. Histogram of the real and imaginary parts of an exemplary OFDM word during the learning task with our setup.

of central limit theorem (CLT). Similarly, if we assume that the elements of the gradient vector in the learning process are i.i.d. and bounded, then the real and imaginary parts of the baseband OFDM word obtained from the gradient vector can be modeled as independent zero-mean stationary Gaussian processes. As a verification, we examine histograms of several OFDM word samples obtained by a certain learning task with our setup. An instance of an exemplary histogram of the OFDM word samples obtained through the 100-th iteration is given in Fig. 2 which is consistent with our assumption. Our extensive experiments further confirm that the corresponding OFDM word samples at different time indexes have almost the same variance. Note that, even if the OFDM words are not Gaussian processes, the Bussgang theorem that will be used to model the nonlinear input-output relationship for DACs and ADCs is still a good approximation as illustrated extensively in the literature, see, e.g., [25], [26].

We denote the autocorrelation matrix of the OFDM words by $\mathbf{C}_{\bar{\mathbf{G}}_m\bar{\mathbf{G}}_m}$ with equal diagonal elements denoted by $\sigma^2_{G_m}$. Using the Bussgang decomposition [29], [30], we can write the quantized signal in two parts: the desired signal component and the quantization distortion which is uncorrelated with the desired signal, that is,

$$\bar{G}^Q_m[n] = Q(\bar{G}_m[n]) = (1-\eta)\bar{G}_m[n] + q_m[n], \tag{9}$$

where $\eta = 1/\text{SQNR}$ is the distortion factor which is the inverse of SQNR, and the variance of the distortion noise is $\sigma^2_{q_m} = \eta(1-\eta)\sigma^2_{G_m}$. When a unit variance Gaussian input is processed by a non-uniform scalar minimum mean-square-error quantizer, the values of corresponding distortion factors are listed in Table I [27], [28].

At the k-th receive chain, after removing the CP, the n-th entry of the received vector is written as

$$Y_k[n] = \sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}G^Q_m[n-\tau_{mkl}] + z_k[n] \tag{10}$$

TABLE I
DISTORTION FACTORS WITH DIFFERENT QUANTIZATION LEVELS [27], [28]

Number of bits	Distortion factor (η)
1	0.3634
2	0.1175
3	0.03454
4	0.009497
5	0.002499

$$= \sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}\Big((1-\eta)\cdot G_m[n-\tau_{mkl}] + q_m[n-\tau_{mkl}]\Big) + z_k[n] \tag{11}$$

$$= (1-\eta)\sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}G_m[n-\tau_{mkl}] + w_k[n], \tag{12}$$

where the total non-Gaussian noise term $w_k[n]$ has variance $\sigma^2_z + \eta(1-\eta)\sigma^2_{G_m}\sum_{m=1}^{M}\sum_{l=1}^{L}|h_{mkl}|^2$.

To perform the demodulation, we take the DFT of (10) which gives

$$r_k[i] = (1-\eta)\sum_{m=1}^{M} H_{mk}[i]g_m[i] + \sum_{m=1}^{M} H_{mk}[i]Q_m[i] + Z_k[i], \tag{13}$$

where $Q_m[i]$ is the DFT of the quantization distortion noise and $H_{mk}[i]$'s are the channel gains from the m-th worker to the k-th receive chain for the i-th subcarrier. $H_{mk}[i]$'s are given by

$$\begin{aligned} H_{mk}[i] &= \sum_{n=0}^{N-1} h_{mk}[n]e^{-j2\pi in/N} \\ &= \sum_{n=0}^{N-1}\left(\sum_{l=1}^{L} h_{mkl}\delta[n-\tau_{mkl}]\right)e^{-j2\pi in/N} \\ &= \sum_{l=1}^{L} h_{mkl}e^{-j2\pi i\tau_{mkl}/N}. \end{aligned} \tag{14}$$

Since the channel taps are zero mean circularly symmetric complex Gaussian (i.e., Rayleigh fading), $H_{mk}[i]$'s are also zero-mean complex Gaussian random variables with variance $\sigma^2_H = \sum_{l=1}^{L}\sigma^2_{h,l}$.

Taking the DFT of the channel noise vector, $Z_k[i]$ is evaluated as

$$Z_k[i] = \sum_{n=0}^{N-1} z_k[n]e^{-j2\pi in/N}. \tag{15}$$

The noise terms are i.i.d. circularly symmetric complex Gaussian, i.e., $Z_k[n] \sim \mathcal{CN}(0, \sigma^2_{Z_k})$ where $\sigma^2_{Z_k} = N\sigma^2_{z_k}$.

We assume that the CSI is available at the PS, hence the received signals from the K antennas can be combined to align

the gradient vectors using

$$y[i] = \frac{1}{(1-\eta)\cdot K}\sum_{k=1}^{K}\Big(\sum_{m=1}^{M}(H_{mk}[i])^*\Big)r_k[i], \tag{16}$$

as in [12], [13]. By substituting (13) into (16), we obtain

$$y[i] = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}|H_{mk}[i]|^2 g_m[i]}_{\text{signal term}} \tag{17a}$$

$$+\underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}(H_{mk}[i])^* H_{m'k}[i]g_{m'}[i]}_{\text{interference term}} \tag{17b}$$

$$+\underbrace{\frac{1}{(1-\eta)K}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}(H_{mk}[i])^* H_{m'k}[i]Q_{m'}[i]}_{\text{distortion noise term}} \tag{17c}$$

$$+\underbrace{\frac{1}{(1-\eta)K}\sum_{k=1}^{K}\sum_{m=1}^{M}|H_{mk}[i]|^2 Q_m[i]}_{\text{second type of distortion noise term}} \tag{17d}$$

$$+\underbrace{\frac{1}{(1-\eta)K}\sum_{k=1}^{K}\Big(\sum_{m=1}^{M}(H_{mk}[i])^*\Big)Z_k[i]}_{\text{channel noise term}}. \tag{17e}$$

There are five different terms in (17): the signal component, interference, distortion noise term, the second type of distortion noise term, and the channel noise.

To analyze the interference term (17b), we write it as a summation of M terms

$$\frac{1}{K}\Bigg[\Big(\sum_{k=1}^{K}\sum_{m=2}^{M}(H_{mk}[i])^* H_{1k}[i]\Big)g_1[i] + \cdots + \Big(\sum_{k=1}^{K}\sum_{\substack{m=1\\m\neq j}}^{M}(H_{mk}[i])^* H_{jk}[i]\Big)g_j[i] + \cdots + \Big(\sum_{k=1}^{K}\sum_{m=1}^{M-1}(H_{mk}[i])^* H_{Mk}[i]\Big)g_M[i]\Bigg], \tag{18}$$

and consider the coefficient of each term $g_j[i]$ separately. Let us define

$$\kappa_j[i] = \frac{1}{K}\sum_{k=1}^{K}\sum_{\substack{m=1\\m\neq j}}^{M}(H_{mk}[i])^* H_{jk}[i], \tag{19}$$

for the coefficient of the j-th interfering gradient $g_j[i]$ in (17b) where $i \in [N]$, and $j \in [M]$. Since $H_{mk}[i]$ and $H_{jk}[i]$ are independent for $j \neq m$, the mean and variance of $\kappa_j[i]$ are calculated as

$$\mathbb{E}\left[\kappa_j[i]\right] = 0, \tag{20a}$$

$$\mathbb{E}\left[|\kappa_j[i]|^2\right] = \frac{(M-1)\sigma_H^4}{K}. \tag{20b}$$

We have M such interference terms in (17b) each for a different worker with zero mean, and variance scaling with $\frac{M-1}{K}$. Hence, the total interference term approaches zero as $K \to \infty$.

To analyze the distortion noise term (17c), we define the coefficient of each uncorrelated distortion term $Q_j[i]$ separately as in the case of (17b) by

$$\delta_{1j}[i] = \frac{1}{(1-\eta)K}\sum_{k=1}^{K}\sum_{\substack{m=1\\m\neq j}}^{M}(H_{mk}[i])^* H_{jk}[i], \tag{21}$$

where $i \in [N]$, and $j \in [M]$.

Similar to the analysis of $\kappa_j[i]$, the mean and variance of $\delta_{1j}[i]$ are calculated as

$$\mathbb{E}\left[\delta_{1j}[i]\right] = 0, \tag{22a}$$

$$\mathbb{E}\left[|\delta_{1j}[i]|^2\right] = \frac{(M-1)\sigma_H^4}{(1-\eta)^2 K}. \tag{22b}$$

This implies that each of the M interfering terms in (17c) goes to zero if K is large enough. Thus, the detrimental effect of the distortion noise term can also be eliminated by employing a large number of receive antennas.

To analyze the second type of distortion noise term (17d), we consider each term $Q_j[i]$ separately for $j \in [M]$, and define the coefficient of the interfering distortion term caused by the j-th one as

$$\delta_{2j}[i] = \frac{1}{(1-\eta)K}\sum_{k=1}^{K}|H_{jk}[i]|^2, \tag{23}$$

where $i \in [N]$, and $j \in [M]$. The mean of $\delta_{2j}[i]$ is

$$\mathbb{E}\left[\delta_{2j}[i]\right] = \frac{\sigma_H^2}{(1-\eta)}. \tag{24}$$

For the variance of $\delta_{2j}[i]$, we have

$$\mathbb{E}\left[\left|\delta_{2j}[i]\right|^2\right] = \frac{1}{(1-\eta)^2 K^2} \cdot \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\mathbb{E}\left[\left|H_{jk_1}[i]\right|^2\left|H_{jk_2}[i]\right|^2\right]. \tag{25}$$

- If $k_1 = k_2$ (case 2.1)

$$\mathbb{E}\left[|\delta_{2j}[i]|^2\right]\big|_{\text{case 2.1}} = \frac{1}{(1-\eta)^2 K^2}\sum_{k=1}^{K}\mathbb{E}\left[|H_{jk}[i]|^4\right] \tag{26}$$

$$= \frac{1}{(1-\eta)^2 K}\mathbb{E}\left[|H_{jk}[i]|^4\right]. \tag{27}$$

- If $k_1 \neq k_2$ (case 2.2)

$$\mathbb{E}\left[|\delta_{2j}[i]|^2\right]\big|_{\text{case 2.2}} = \frac{1}{(1-\eta)^2 K^2}\sum_{k_1=1}^{K}\sum_{\substack{k_2=1\\k_2\neq k_1}}^{K}\mathbb{E}\left[|H_{jk_1}[i]|^2\right]\mathbb{E}\left[|H_{jk_2}[i]|^2\right] \tag{28}$$

$$= \frac{(K^2-K)\sigma_H^4}{(1-\eta)^2 K^2} \tag{29}$$

$$\approx \frac{\sigma_H^4}{(1-\eta)^2}, \tag{30}$$

for $K \gg 1$. Thus, the mean and variance of the second distortion term of the j-th worker is calculated as

$$\mathbb{E}\left[\delta_{2j}[i]\right] = \frac{\sigma_H^2}{(1-\eta)}, \tag{31a}$$

$$\text{Var}(\delta_{2j}[i]) \approx \frac{1}{(1-\eta)^2 K}\mathbb{E}\left[|H_{jk}[i]|^4\right]. \tag{31b}$$

Note that $\delta_{2j}[i]$ has a finite mean and its variance approaches zero as $K \to \infty$. We know that the mean of the distortion term, $Q_j[i]$ for all $j \in [M]$, is zero. Accordingly, using the law of large numbers, the summation will converge to the mean of $Q_j[i]$, which is zero, for a sufficiently large M.

Using the law of large numbers, as the number of antennas at the PS $K \to \infty$, the signal term can be approximated as

$$y_{\text{sig}}[i] = \sigma_H^2 \sum_{m=1}^{M} g_m[i]. \tag{32}$$

Thus, with low-resolution DACs at the workers, the PS can recover the i-th entry of the desired signal using

$$\frac{1}{M}\sum_{m=1}^{M} g_m[i] = \begin{cases} \dfrac{y^R[i]}{M\sigma_H^2}, & \text{if } 1 \le i \le s, \\ \dfrac{y^I[i-s]}{M\sigma_H^2}, & \text{if } s < \ i \le 2s. \end{cases} \tag{33}$$

This result clearly shows that the destructive effect of low-resolution DACs can be effectively alleviated using a sufficient number of PS antennas. Thus, the convergence of the learning process is guaranteed even if we employ low-cost low-resolution DACs at the workers, which significantly reduces the cost of designing distributed learning systems with a high number of workers. On the other hand, using a very large number of PS antennas will increase both the design cost and energy consumption, hence it may not be efficient. For further assessment, we can consider the coefficients of the distortion terms. For the distortion noise term given in (17c), we have M contributing terms each with zero mean and variance $\frac{(M-1)\sigma_H^4}{(1-\eta)^2 K}$. To reduce the effects of these terms on the learning accuracy, it is desired to have this variance close to zero. Clearly, this variance depends on several parameters; hence, to evaluate the overall performance, we should not only consider the number of receive antennas K, but also the channel variance σ_H^2, number of workers M, and distortion factor $\eta \in [0,\ 1]$. For example, if we have a high-resolution DAC, η will be small; hence, using a smaller number of receive antennas may be sufficient to cancel out the resulting impairments. However, when the resolution is very low, e.g., for a one-bit DAC, η will be large, and we will need a higher number of receive antennas due to the $\frac{1}{(1-\eta)^2}$ term. A similar approach can also be used to analyze the second type of distortion noise term given in (17d) for which we have M contributing terms each with variance $\frac{1}{(1-\eta)^2 K}\mathbb{E}\left[|H_{jk}[i]|^4\right]$. In other words, there is a trade-off between the DAC resolution and the number of receive antennas, and the overall performance is also affected by the channel statistics.

IV. DSGD With Low-Resolution ADCs at the PS

In this section, we consider a system where the workers transmit the OFDM words corresponding to the local gradients with full-resolution through a multipath fading channel while the PS employs low-resolution ADCs at each receive antenna, and analyze the convergence of the federated learning algorithm.

At each receive chain, after removing the CP, the n-th entry of the received OFDM word $\mathbf{Y}_k$ is

$$Y_k[n] = \sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl} G_m[n-\tau_{mkl}] + z_k[n]. \tag{34}$$

The $(k,\ k')$-th element of the auto-correlation matrix of $\mathbf{Y}[n] = [Y_1[n] \cdots Y_K[n]]$ received by different antennas can be written as

$$\mathbf{C}_{\mathbf{YY}}[k,k'] = \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{m'=1}^{M}\sum_{l=1}^{L}\sum_{l'=1}^{L} h_{mkl} h^*_{m'k'l'} G_m[n-\tau_{mkl}] \cdot G^*_{m'}[n-\tau_{m'k'l'}]\Big] + \sigma_z^2 \mathbb{1}_{\{k=k'\}} \tag{35}$$

$$= \sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{l'=1}^{L} h_{mkl} h^*_{mk'l'} \mathbb{E}\big[G_m[n-\tau_{mkl}] \cdot G_m[n-\tau_{mk'l'}]\big] + \sigma_z^2 \mathbb{1}_{\{k=k'\}}. \tag{36}$$

The variance of the received signal at the k-th antenna $Y_k[n]$ is given by

$$\sigma^2_{Y_k} = \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{m'=1}^{M}\sum_{l=1}^{L}\sum_{l'=1}^{L} h_{mkl} h^*_{m'kl'} \cdot G_m[n-\tau_{mkl}] G^*_{m'}[n-\tau_{m'kl'}]\Big] + \sigma_z^2 \tag{37}$$

$$= \sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{l'=1}^{L} h_{mkl} h^*_{mkl'} \cdot \mathbb{E}\left[G_m[n-\tau_{mkl}] G^*_m[n-\tau_{mkl'}]\right] + \sigma_z^2, \tag{38}$$

which only depends on k.

A complex-valued low-resolution ADC employed at each receive antenna performs quantization. As in the case with low-resolution DACs described in the previous section, we describe b-bit quantization with quantization function $Q_b(\cdot)$ that independently quantizes the real and imaginary parts into $\beta = 2^b$ reconstruction levels such that the quantizer output is chosen as $Q_b(x) = \mathbb{E}[X|Q_b(X)]$.

With element-wise quantization, we can decompose the quantized signal into two parts as the desired signal component and quantization distortion which is uncorrelated with the desired signal. Analytically, we can write the quantized signal as

$$R_k[n] = (1-\eta_k)\Big(\sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl} \cdot G_m[n-\tau_{mkl}] + z_k[n]\Big) + w_q^k[n], \tag{39}$$

where η_k is the distortion factor which is the inverse of the SQNR due to quantization of $\mathbf{Y}_k$. To determine η_k, one can

use Table I. $w_q^k[n]$ is a non-Gaussian distortion noise at the k-th antenna whose variance is $\sigma^2_{w_q^k} = \eta_k(1-\eta_k)\sigma^2_{Y_k}$.

The receive antennas at the PS are equipped with identical ADCs. As explained in [30], while it may be tempting to think that the quantization noise terms at different ADCs are uncorrelated, this is generally not the case since each antenna receives different (delayed) linear combinations of the same set of OFDM words generated at the workers. On the other hand, as shown in [31], the distortion can be safely approximated as uncorrelated for massive MIMO systems with a sufficient number of users. We have also validated this approximation for our system, and observed that the correlation across the antennas of the PS is near-zero, even for the one-bit ADC case. Therefore, the correlations can be ignored as in the additive quantization noise model (AQNM), leading to a tractable scheme [32]. We further note that there are different studies on low-resolution ADCs which also neglect the distortion correlation among antennas as in our approach [27], [33], [34]. For zero-mean Gaussian processes, this approach is equivalent to the Bussgang decomposition, except that it ignores the correlation among the elements of the distortion term.

If we define the total effective noise due to the channel and the quantization process as

$$w_k[n] = (1-\eta_k)z_k[n] + w_q^k[n], \tag{40}$$

the outputs of the complex ADCs can be written as

$$R_k[n] = (1-\eta_k)\sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}G_m[n-\tau_{mkl}] + w_k[n], \tag{41}$$

where $w_k[n]$ is non-Gaussian total noise with variance $\sigma^2_{w_k} = \sigma^2_{w_q^k} + (1-\eta_k)^2\sigma_z^2$, and it is assumed to be uncorrelated across the antennas.

To perform the OFDM demodulation, we take the DFT of (41) which results in

$$r_k[i] = (1-\eta_k)\sum_{m=1}^{M} H_{mk}[i]g_m[i] + W_k[i], \tag{42}$$

where $H_{mk}[i]$'s are the channel gains from the m-th worker to the k-th receive chain for the i-th subcarrier, given by (14), which are zero-mean Gaussian random variables with variance $\sigma_H^2 = \sum_{l=1}^{L}\sigma_{h,l}^2$.

Taking the DFT of the effective noise, $W_k[i]$ is given as

$$W_k[i] = \sum_{n=0}^{N-1} w_k[n]e^{-j2\pi in/N}. \tag{43}$$

We know that the channel noise is i.i.d., and we assume that the distortion noise decorrelates sufficiently fast. Hence, $W_k[i]$ converges absolutely to a Gaussian random variable by an application of the central limit theorem (CLT) [35], i.e., $W_k[n] \sim \mathcal{CN}(0, \sigma^2_{W_k})$ where $\sigma^2_{W_k} = N\sigma^2_{w_k}$.

Assuming that the CSI is available at the PS as in the previous section, the received signals from the K antennas can be combined to align the gradient vectors by

$$y[i] = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{1-\eta_k}\left(\sum_{m=1}^{M}(H_{mk}[i])^*\right)r_k[i]. \tag{44}$$

By substituting (42) into (44), we obtain

$$y[i] = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}|H_{mk}[i]|^2 g_m[i]}_{\text{signal term}} \tag{45a}$$

$$+ \underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{m'=1,m'\neq m}^{M}(H_{mk}[i])^* H_{m'k}[i]g_{m'}[i]}_{\text{interference term}} \tag{45b}$$

$$+ \underbrace{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{1-\eta_k}\left(\sum_{m=1}^{M}(H_{mk}[i])^*\right)W_k[i]}_{\text{noise term}}. \tag{45c}$$

There are three different terms in (45): the signal component, the interference and the noise. Using the law of large numbers, as the number of antennas at the PS $K \to \infty$, the signal term approaches

$$y_{\text{sig}}[i] = \sigma_H^2\sum_{m=1}^{M} g_m[i]. \tag{46}$$

Thus, the PS can recover the i-th entry of the desired signal

$$\frac{1}{M}\sum_{m=1}^{M} g_m[i] = \frac{y_{\text{sig}}[i]}{M\sigma_H^2}. \tag{47}$$

To analyze the interference term (45b), we follow the same approach as in the previous section where each of the M interfering terms is analyzed separately. We define the term due to the j-th interfering worker as

$$\kappa_j[i] = \frac{1}{K}\sum_{k=1}^{K}\sum_{\substack{m=1\\ m\neq j}}^{M}(H_{mk}[i])^* H_{jk}[i], \tag{48}$$

where $i \in [N]$, and $j \in [M]$. Since $H_{mk}[i]$ and $H_{jk}[i]$ are independent for $j \neq m$, the mean and variance of $\kappa_j[i]$ are calculated as

$$\mathbb{E}\left[\kappa_j[i]\right] = 0, \tag{49a}$$

$$\mathbb{E}\left[|\kappa_j[i]|^2\right] = \frac{(M-1)\sigma_H^4}{K}. \tag{49b}$$

Accordingly, for fixed gradient values, each of the M interference terms in (45b) has zero mean and their variances scale with $\frac{M-1}{K}$. Thus, similar to the ideal case (where the receive chains are equipped with infinite resolution ADCs as considered in [12]), the interference term approaches zero as $K \to \infty$. In other words, using a sufficiently large number of antennas at the PS eliminates the destructive effects of the interference on the learning process, and the estimate for the gradient vector is obtained as

$$\frac{1}{M}\sum_{m=1}^{M} g_m[i] = \begin{cases} \dfrac{y^R[i]}{M\sigma_H^2}, & \text{if } 1 \le i \le s, \\ \dfrac{y^I[i-s]}{M\sigma_H^2}, & \text{if } s < i \le 2s, \end{cases} \tag{50}$$

for $i \in [d]$. This result clearly shows that the convergence of the learning process is guaranteed even if we employ low-cost low-resolution ADCs at the receiver.

V. DSGD With Low-Resolution DACs and ADCs

We now consider a system where the workers and the PS employ low-resolution DACs and ADCs, respectively. Each worker uses a finite resolution DAC to quantize the OFDM words, and transmits them through a multipath fading channel. The PS receives the signal from multiple antennas where finite resolution ADCs are employed at each receive chain. The aim is to obtain an estimate of the gradients using the received signals, which are distorted by ADCs and DACs as well as the multipath fading channel impairments. We analyze the impact of employing finite resolution ADCs and DACs jointly on the convergence of the learning algorithm. We accomplish this by using the Bussgang decomposition and AQNM model for the quantization operation at the workers and the PS, respectively.

Each worker calculates their local gradients and their corresponding OFDM words $\bar{\mathbf{G}}_m \in \mathbb{C}^{N+N_{cp}}$. As in Section III, each worker uses a finite resolution DAC, and quantizes the OFDM words corresponding to the local gradients. The n-th element of the transmitted signal by the m-th worker is

$$\bar{G}_m^Q[n] = Q(\bar{G}_m[n]) = (1-\eta)\bar{G}_m[n] + q_m[n] \tag{51}$$

using the Bussgang decomposition. Here $\eta = 1/\text{SQNR}$ due to the quantization of $\bar{G}_m[n]$, and the variance of the distortion noise is $\sigma^2_{q_m} = \eta(1-\eta)\sigma^2_{G_m}$.

The quantized signals pass through a multipath fading channel whose impulse response is given in (5). After removing the CP, the received signal at the input of the finite resolution ADC of the k-th antenna of the PS is

$$U_k[n] = \sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}\Big((1-\eta)G_m[n-\tau_{mkl}] + q_m[n-\tau_{mkl}]\Big) + z_k[n]. \tag{52}$$

The mean of $U_k[n]$ is zero, and its variance is given by

$$\begin{aligned}\sigma^2_{U_k} &= \sum_{m=1}^{M}\sum_{l=1}^{L} |h_{mkl}|^2 \left((1-\eta)^2 + \eta(1-\eta)\right)\sigma^2_{G_m} \\ &\quad + (1-\eta)^2 \sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{l'=1,l'\neq l}^{L} h_{mkl}h^*_{mkl'} \\ &\quad \cdot \mathbb{E}\Big[G_m[n-\tau_{mkl}]G_m[n-\tau_{mkl'}]\Big] + \sigma_z^2,\end{aligned} \tag{53}$$

which only depends on the receive antenna index k.

The PS employs finite resolution ADCs at each receive antenna. The quantization operation of the ADC can be modeled as a linear operation using an AQNM model where the correlation of distortion noise across the antennas is ignored. The corresponding quantized signal at the k-th antenna is written as

$$\begin{aligned}R_k[n] = (1-\eta_k)\Bigg(&\sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}(1-\eta)G_m[n-\tau_{mkl}] \\ &+ \sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}q_m[n-\tau_{mkl}] + z_k[n]\Bigg) + v_q[n],\end{aligned} \tag{54}$$

where η_k is the distortion factor due to quantization of the received signal at the k-th antenna ($\mathbf{U}_k$), and calculated through the SQNR of the corresponding quantization operation as $\eta_k = 1/\text{SQNR}$. $v_q[n]$ is a non-Gaussian distortion noise whose variance is $\sigma^2_{v_q} = \eta_k(1-\eta_k)\sigma^2_{U_k}$.

The total effective non-Gaussian noise due to the channel and quantization with ADC at the PS is

$$p_k[n] = (1-\eta_k)z_k[n] + v_q[n], \tag{55}$$

with variance $\sigma^2_{p_k} = (1-\eta_k)^2\sigma_z^2 + \sigma^2_{v_q}$, and the output of the complex ADC can be rewritten as

$$\begin{aligned}R_k[n] = (1-\eta_k)(1-\eta)&\sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}G_m[n-\tau_{mkl}] \\ &+ (1-\eta_k)\sum_{m=1}^{M}\sum_{l=1}^{L} h_{mkl}q_m[n-\tau_{mkl}] + p_k[n].\end{aligned} \tag{56}$$

For demodulation, we take the DFT of (56), which results in

$$\begin{aligned}r_k[i] = (1-\eta_k)(1-\eta)&\sum_{m=1}^{M} H_{mk}[i]g_m[i] \\ &+ (1-\eta_k)\sum_{m=1}^{M} H_{mk}[i]Q_m[i] + P_k[i],\end{aligned} \tag{57}$$

where $H_{mk}[i]$'s are as defined in (14), and $Q_m[i]$ is the DFT of the quantization distortion noise.

Taking the DFT of the effective noise, $P_k[i]$ is evaluated as $P_k[i] = \sum_{n=0}^{N-1} p_k[n]e^{-j2\pi in/N}$. With a similar approach to the one used in Section IV, under some mild assumptions, $P_k[i]$ converges absolutely to a Gaussian random variable by an application of CLT [35].

Since the CSI is only available at the PS as in [12], the received signals can be combined to align the gradient vectors as

$$y[i] = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{(1-\eta)(1-\eta_k)}\left(\sum_{m=1}^{M} H_{mk}[i]\right)^* r_k[i]. \tag{58}$$

This quantity can be written as the sum of five different terms as in Section III:

$$y[i] = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}|H_{mk}[i]|^2 g_m[i]}_{\text{signal term}} \tag{59a}$$

$$+ \underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}(H_{mk}[i])^* H_{m'k}[i]g_{m'}[i]}_{\text{interference term}} \tag{59b}$$

$$+ \underbrace{\frac{1}{(1-\eta)K}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}(H_{mk}[i])^* H_{m'k}[i]Q_{m'}[i]}_{\text{distortion noise term}} \tag{59c}$$

$$+\underbrace{\frac{1}{(1-\eta)K}\sum_{k=1}^{K}\sum_{m=1}^{M}|H_{mk}[i]|^2 Q_m[i]}_{\text{second type of distortion noise term}} \tag{59d}$$

$$+\underbrace{\frac{1}{(1-\eta)K}\sum_{k=1}^{K}\frac{1}{(1-\eta_k)}\left(\sum_{m=1}^{M}(H_{mk}[i])^*\right)P_k[i]}_{\text{noise term}}, \tag{59e}$$

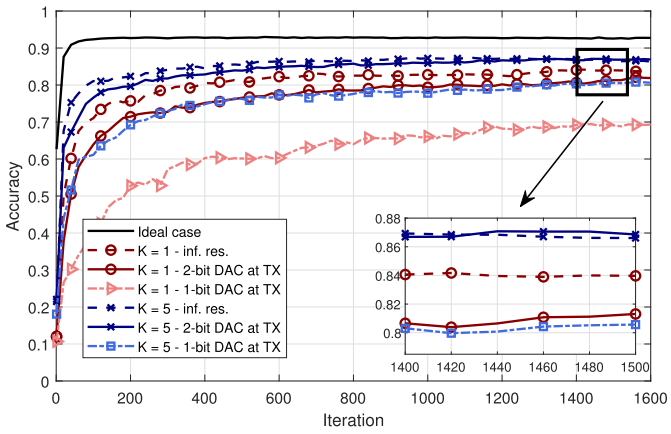
which are the same as the terms given in (17) except for the last noise term. As in Section IV, the noise term, $P_k[i]$, includes both the channel noise and the quantization noise due to ADCs, and it is with zero mean and finite variance. The analyses of the interference term (59b), distortion noise term (59c), and the second type of distortion noise term (59d) are the same as those of (17b), (17c), and (17d), respectively. Hence, similar arguments on the convergence of the learning algorithm with finite resolution DACs are also valid for the combined effects of DACs and ADCs. In other words, using a sufficiently large number of antennas at the PS, the gradients can be recovered via (33). The main conclusion is that we can design a federated learning system with a large number of workers and receive antennas, and still have extremely low hardware cost and energy consumption. This is remarkable since it shows the practicality of the federated learning over realistic wireless channels with very low-cost hardware.

VI. Numerical Examples

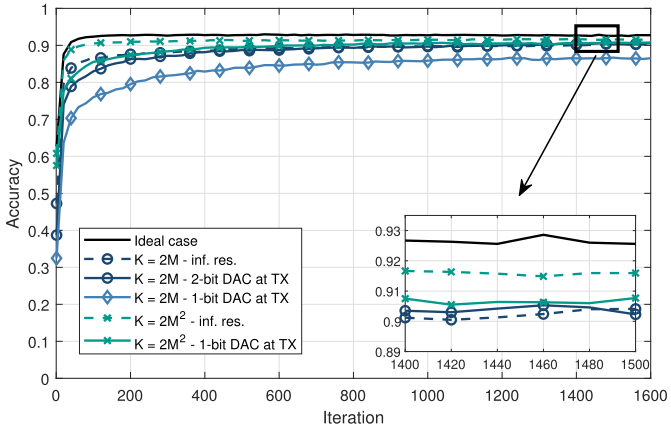
We now evaluate the performance of blind federated learning with realistic channel effects and hardware limitations via simulations. Our main objective is to verify the theoretical expectations on the low-cost federated learning systems over wireless channels via simulations. We use the MNIST dataset [36] with 60000 training and 10000 test samples to train a single layer neural network using the Adam optimizer [37]. At the beginning of the training process, each worker caches $B = 1000$ training samples randomly. The number of parameters is $d = 7850$.

Our system consists of $M = 20$ workers connected to a PS through a multipath fading channel with $L = 3$ taps and $\sigma_{h,l}^2 = 1/L$, hence we have a normalized uniform multipath delay profile where each tap experiences Rayleigh fading. We consider an OFDM setup with $f_c = 3$ GHz carrier frequency, and the number of subcarriers is $N_{\text{cp}} = 4096$ where the subcarrier spacing is $\Delta f = 80$ kHz. We take the sampling period as $T_s = T_w/N$ where $T_w = \frac{1}{\Delta f} = 12.5\ \mu$s is the OFDM word duration without the CP. As given in [38], the maximum delay spread of a typical urban area is 3.5 μs. Consider a wireless network in an urban area where the delay spread is 3.05 μs which is approximately $1000T_s$. We assume that the first tap has no delay and coherence time corresponds to $1000T_s$. Also, time delays are uniformly spaced, i.e., $\tau_{mk1} = 0$, $\tau_{mk2} = 500T_s$, $\tau_{mk3} = 1000T_s$ for $\forall m, k$[1]. The cyclic prefix length is set to $N_{\text{cp}} = 1024$,

[1]We select this multipath delay profile for ease of illustration. More realistic multipath delay profiles, e.g., exponential delay profiles, can be selected, but doing so will not change our main conclusions.



(a) Number of receive antennas $K = 1, 5$.

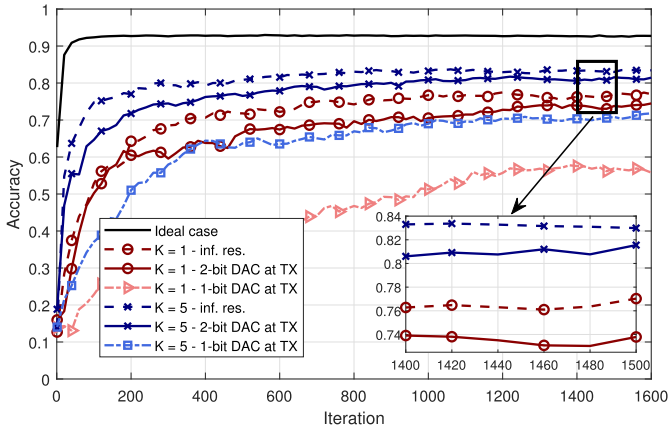


(b) Number of receive antennas $K = 2M, 2M^2$.

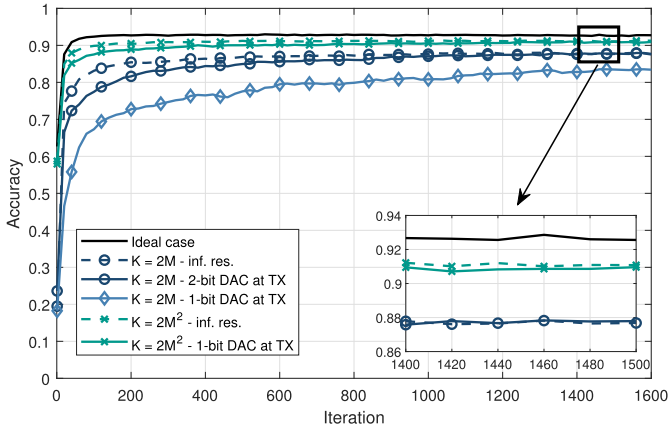
Fig. 3. Test accuracy of the system with low-resolution DACs for channel noise variance $\sigma_z^2 = 8 \times 10^{-4}$.

which is enough to remove the intersymbol interference caused by the multipath. The average transmit power of the OFDM word transmitted by the m-th worker is calculated as $P_T = \frac{1}{T}\sum_{t=1}^{T}\left\|\bar{\mathbf{G}}_m^t\right\|_2^2$, which gives $P_T = 1.3267 \times 10^{-4}$ for this setup, where T is the total iteration count. In our theoretical analysis, we model the OFDM words with the autocorrelation matrix $\mathbf{C}_{\bar{\mathbf{G}}_m\bar{\mathbf{G}}_m}$ with equal nonzero diagonal elements denoted by σ_G^2, and zero off-diagonal elements. In our simulations, we do not make any assumption on the statistics of the gradients; we simply use the estimates through the realistic channel simulations.

In Figs. 3a and 3b, the test accuracy for a system where each worker is equipped with a low-resolution DAC and different number of antennas $K \in \{1,\ 5,\ M,\ 2M^2\}$ at the receiver side is illustrated for $\sigma_z^2 = 8 \times 10^{-4}$. As the number of receive antennas increases, the test accuracy approaches that of the infinite resolution case since the variance of the distortion noise and interference decrease. At iteration $T = 1600$, the accuracy loss with one-bit DAC compared to infinite resolution case is 17.62%, 6.62%, 4.07%, and 0.37% for $K = 1$, $K = 5$, $K = 2M$, and $K = 2M^2$, respectively. Furthermore, the low complexity system achieves almost the

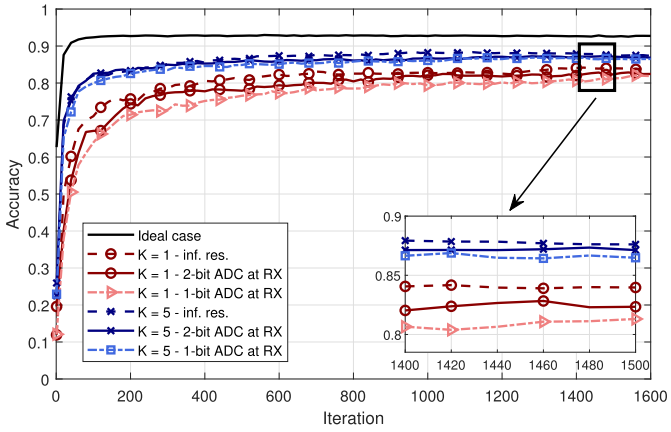


(a) Number of receive antennas $K = 1, 5$.

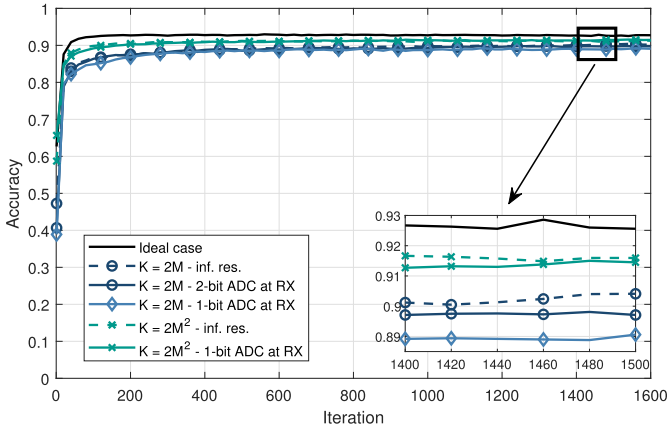


(b) Number of receive antennas $K = 2M, 2M^2$.

Fig. 4. Test accuracy of the system with low-resolution DACs for channel noise variance $\sigma_z^2 = 4 \times 10^{-3}$.



(a) Number of receive antennas $K = 1, 5$.



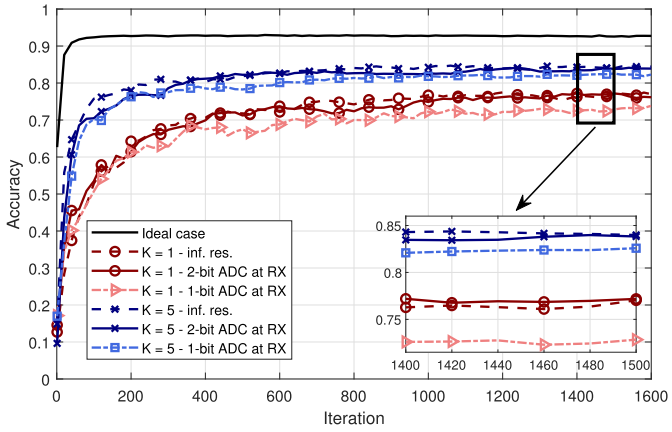
(b) Number of receive antennas $K = 2M, 2M^2$.

Fig. 5. Test accuracy of the system with low-resolution ADCs for channel noise variance $\sigma_z^2 = 8 \times 10^{-4}$.

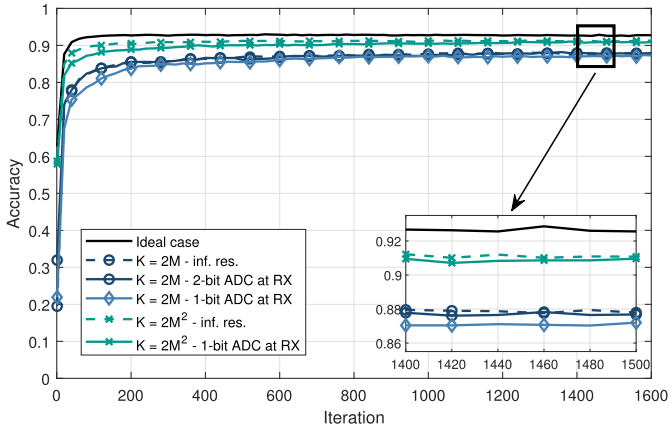
same accuracy with the infinite resolution case when two-bit DACs are employed (except for $K = 1$ which has an accuracy loss of 2.64%). In Figs. 4a and 4b, we increase the channel noise variance to $\sigma_z^2 = 4 \times 10^{-3}$, i.e., there is a 14 dB SNR reduction. Since the effect of the noise term is increased, as expected, the performance of the learning algorithm deteriorates. However, as shown in Figs. 4a and 4b, the convergence is still achieved, and the accuracy loss of the one-bit DAC case compared to infinite resolution case is 27.54%, 13.95%, 4.71%, and 0.8% for $K = 1$, $K = 5$, $K = 2M$, and $K = 2M^2$, respectively. With two-bit DACs, the accuracy loss decreases to 3.26% and 2.40% for $K = 1$ and $K = 5$, respectively, while it gives almost the same performance when the number of PS antennas is $K = 2M$ and $K = 2M^2$. These results clearly illustrate that when a moderate number of receive antennas are employed, low-resolution, even two-bit, DACs can achieve a learning performance comparable with the infinite resolution case.

In Figs. 5a and 5b, the test accuracy for different number of PS antennas $K \in \{1,\ 5,\ M,\ 2M^2\}$ each equipped with a low-resolution ADC is illustrated for a system with $\sigma_z^2 = 8 \times 10^{-4}$, and compared with the error-free shared link case. As expected, using higher number of receive antennas results in an improved learning accuracy. Indeed the results are very close to those of the infinite resolution case, especially with two-bit ADCs, while there is a minor drop on accuracy with one-bit ADCs. For instance, after the 1600-th iteration, using one-bit ADCs causes only 2.64%, 0.95%, and 0.13%, accuracy loss compared to infinite resolution case for $K = 1$, $K = 5$, and $K = 2M$, respectively. Furthermore, the system achieves the performance of the infinite resolution scenario with $K = 2M^2$ PS antennas. These results are due to the fact that increasing the number of antennas reduces the interference dramatically which makes the combined signal a very good estimate of the gradient vector, even with low-resolution ADCs.

Without changing any other parameters of the setup described above, we increase the noise variance to $\sigma_z^2 = 4 \times 10^{-3}$ in Figs. 6a and 6b. As in the previous case, for the two-bit ADC case, the performance of the proposed scheme is very close to the error-free case for a large number of receive antennas. When the number of antennas is decreased, with the detrimental effects of the channel noise and interference caused by the multipath fading channel, the accuracy decreases. However, even for this high level of channel noise, using one-bit ADCs causes only 4.09%, 2.55%, 0.37%, and

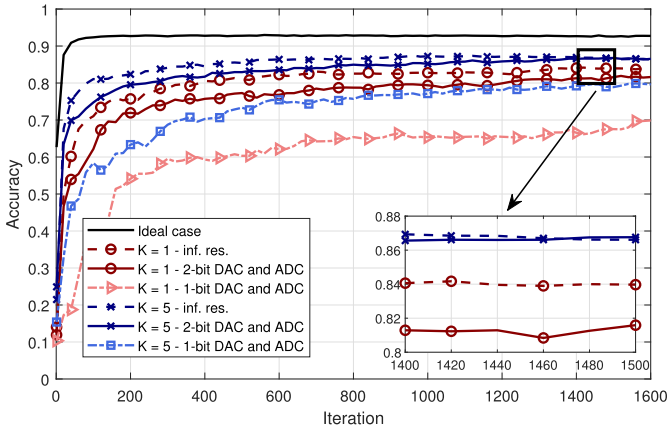


(a) Number of receive antennas $K = 1, 5$.

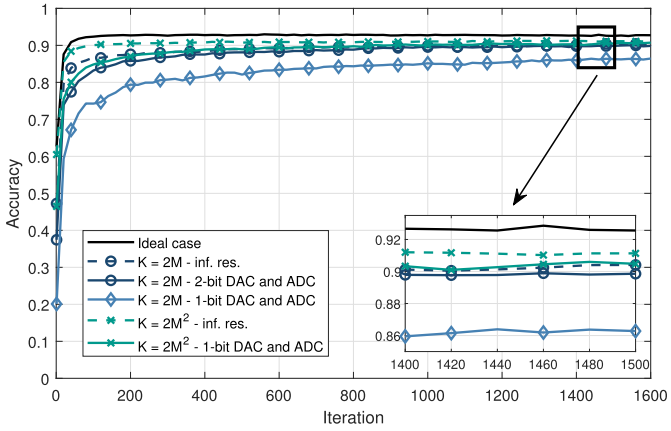


(b) Number of receive antennas $K = 2M, 2M^2$.

Fig. 6. Test accuracy of the system with low-resolution ADCs for channel noise variance $\sigma_z^2 = 4 \times 10^{-3}$.



(a) Number of receive antennas $K = 1, 5$.



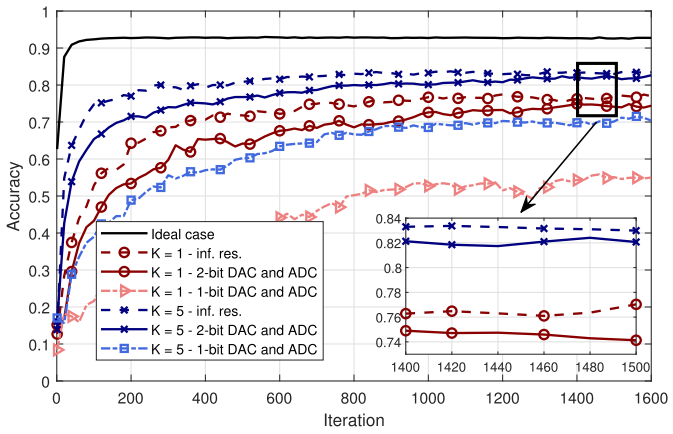
(b) Number of receive antennas $K = 2M, 2M^2$.

Fig. 7. Test accuracy of the system with low-resolution DACs and ADCs for channel noise variance $\sigma_z^2 = 8 \times 10^{-4}$.

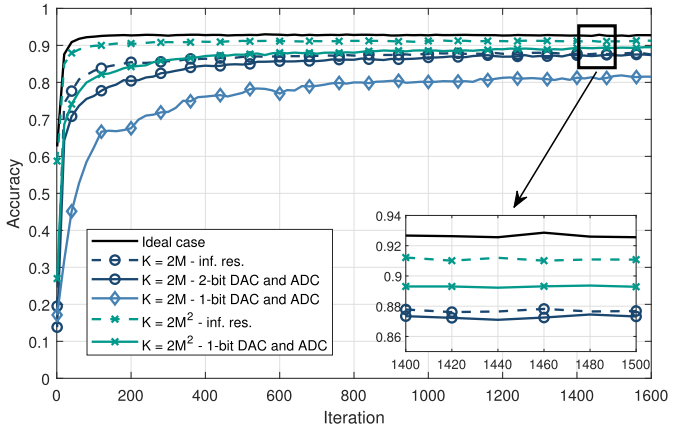
0.32% accuracy loss compared to the infinite resolution case for $K = 1$, $K = 5$, $K = 2M$, and $K = 2M^2$, respectively, after the 1600-th iteration.

In Figs. 7a and 7b, we consider a system which employs both low-resolution DACs at the workers and one-bit ADCs at the PS antennas with channel noise variance $\sigma_z^2 = 8 \times 10^{-4}$. As expected, using low-resolution DAC and ADC at the same time increases the amount of interference in the gradient estimates at the PS, which decreases the learning accuracy of the distributed system. However, the combined effect of the interference terms is still negligible, especially for sufficiently large number of receive antennas. After the 1600-th iteration, the use of one-bit DACs and ADCs simultaneously causes only 17.91%, 7.76%, 4.18%, and 0.39% accuracy loss compared to the infinite resolution case for $K = 1$, $K = 5$, $K = 2M$, and $K = 2M^2$, respectively. When $K = 1$, using two-bit DACs and ADCs results in a 2.95% accuracy loss while the performance is almost the same as that of the infinite resolution case when the number of PS antennas is higher. In the same system, we increase the channel noise variance to $\sigma_z^2 = 4 \times 10^{-3}$, and show the corresponding results in Figs. 8a and 8b. We observe that increasing the noise level causes 28.56%, 15.66%, 6.87%, and 1.91% accuracy loss compared to the infinite resolution case for $K = 1$, $K = 5$, $K = 2M$, and $K = 2M^2$, respectively after the 1600-th iteration (with one-bit DACs and ADCs). With two-bit DACs and ADCs, the accuracy loss decreases to 3.35% and 2.57% for $K = 1$ and $K = 5$, respectively.

Finally, in Fig. 9, we compare the effect of one-bit quantization on the transmitter and receiver sides, both separately and jointly, with a fixed number of receive antennas $K = 5$. As expected, the test accuracy of the system with one-bit DAC workers and infinite resolution ADCs at the PS is lower than that for the case of infinite resolution DACs at the workers and one-bit ADCs at the PS. This is because, using DACs at the workers results in higher interference than using ADCs at the PS, and the performance is deteriorated. However, the convergence of the learning algorithm is preserved. Another important implication of our results is that even though our analysis is based on a certain assumption on the statistics of the gradients, the simulation results (which are obtained without using the Gaussian assumption on the OFDM words) are consistent with our theoretical expectations. Hence, with a slight sacrifice on the accuracy rate of the learning algorithm, power and hardware efficient systems (at both transmitter and receiver sides) can be designed and implemented for



(a) Number of receive antennas $K = 1, 5$.



(b) Number of receive antennas $K = 2M, 2M^2$.

Fig. 8. Test accuracy of the system with low-resolution DACs and ADCs for channel noise variance $\sigma_z^2 = 4 \times 10^{-3}$.

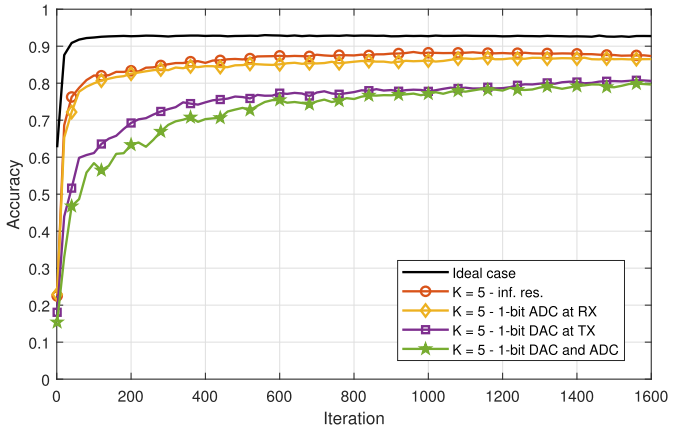


Fig. 9. Test accuracy of the system with seperate one-bit DACs at the workers, one-bit ADCs at the PS antennas, and joint DACs and ADCs where the channel noise variance is $\sigma_z^2 = 8 \times 10^{-4}$, and $K = 5$.

distributed learning at the wireless edge over realistic wireless channels.

VII. Conclusion

We have investigated blind federated learning at the wireless edge with OFDM based transmission and low-resolution, even one-bit, DACs and ADCs at the transmitter and receiver sides, respectively, for a practical and inexpensive system design, and reduced power consumption. Our analytical results illustrate that with low-resolution DACs at the transmitter and ADCs at the receiver, the convergence of the distributed learning algorithms based on SGD is guaranteed when the number of receive antennas is increased as in the ideal case of infinite resolution DACs and ADCs. Moreover, the convergence is still attained with the joint use of DACs and ADCs which reduces the implementation costs further. The results are also valid for the extreme case of one-bit DACs and ADCs. Through extensive numerical examples, it is also illustrated that using a moderate number of antennas with low-resolution DACs and ADCs, e.g., using 5 antennas at the PS, can closely approach the performance of the infinite resolution case. It is also observed that, in case of low channel noise, the learning performance is decreased only slightly even for the extreme case of one-bit ADCs and DACs.

References

[1] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling Up Machine Learning: Parallel and Distributed Approaches*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[2] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, "Project Adam: Building an efficient and scalable deep learning training system," in *Proc. 11th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, Broomfield, CO, USA, Oct. 2014, pp. 571–582.

[3] 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, 2020.

[4] 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.

[5] 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.

[6] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.

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

[8] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 1709–1720.

[9] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, Singapore, Jun. 2014, pp. 1058–1062.

[10] S.-Y. Zhao, H. Gao, and W.-J. Li, "Quantized epoch-SGD for communication-efficient distributed learning," 2019, *arXiv:1901.03040*. [Online]. Available: http://arxiv.org/abs/1901.03040

[11] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," 2020, *arXiv:2006.10672*. [Online]. Available: http://arxiv.org/abs/2006.10672

[12] M. M. Amiri, T. M. Duman, and D. Gunduz, "Collaborative machine learning at the wireless edge with blind transmitters," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Ottowa, ON, Canada, Nov. 2019, pp. 1–5.

[13] M. M. Amiri, T. M. Duman, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, early access, Mar. 19, 2021, doi: 10.1109/TWC.2021.3065920.

[14] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[15] C. Mollen, J. Choi, E. G. Larsson, and R. W. Heath, Jr., "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 87–100, Jan. 2017.

[16] D. Dardari, "Joint clip and quantization effects characterization in OFDM receivers," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 53, no. 8, pp. 1741–1748, Aug. 2006.
[17] J. Zhang, L. Dai, X. Li, Y. Liu, and L. Hanzo, "On low-resolution ADCs in practical 5G millimeter-wave massive MIMO systems," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 205–211, Jul. 2018.
[18] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, Jun. 2017.
[19] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. Lee Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.
[20] J. Xu, W. Xu, F. Shi, and H. Zhang, "User loading in downlink multiuser massive MIMO with 1-bit DAC and quantized receiver," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Toronto, ON, Canada, Sep. 2017, pp. 1–5.
[21] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Massive MU-MIMO-OFDM downlink with one-bit DACs and linear precoding," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–6.
[22] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539–550, Apr. 1999.
[23] H.-S. Lee and C. G. Sodini, "Analog-to-digital converters: Digitizing the analog world," *Proc. IEEE*, vol. 96, no. 2, pp. 323–334, Feb. 2008.
[24] S. Wei, D. L. Goeckel, and P. A. Kelly, "Convergence of the complex envelope of bandlimited OFDM signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4893–4904, Oct. 2010.
[25] S. Jacobsson, U. Gustavsson, G. Durisi, and C. Studer, "Massive MU-MIMO-OFDM uplink with hardware impairments: Modeling and analysis," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 1829–1835.
[26] S. R. Aghdam and T. Eriksson, "On the performance of distortion-aware linear receivers in uplink massive MIMO systems," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Oulu, Finland, Aug. 2019, pp. 208–212.
[27] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Dec. 2015.
[28] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-6, no. 1, pp. 7–12, Mar. 1960.
[29] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. 216, Mar. 1952.
[30] O. T. Demir and E. Bjornson, "The bussgang decomposition of nonlinear systems: Basic theory and MIMO extensions," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 131–136, Jan. 2021.
[31] E. Bjornson, L. Sanguinetti, and J. Hoydis, "Hardware distortion correlation has negligible impact on UL massive MIMO spectral efficiency," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1085–1098, Feb. 2019.
[32] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Robust predictive quantization: Analysis and design via convex optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 618–632, Dec. 2007.
[33] O. Orhan, E. Erkip, and S. Rangan, "Low power analog-to-digital conversion in millimeter wave systems: Impact of resolution and bandwidth on performance," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2015, pp. 191–198.
[34] J. Zhang, L. Dai, Z. He, B. Ai, and O. A. Dobre, "Mixed-ADC/DAC multipair massive MIMO relaying systems: Performance analysis and power optimization," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 140–153, Jan. 2019.
[35] P. Billingsley, *Probability and Measure*. Hoboken, NJ, USA: Wiley, 2008.
[36] Y. LeCun (1998). *The MNIST Database of Handwritten Digits*. [Online]. Available: http://yann. lecun.com/exdb/mnist/
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[38] W. C. Lee, *Mobile Communications Engineering: Theory and Applications*. New York, NY, USA: McGraw-Hill, 1998.

Busra Tegin (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2017 and 2020, respectively, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Electronics Engineering. She is currently a Research Engineer with the Turkey R&D Center, Huawei Technologies Company Ltd., Istanbul, Turkey. Her research interest includes the general area of wireless communications, with an emphasis in federated learning and distributed computing.

Tolga M. Duman (Fellow, IEEE) received the B.S. degree from Bilkent University, Ankara, Turkey, in 1993, and the M.S. and Ph.D. degrees from Northeastern University, Boston, MA, USA, in 1995 and 1998, respectively, all in electrical engineering. He is currently a Professor with the Department of Electrical and Electronics Engineering, Bilkent University. Prior to joining Bilkent University in September 2012, he was a Professor with the School of ECEE, Arizona State University. His current research interests are in systems, with particular focus on communication and signal processing, including wireless and mobile communications, coding/modulation, coding for wireless communications, data storage systems, and underwater acoustic communications. He was a recipient of the National Science Foundation CAREER Award and the IEEE Third Millennium Medal. He is currently the Editor-in-Chief of IEEE TRANSACTIONS ON COMMUNICATIONS.