A Compressive Sensing Approach for Federated Learning over Massive MIMO Communication Systems

Yo-Seb Jeon, Mohammad Mohammadi Amiri, Jun Li, and H. Vincent Poor

***Abstract*—Federated learning is a privacy-preserving approach to train a global model at a central server by collaborating with wireless devices, each with its own local training data set. In this paper, we present a compressive sensing approach for federated learning over massive multiple-input multiple-output communication systems in which the central server equipped with a massive antenna array communicates with the wireless devices. One major challenge in system design is to reconstruct local gradient vectors accurately at the central server, which are computed-and-sent from the wireless devices. To overcome this challenge, we first establish a transmission strategy to construct sparse transmitted signals from the local gradient vectors at the devices. We then propose a compressive sensing algorithm enabling the server to iteratively find the linear minimum-mean-square-error (LMMSE) estimate of the transmitted signal by exploiting its sparsity. We also derive an analytical threshold for the residual error at each iteration, to design the stopping criterion of the proposed algorithm. We show that for a sparse transmitted signal, the proposed algorithm requires less computationally complexity than LMMSE. Simulation results demonstrate that the presented approach outperforms conventional linear beamforming approaches and reduces the performance gap between federated learning and centralized learning with perfect reconstruction.**

***Index Terms*—Federated learning, distributed stochastic gradient descent, massive multiple-input multiple-output (MIMO), compressive sensing, multi-antenna technique**

I. Introduction

Machine learning has attracted significant interest as a breakthrough for emerging applications of wireless communications [1]–[6]. The fundamental idea of machine learning for wireless communications is to learn a model (e.g., input-output relation) based on large amounts of data and computing power. It has been demonstrated that the learned model can be exploited either to improve the performance of conventional model-based techniques (e.g. [3], [4]) or to describe an unknown input-output relation whose characterization was otherwise challenging due to mathematical intractability (e.g., [5], [6]). The most common and widely adopted form of machine learning is *centralized learning* in which a central server equipped with sufficient storage and computing power has full access to the entire data set. Unfortunately, centralized learning is infeasible in many applications of wireless communications. The major reason is that data sets are usually generated at wireless devices, but transmitting them to the central server is highly limited by both the amount of radio resources and communication latency allowed in the applications. Particularly, this problem becomes more severe as the data size and the model complexity increase. In addition, sending the data may not be allowed in privacy-sensitive applications such as social networking, e-health, and financial services.

Recently, federated learning has drawn increasing attention as a viable solution to overcome the limitations of centralized learning [7]–[10], in which a global model at a central server is collaboratively trained by multiple devices each with its own local data set. The major advantage of federated learning is a significant reduction in communication overhead as the devices only send an updated model instead of the whole data set. In addition, this approach preserves the privacy of the devices because the data is kept where it is generated while a central server has no direct access to the local data sets. Distributed machine learning also provides similar advantages, but federated learning focuses on a more practical setting which may include unbalanced and non-identically-distributed data sets, unreliable communication, and massively distributed data [7]. Thanks to the advantages and the practicality of federated learning, it has been adopted as a key enabler for emerging applications of wireless communications [11]–[13]. For example, in [12], federated learning is applied to learn the statistical properties of vehicular users in wireless vehicular networks. Another example is introduced in [13] which applies federated learning to learn the locations and orientations of the users in wireless virtual reality networks.

There also exist recent studies that seek to enable and optimize federated learning over wireless communication systems by considering the physical characteristics of wireless channels [14]–[23]. A device scheduling problem is studied in [14]–[18] based on various scheduling criteria, while a joint resource allocation and user scheduling problem is tackled in [19]. Transmission and reception techniques for federated

Y.-S. Jeon is with the Department of Electrical Engineering, POSTECH, Pohang, Gyeongbuk 37673, South Korea (e-mail: yoseb.jeon@postech.ac.kr).

M. M. Amiri and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 (e-mails: mamiri@princeton.edu, poor@princeton.edu).

J. Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, 210094, CHINA. He is also with the School of Computer Science and Robotics, National Research Tomsk Polytechnic University, Tomsk, 634050, RUSSIA (e-mail: jun.li@njust.edu.cn).

This work was supported in part by the National Research Foundation of Korea (NRF) grant (No. NRF-2020R1G1A1099962) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No. 2016-0-00123, Development of Integer-Forcing MIMO Transceivers for 5G & Beyond Mobile Communication Systems) funded by the Korea government (MSIT), in part by National Natural Science Foundation of China under Grant 61872184, and in part by the U.S. National Science Foundation under Grants CCF-0939370 and CCF-1908308.

learning are developed for a simple Gaussian multiple access channel (MAC) in [20] and for a fading MAC in [21]. A transmission technique for the fading MAC is also proposed in [22] jointly with a device scheduling method. To improve the robustness against channel fading and noise effects in wireless environments, the use of multiple antennas at a central server and/or wireless devices is considered in [24]–[27]. The effect of multiple antennas on the performance of federated learning is investigated in [24] under the assumption that receive beamforming is employed by a multi-antenna central server. In this work, the use of the receive beamforming is shown to be an effective approach when the server is equipped with a sufficiently large number of antennas. Under the same assumption as in [24], training time optimization for federated learning is studied in [25]. Transmit and/or receive beamforming for over-the-air computation are considered in [26], [27] which can be utilized to reduce latency in federated learning. Specifically, joint optimization for device selection and receive beamforming is proposed in [26], while joint design of transmit and receive beamforming is proposed in [27]. The common limitation of the existing works in [24]–[27], however, is that they only consider linear beamforming approaches to design the reception technique at the multi-antenna server. Unfortunately, such linear beamforming approach is not optimal in terms of the estimation performance at the server in general; thereby, further investigation on the reception technique is still needed to maximize the estimation performance at the multi-antenna central server in federated learning over wireless communication systems.

In this work, we study federated learning over massive multiple-input multiple-output (MIMO) communication systems, in which a global model is trained through collaboration of multiple wireless devices, each with its own local training data set, with a central server equipped with a massive number of antennas. One of the challenges in system design is to reconstruct local gradient vectors accurately at the central server, which are computed-and-sent from the wireless devices. Our key observation to overcome this challenge is that the local gradient vectors are likely to be sparse in the considered federated learning framework. Motivated by this observation, we present a compressive sensing approach that exploits the sparsity of the local gradient vectors, to enable efficient reconstruction of them at the central server. To the best of our knowledge, this work is the first attempt to establish a compressive sensing approach for a multi-antenna central server in federated learning, to exploit the sparsity in a spatial-device domain. Note that the existing works in [20], [21] also consider a compressive sensing approach, but for a single-antenna server, while exploiting the sparsity in a different domain. The major contributions of this paper are summarized as follows:

- We establish a transmission strategy to construct sparse transmitted signals from the local gradient vectors at the wireless devices. The basic idea of this strategy is to permute the local gradient vectors using different patterns across the wireless devices. The major advantage of our transmission strategy is that when the local gradient vectors are sparse, only a small subset of devices simultaneously transmit non-zero gradient elements at each radio resource, which results in a sparse transmitted signal. Using simulations, we demonstrate that the use of our transmission strategy significantly improves the sparsity of the transmitted signal in the considered federated learning framework.
- We propose a compressive sensing algorithm enabling the central server to efficiently estimate the transmitted signal by exploiting its sparsity. The basic idea of the proposed algorithm is to iteratively find the linear minimum-mean-square-error (LMMSE) estimate of non-zero elements of the transmitted signal. In the proposed algorithm, to properly determine the LMMSE estimate at the server, we establish a statistical model for the transmitted signal based on a large-scale approximation and a statistical feature obtained from our transmission strategy. We also derive an analytical threshold for the residual error at each iteration, to design the stopping criterion of the proposed algorithm. Simulation results demonstrate that the proposed compressive sensing algorithm with our transmission strategy efficiently reduces the performance gap between federated learning and centralized learning with perfect reconstruction, when the size of the mini-batch employed at each device is relatively small.
- We compare our compressive sensing approach with linear beamforming approaches that can be employed to reconstruct local gradient vectors at the multi-antenna central server. To this end, we introduce proper modifications of the conventional maximum ratio combining (MRC) and LMMSE methods and present their limitations for the use in federated learning over a massive MIMO system. We then compare the computational complexity of the proposed algorithm with those of the linear beamforming methods. Our key finding is that the proposed algorithm requires less complexity than the LMMSE method when the transmitted signal is sparse, which is verified through both analytical and numerical results. We also demonstrate that the proposed algorithm outperforms linear beamforming methods in terms of the classification accuracy of federated learning, using simulations.

Notation: Upper-case and lower-case boldface letters denote matrices and column vectors, respectively. $\mathbb{E}[\cdot]$ is the statistical expectation, $\mathbb{P}(\cdot)$ is the probability, $(\cdot)^{\mathsf{T}}$ is the transpose, $(\cdot)^{\mathsf{H}}$ is the conjugate transpose, $\lceil\cdot\rceil$ is the ceiling function, $\lfloor\cdot\rfloor$ is the floor function, and $|\cdot|$ is the absolute value. $\mathsf{Re}\{\cdot\}$ and $\mathsf{Im}\{\cdot\}$ denote real and imaginary components, respectively. $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$. $(\mathbf{a})_i$ represents the i-th element of vector $\mathbf{a}$. $\|\mathbf{a}\| = \sqrt{\mathbf{a}\mathbf{a}^{\mathsf{H}}}$ is the Euclidean norm of vector $\mathbf{a}$. $\mathcal{CN}(\boldsymbol{\mu}, \mathbf{R})$ represents the distribution of a circularly symmetric complex Gaussian random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$. $\mathbf{0}_n$ and $\mathbf{1}_n$ are an n-dimensional vectors whose elements are zero and one, respectively. $\mathbb{R}$ is the set of real numbers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.3038407, IEEE Transactions on Wireless Communications

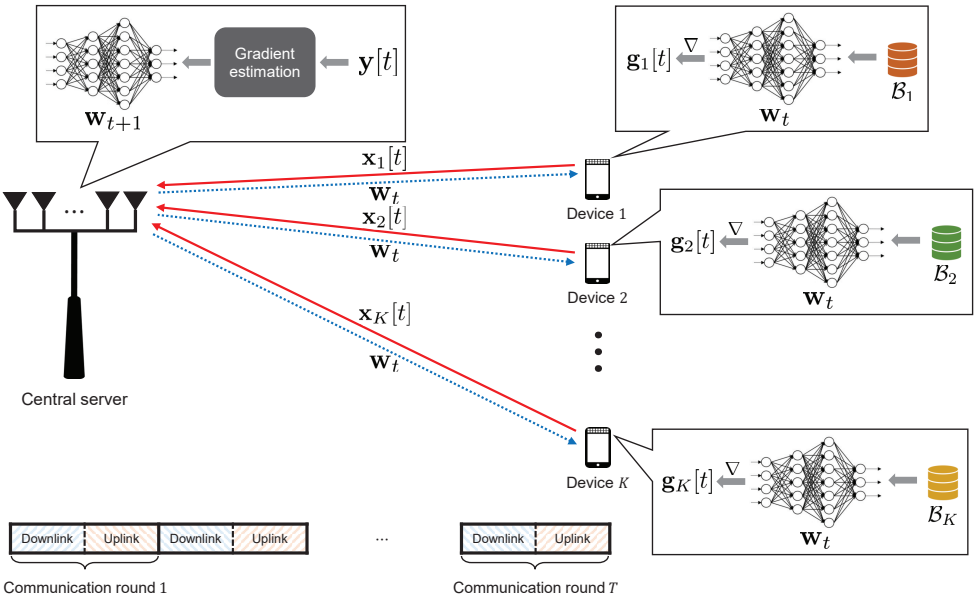


Fig. 1. An illustration of federated learning over a TDD massive MIMO communication system.

II. System Model

We consider federated learning over a time-division-duplex (TDD) massive MIMO communication system in which a central server equipped with M antennas trains a global model by collaborating with K single-antenna wireless devices, as illustrated in Fig. 1. In this system, the server and the wireless devices share a global model (e.g., neural network) represented by a parameter vector $\mathbf{w} \in \mathbb{R}^{N_{\rm w}}$, but only the devices have training data samples to train the global model. We denote $\mathcal{B}_k$ as a local training data set available at device k which consists of $|\mathcal{B}_k|$ training data samples for $k \in \mathcal{K} = \{1, \ldots, K\}$. We model the wireless channel from the devices to the server by an L-tap channel impulse response (CIR). We assume that this CIR is perfectly known at the server via uplink channel training[1] performed once per each uplink transmission and remains constant until the next downlink transmission. Note that this is a common assumption in TDD massive MIMO systems [28], [29]. We adopt an orthogonal frequency division multiplexing (OFDM) modulation with $N_{\rm sub}$ subcarriers to obtain parallel subchannels without inter-symbol interference.

We assume that the global model is trained using a gradient-based iterative algorithm (e.g., stochastic gradient descent algorithm or Adam optimizer [30]). Each iteration of the algorithm corresponds to one communication round that consists of uplink and downlink phases, as illustrated in Fig. 1. Let T be the number of iterations and $\mathbf{w}_t$ be the parameter vector at iteration t of the algorithm. During the uplink phase at communication round t, the server broadcasts $\mathbf{w}_t$ to the wireless devices. Then during the uplink phase, the wireless devices transmit their local gradient vectors to the server. In this work, we focus only on a transmission/reception strategy for the uplink phase, while assuming that the broadcasting of $\mathbf{w}_t$ in the downlink phase is error free, as assumed in most literature [20]–[22], [24], [26]. Under this assumption, all devices have a globally consistent parameter vector $\mathbf{w}_t$ for all $t \in \mathcal{T} = \{1, \ldots, T\}$.

We present the transmission procedure of each wireless device during the uplink phase. In this work, we assume that the local gradient vector at each device is computed over only a small fraction of its local data set, considering a limited computing power and stringent latency constraint at the wireless devices in practical communication systems. Let $\mathcal{B}_k[t] \subset \mathcal{B}_k$ be a set of the samples selected by device k to compute the local gradient vector at communication round t. Then a local gradient vector at device k is computed as

$$\mathbf{g}_k[t] = \frac{1}{|\mathcal{B}_k[t]|} \sum_{b \in \mathcal{B}_k[t]} \nabla f(\mathbf{w}_t, b), \tag{1}$$

where $\nabla f(\cdot, b)$ is the gradient of a loss function computed for the training data sample $b \in \mathcal{B}_k$ defined by the learning task. After computing the local gradient vector, a transmitted signal at device k can be constructed as

$$\mathbf{x}_k[t] = \sqrt{\frac{N_{\rm w}}{\|\mathbf{g}_k[t]\|^2}}\mathbf{g}_k[t]. \tag{2}$$

The scaling operation in (2) is adopted to ensure that every device has the same transmit power of $\|\mathbf{x}_k[t]\|^2 = N_{\rm w}$ at each communication round. We assume that each uplink phase consists of $\lceil \frac{N_{\rm w}}{N_{\rm sub}} \rceil$ OFDM symbols, each with $N_{\rm sub}$ subcarriers, to support the transmission of $N_{\rm w}$ elements. Under this assumption, the n-th element of $\mathbf{x}_k[t]$, namely $x_k[t,n] \in \mathbb{R}$, is transmitted using the n-th radio resource of the uplink phase, corresponding to the f_n-th subcarrier of the u_n-th OFDM symbol where $f_n = n - (u_n - 1)N_{\rm sub}$ and $u_n = \lceil \frac{n}{N_{\rm sub}} \rceil$.

We now describe the reception procedure of the server during the uplink phase. The received signal associated with

[1]To enable accurate uplink CSI at the server, each device may need to send more than K pilot signals, but this overhead is still negligible compared to the overhead required for gradient vector transmission since typically $N_{\rm w} \gg K$ in the federated learning framework.

the n-th radio resource of the uplink phase is expressed as

$$\mathbf{y}^{\mathrm{c}}[t,n] = \sum_{k=1}^{K} \mathbf{h}_k^{\mathrm{c}}[t,n] x_k[t,n] + \mathbf{z}^{\mathrm{c}}[t,n], \tag{3}$$

where $\mathbf{h}_k^{\mathrm{c}}[t,n]$ is the channel frequency response vector of device k and $\mathbf{z}^{\mathrm{c}}[t,n] \in C^M$ is the noise signal at the f_n-th subcarrier of the u_n-th OFDM symbol. We assume that the noise signal at each radio resource is independent and identically distributed (i.i.d.) as $\mathcal{CN}(\mathbf{0}_M, \sigma_{\mathrm{c}}^2 \mathbf{I}_M)$. The real-domain equivalent representation of $\mathbf{y}^{\mathrm{c}}[t,n]$ is given by

$$\mathbf{y}[t,n] = \sum_{k=1}^{K} \mathbf{h}_k[t,n] x_k[t,n] + \mathbf{z}[t,n], \tag{4}$$

where

$$\begin{aligned}
\mathbf{y}[t,n] &= \left[\mathsf{Re}(\mathbf{y}^{\mathrm{c}}[t,n])^{\mathsf{T}}, \mathsf{Im}(\mathbf{y}^{\mathrm{c}}[t,n])^{\mathsf{T}}\right]^{\mathsf{T}}, \\
\mathbf{h}_k[t,n] &= \left[\mathsf{Re}(\mathbf{h}_k^{\mathrm{c}}[t,n])^{\mathsf{T}}, \mathsf{Im}(\mathbf{h}_k^{\mathrm{c}}[t,n])^{\mathsf{T}}\right]^{\mathsf{T}}, \\
\mathbf{z}[t,n] &= \left[\mathsf{Re}(\mathbf{z}^{\mathrm{c}}[t,n])^{\mathsf{T}}, \mathsf{Im}(\mathbf{z}^{\mathrm{c}}[t,n])^{\mathsf{T}}\right]^{\mathsf{T}}.
\end{aligned}$$

The above representation can be rewritten as

$$\mathbf{y}[t,n] = \mathbf{H}[t,n]\mathbf{x}[t,n] + \mathbf{z}[t,n], \tag{5}$$

where

$$\begin{aligned}
\mathbf{H}[t,n] &= \left[\mathbf{h}_1[t,n], \mathbf{h}_2[t,n], \cdots, \mathbf{h}_K[t,n]\right], \\
\mathbf{x}[t,n] &= \left[x_1[t,n], x_2[t,n], \cdots, x_K[t,n]\right]^{\mathsf{T}}.
\end{aligned}$$

Note that $\mathbf{z}[t,n] \sim \mathcal{CN}(\mathbf{0}_{2M}, \sigma^2 \mathbf{I}_{2M})$ with $\sigma^2 = \frac{\sigma_{\mathrm{c}}^2}{2}$. Based on the received signals $\{\mathbf{y}[t,n]\}_{n=1}^{N_{\mathrm{w}}}$, the server estimates the transmitted signals $\{\mathbf{x}[t,n]\}_{n=1}^{N_{\mathrm{w}}}$ to obtain the information of the local gradient vectors sent from the wireless devices. A detailed estimation process will be discussed in the sequel. Let $\hat{x}_k[t,n]$ be the estimate of the transmitted signal sent from device k at the n-th radio resource. Then by aggregating all the estimates of the transmitted signals, the local gradient vector sent from device k can be reconstructed as

$$\hat{\mathbf{g}}_k[t] = \sqrt{\frac{\|\mathbf{g}_k[t]\|^2}{N_{\mathrm{w}}}} \hat{\mathbf{x}}_k[t], \tag{6}$$

where $\hat{\mathbf{x}}_k[t] = \left[\hat{x}_k[t,1], \cdots, \hat{x}_k[t,N_{\mathrm{w}}]\right]^{\mathsf{T}}$ for $k \in \mathcal{K}$. In (6), we assume that the norm of the local gradient vector $\|\mathbf{g}_k[t]\|$ for all $k \in \mathcal{K}$ is known[2] at the server. After reconstructing all the local gradient vectors, the server aggregates these vectors to obtain the global gradient vector defined as

$$\bar{\mathbf{g}}[t] = \frac{1}{\sum_{j=1}^{K} |\mathcal{B}_j[t]|} \sum_{k=1}^{K} |\mathcal{B}_k[t]| \, \hat{\mathbf{g}}_k[t]. \tag{7}$$

The computing power of each device may not change during the training process, so we assume that $\{|\mathcal{B}_k[t]|\}_{t \in \mathcal{T}}$ is fixed and known at the central server. The global gradient vector in (7) is utilized to update the parameter vector $\mathbf{w}_t$. For example, if the central server adopts a gradient descent algorithm, the update of the parameter vector is expressed as

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \bar{\mathbf{g}}[t], \tag{8}$$

where η_t represents the learning rate at iteration t.

In this paper, we address the mismatch between the local gradient vectors sent from the wireless devices and their estimates reconstructed at the server, i.e., $\hat{\mathbf{g}}_k[t] \neq \mathbf{g}_k[t]$, in federated learning over the massive MIMO systems. The main causes of this mismatch are 1) inter-user interference caused by simultaneous transmission of multiple devices, 2) channel fading naturally occurred in wireless channels, and 3) noise signal in RF chain. This mismatch may harm both the learning accuracy and the convergence rate of federated learning as will be demonstrated in Sec. V. Therefore, it is essential to reduce this mismatch by developing a proper reception technique at the central server that enables accurate estimation of the local gradient vectors.

III. A Compressive Sensing Approach for Efficient Reconstruction of Local Gradient Vectors

In this section, we present a compressive sensing approach that allows the central server to efficiently reconstruct the local gradient vectors sent from the wireless devices. To this end, we first discuss the sparsity of the local gradient vectors and then establish a transmission strategy to construct sparse transmitted signals from the local gradient vectors at the wireless devices. Based on this strategy, we propose a compressive sensing algorithm enabling the central server to estimate the transmitted signal by exploiting its sparsity.

A. Motivation: Sparsity of Local Gradient Vectors

Our compressive sensing approach is motivated by the sparsity of the local gradient vectors in federated learning over wireless communication systems. First of all, the average magnitudes of gradient elements are expected to reduce as a training algorithm (e.g., gradient descent algorithm) proceeds; thereby, the number of zero gradient elements may increase over time. Particularly this number is much larger when employing a ReLU activation function since the gradient of the ReLU function is zero for any negative-valued input. We have also observed that the number of zero gradient elements are likely to increase as the size of the mini-batch training data samples utilized to compute the local gradient vector at each device decreases. The rationale behind this observation is that zero gradient elements computed for each training data sample would be associated with a significantly different subset of weights across different samples. This observation is particularly relevant for federated learning over wireless communications since the size of the mini-batch samples at wireless device (e.g., smartphone or IoT device) is expected to be relatively small due to a limited computing power and/or a stringent latency constraint. All these observations imply that the local gradient vectors are likely to be sparse in federated learning over practical communication systems. It is also noticeable that transmitted signals whose magnitudes are much

[2]To convey the information of the norm of the local gradient vector, each device needs to send one additional real value, but this overhead is still negligible compared to the overhead required for gradient vector transmission since typically $N_{\mathrm{w}} \gg 1$ in the federated learning framework.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.3038407, IEEE Transactions on Wireless Communications

smaller than the noise level of a communication system can be treated as zero signals since they have negligible impacts on the performance of federated learning. Meanwhile, only a few of the largest gradient elements will remain large in the transmission signal at the wireless devices since each device normalizes the local gradient vector before its transmission, as can be seen in (2). These observations imply that even if the local gradient vectors may not be exactly sparse in certain cases (e.g., during the initial training iterations), an approximate sparsity can still be observed at the central server.

B. Transmission Strategy: Random Permutation

Even when the local gradient vectors are sparse, the transmitted signal may not be sparse if a naive transmission procedure (e.g., the procedure introduced in Sec. II) is employed by the wireless devices. For example, consider an extreme case where only the n^*-th element of the local gradient vector is non-zero at all the devices (i.e., $g_k[t, n^\star] \neq 0$ for $k \in \mathcal{K}$). In this case, the transmitted signal at the n^*-th radio resource, $\mathbf{x}[t, n^*]$, is not sparse because $x_k[t, n^*]$ is non-zero for all $k \in \mathcal{K}$ from (2). More generally, if all the local training data sets have an identical distribution, the wireless devices may have similar sparsity patterns [20], [21]; in this case, the transmitted signals constructed from (2) may not be sparse.

To prevent the loss of the sparsity, we establish a new transmission strategy that allows the wireless devices to construct sparse transmitted signals by preserving the sparsity of the local gradient vectors. The basic idea is to permute the local gradient vectors using different patterns across the wireless devices when constructing the transmitted signal. This strategy is enabled through linear projection using different permutation matrices at different devices. More precisely, the transmitted signal of device k is determined as

$$\mathbf{x}_k[t] = \sqrt{\frac{N_{\mathrm{w}}}{\|\mathbf{g}_k[t]\|^2}}\mathbf{P}_k\mathbf{g}_k[t], \tag{9}$$

where $\mathbf{P}_k \in \{0,1\}^{N_{\mathrm{w}} \times N_{\mathrm{w}}}$ is the permutation matrix employed at device k such that $\mathbf{P}_k\mathbf{P}_k^{\mathsf{T}} = \mathbf{I}_{N_{\mathrm{w}}}$. Then the server reconstructs[3] the local gradient vector sent from device k as

$$\hat{\mathbf{g}}_k[t] = \sqrt{\frac{\|\mathbf{g}_k[t]\|^2}{N_{\mathrm{w}}}}\mathbf{P}_k^{\mathsf{T}}\hat{\mathbf{x}}_k[t]. \tag{10}$$

Our strategy in (9) implies that the local gradient elements transmitted at each radio resource are associated with not only different devices but also different weights of the neural network. Therefore, when the local gradient vectors are sparse, it is likely that only a small number of devices simultaneously transmit non-zero local gradient elements at each radio resource, which results in the sparsity of the transmitted signal $\mathbf{x}[t, n]$. Fig. 2 illustrates the transmitted signals and the corresponding received signals when employing our random permutation strategy.

[3]We assume that the central server has the information of the permutation matrices employed at wireless devices. One possible approach to realize this assumption is to share a *pre-determined* generator between the server and the devices, which generates different permutation matrices for different inputs. Then each device only needs to send the selected input of the generator, once at the beginning of the training process.

We also demonstrate the sparsity of the transmitted signal through numerical simulations. As a performance metric to evaluate the sparsity, we consider a *magnitude ratio* $\xi_k[t, n]$ defined as the ratio of each element's magnitude to the maximum magnitude in the transmitted signal $\mathbf{x}[t, n]$, i.e., $\xi_k[t, n] \triangleq \frac{|x_k[t,n]|}{\max_{k'}|x_{k'}[t,n]|}$. Fig. 3 shows the cumulative distribution function of $\xi_k[t, n]$ with and without our random permutation strategy when $M = 25$ and $K = 100$ for the simulation setting described in Sec. V-A. It should be noticed that the distribution function at $\xi_k[t, n] = 0$ represents the ratio of the number of zero elements to the number of total elements in each transmitted signal. It is observed that our random permutation strategy significantly increases the number of zero elements in the transmitted signal for both the stochastic setting ($|\mathcal{B}_k[t]| = 1$) and the mini-batch setting ($|\mathcal{B}_k[t]| \sim \mathrm{Uni}[1, 30]$). Particularly for the stochastic setting, only 10% of the elements of the transmitted signal are non-zero when applying our strategy. Another important observation is that the sparsity level is higher for the stochastic setting than for the mini-batch setting. This result demonstrates that the sparsity of the transmitted signal increases as the batch size decreases, as we already discussed in Sec. III-A.

C. Reception Strategy: Compressive Sensing

The goal is to design an efficient signal recovery algorithm at the central server to estimate the transmitted signal $\mathbf{x}[t, n]$ from the received signal $\mathbf{y}[t, n]$ in (5). To achieve this goal, we propose to exploit the sparsity of the transmitted signal by using a compressive sensing approach. The intuition behind this idea is that estimating a *sparse* transmitted signal $\mathbf{x}[t, n]$ from the received signal $\mathbf{y}[t, n] = \mathbf{H}[t, n]\mathbf{x}[t, n] + \mathbf{z}[t, n]$ can be interpreted as compressive sensing to estimate an unknown *sparse* signal from its linear measurement with noise [31]. Based on this intuition, we propose a compressive sensing algorithm that iteratively finds the LMMSE estimate of the transmitted signal by exploiting its sparsity. This algorithm can be separately applied to the received signal at each radio resource of each uplink phase. Therefore, to simplify the notation, we omit the indexes t and n in the rest of this subsection.

1) Definitions: We start by defining some terminologies and notations employed in the proposed algorithm. A *true support set* $\tilde{\mathcal{K}} \subset \mathcal{K}$ is a set of device indices that have non-zero gradient values, i.e., $\tilde{\mathcal{K}} \triangleq \{x_k \neq 0 \mid k \in \mathcal{K}\}$. A *support set* at iteration i, denoted by $\mathcal{S}_i \subset \mathcal{K}$, is a set of device indices that have been selected as the members of the true support set until iteration i. An estimated transmitted signal at iteration i, denoted by $\hat{\mathbf{x}}^{(i)} \in \mathbb{R}^i$, is the estimate of the transmitted signal when assuming $\tilde{\mathcal{K}} = \mathcal{S}_i$. A *residual vector* at iteration i is defined as $\mathbf{r}_i = \mathbf{y} - \mathbf{H}^{(i)}\hat{\mathbf{x}}^{(i)}$ which is a residual part of the received signal after subtracting the effect of the estimated transmitted signal at iteration i, where $\mathbf{H}^{(i)}$ is defined as $\mathbf{H}^{(i)} = \left[\mathbf{h}_{\mathcal{S}_i(1)}, \cdots, \mathbf{h}_{\mathcal{S}_i(i)}\right]$.

2) Support Set Update: At iteration i, the proposed algorithm selects the index of the device whose normalized channel has the maximum correlation with the residual vector $\mathbf{r}_{i-1}$ at

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.3038407, IEEE Transactions on Wireless Communications

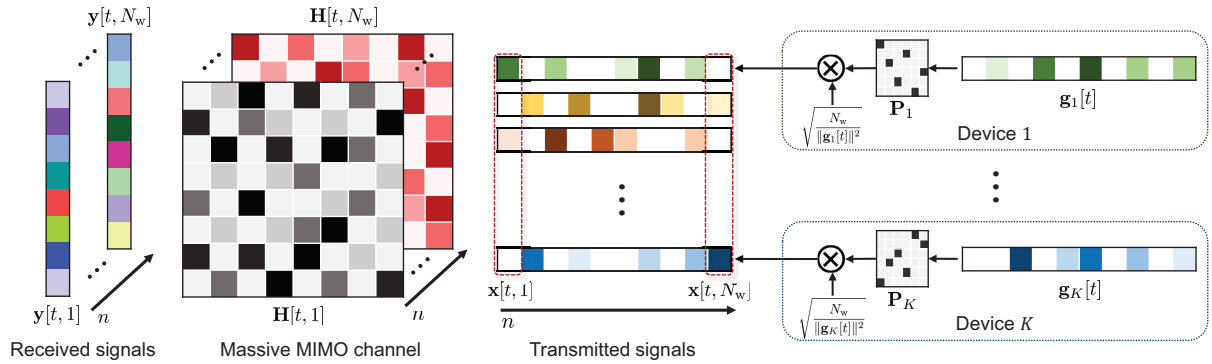


Fig. 2. An illustration of transmitted signals and the corresponding received signals when employing the random permutation strategy in Sec. III-B.

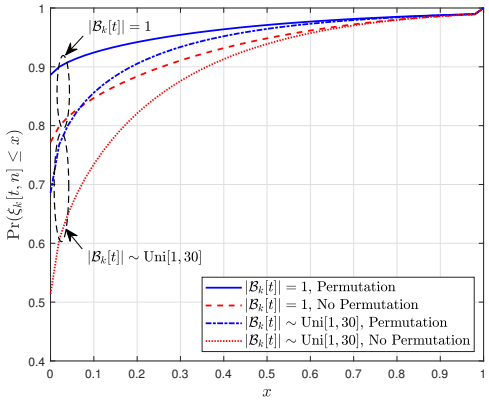


Fig. 3. Cumulative distribution function of the magnitude ratio $\xi_k[t, n]$ with and without our random permutation strategy when $M = 25$ and $K = 100$.

the previous iteration. Our strategy for this step is to use the following selection criterion:

$$k_i^\star = \underset{k \in \mathcal{S}_{i-1}^c}{\operatorname{argmax}} \left| \tilde{\mathbf{h}}_k^{\mathsf{T}} \mathbf{r}_{i-1} \right|^2, \tag{11}$$

where $\tilde{\mathbf{h}}_k = \frac{1}{\|\mathbf{h}_k\|} \mathbf{h}_k$. The promising feature of the above criterion is that it correctly finds the device index with the maximum effective SNR, defined as $\rho_k = \|\mathbf{h}_k\|^2 |x_k|^2 / \sigma^2$, when the number of the antennas at the server is sufficiently large. This fact can be readily shown by characterizing the correlation between the channel vector of device k and the residual vector, given by

$$\begin{aligned} \frac{1}{\|\mathbf{h}_k\|} \tilde{\mathbf{h}}_k^{\mathsf{T}} \mathbf{r}_{i-1} &= \frac{1}{\|\mathbf{h}_k\|^2} \mathbf{h}_k^{\mathsf{T}} \left(\mathbf{y} - \mathbf{H}^{(i-1)} \hat{\mathbf{x}}^{(i-1)} \right) \\ &= x_k + \sum_{s' \in \mathcal{S}_{i-1}^c \setminus \{k\}} \frac{\mathbf{h}_k^{\mathsf{T}} \mathbf{h}_{s'}}{\|\mathbf{h}_k\|^2} x_{s'} \\ &\quad + \sum_{s \in \mathcal{S}_{i-1}} \frac{\mathbf{h}_k^{\mathsf{T}} \mathbf{h}_s}{\|\mathbf{h}_k\|^2} \left(x_s - \hat{x}_s \right) + \frac{\mathbf{h}_k^{\mathsf{T}} \mathbf{z}}{\|\mathbf{h}_k\|^2}, \end{aligned} \tag{12}$$

where $\hat{\mathbf{x}}^{(i-1)} = \left[\hat{x}_{\mathcal{S}_{i-1}(1)}, \cdots, \hat{x}_{\mathcal{S}_{i-1}(i-1)} \right]^{\mathsf{T}}$. By the law of large numbers, the correlation in (12) approaches x_k as M increases; thereby, the metric in (11) can be approximated as

$$\left| \tilde{\mathbf{h}}_k^{\mathsf{T}} \mathbf{r}_{i-1} \right| \approx \|\mathbf{h}_k\| |x_k| \propto \sqrt{\rho_k}, \tag{13}$$

for $M \gg 1$. Since we focus on the massive MIMO system with $M \gg 1$, the proposed algorithm at iteration i is expected to find the i-th dominant element in the true support set $\tilde{\mathcal{K}}$. Once the best index $k_i^\star$ is selected by the criterion in (11), the support set $\mathcal{S}_{i-1}$ is updated by adding the selected index, i.e., $\mathcal{S}_i \leftarrow \mathcal{S}_{i-1} \cup \{k_i^\star\}$.

3) Transmitted Signal Estimation: After updating the support set at iteration i, the proposed algorithm estimates the transmitted signal associated with the current support set $\mathcal{S}_i$. Our strategy for this step is to find the LMMSE estimate that has the minimum MSE with respect to the true transmitted signal. For a given support set $\mathcal{S}_i$, the received signal at the server can be rewritten as

$$\mathbf{y} = \sum_{s \in \mathcal{S}_i} \mathbf{h}_s x_s + \sum_{s' \in \mathcal{S}_i^c} \mathbf{h}_{s'} x_{s'} + \mathbf{z}. \tag{14}$$

By the support set update rule in (11), the first term of the right hand side (RHS) of (14) may consist of signals with the i largest magnitudes. Motivated by this observation, we approximate the received signal in (14) by assuming that the magnitude of $\sum_{s \in \mathcal{S}_i} \mathbf{h}_s x_s$ is much larger than the magnitude of $\sum_{s' \in \mathcal{S}_i^c} \mathbf{h}_{s'} x_{s'}$, expressed as

$$\mathbf{y} \approx \sum_{s \in \mathcal{S}_i} \mathbf{h}_s x_s + \mathbf{z} = \mathbf{H}^{(i)} \mathbf{x}^{(i)} + \mathbf{z}, \tag{15}$$

where $\mathbf{x}^{(i)} = [x_{k_1^\star}, \cdots, x_{k_i^\star}]^{\mathsf{T}}$. From the above approximation, the LMMSE estimate for $\mathbf{x}^{(i)}$ is computed as [37]

$$\begin{aligned} &\hat{\mathbf{x}}^{(i)} \\ &= \mathbf{R}_{\mathbf{x}}^{(i)} (\mathbf{H}^{(i)})^{\mathsf{T}} \left(\mathbf{H}^{(i)} \mathbf{R}_{\mathbf{x}}^{(i)} (\mathbf{H}^{(i)})^{\mathsf{T}} + \sigma^2 \mathbf{I}_{2M} \right)^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}) + \boldsymbol{\mu}_{\mathbf{x}}^{(i)}, \end{aligned} \tag{16}$$

provided that $\mathbb{E}[(\mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{x}}^{(i)})(\mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{x}}^{(i)})^{\mathsf{T}}] = \mathbf{R}_{\mathbf{x}}^{(i)}$, $\mathbb{E}[\mathbf{x}^{(i)}] = \boldsymbol{\mu}_{\mathbf{x}}^{(i)}$, and $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}_{\mathbf{y}}$.

Unfortunately, the LMMSE estimate in (16) cannot be computed at the server since it does not have the information of the mean and the covariance of the transmitted signal. To overcome this difficulty, we model these statistics based on relevant observations and approximations. First of all, from the large-scale approximation in (13), we can approximate $|x_{k_i^\star}|^2$ as

$$|x_{k_i^\star}|^2 \approx \frac{1}{\|\mathbf{h}_{k_i^\star}\|^2} |\tilde{\mathbf{h}}_{k_i^\star}^{\mathsf{T}} \mathbf{r}_{i-1}|^2 \triangleq \alpha_{k_i^\star}. \tag{17}$$

Note that the above approximation is tight in the massive MIMO system because $\alpha_{k_i^\star}$ approaches $|x_{k_i^\star}|^2$ as the number of antennas at the server increases, as shown in (13). Since we only have the information of $|x_{k_i^\star}|^2 \approx \alpha_{k_i^\star}$, a simple, yet reasonable, model for $x_{k_i^\star}$ would be a discrete random variable with a probability mass function:

$$\mathbb{P}(x_{k_i^\star} = x) = \begin{cases} \frac{1}{2}, & x = \sqrt{\alpha_{k_i^\star}}, \\ \frac{1}{2}, & x = -\sqrt{\alpha_{k_i^\star}}. \end{cases} \tag{18}$$

Under this model, the mean and the variance of $x_{k_i^\star}$ is given by $\mathbb{E}[x_{k_i^\star}] = 0$ and $\mathbb{E}[|x_{k_i^\star}|^2] = \alpha_{k_i^\star}$, and consequently, we have $\boldsymbol{\mu}_{\mathbf{x}}^{(i)} = \mathbf{0}_i$, $\boldsymbol{\mu}_{\mathbf{y}} = \mathbf{H}^{(i)}\boldsymbol{\mu}_{\mathbf{x}}^{(i)} = \mathbf{0}_i$, and $\mathbf{R}_{\mathbf{x}}^{(i)} = \mathbb{E}[\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^{\mathsf{T}}]$. Now, recall from the discussions in Sec. III-B that the elements of the transmitted signal are statistically uncorrelated (i.e., $\mathbb{E}[x_k x_j] = 0$ for $k \neq j$ with $k, j \in \mathcal{K}$) because they are associated with different parameters by the use of the random permutation strategy. Utilizing this fact, we may assume that all the non-diagonal elements of $\mathbf{R}_{\mathbf{x}}^{(i)}$ are zero, which yields

$$\begin{aligned} \mathbf{R}_{\mathbf{x}}^{(i)} &= \mathsf{diag}\big(\mathbb{E}[|x_{k_1^\star}|^2], \ldots, \mathbb{E}[|x_{k_i^\star}|^2]\big) \\ &= \mathsf{diag}\big(\alpha_{k_1^\star}, \ldots, \alpha_{k_i^\star}\big) \triangleq \mathbf{D}_{\alpha}^{(i)}, \end{aligned} \tag{19}$$

where $\mathsf{diag}(a_1, \cdots, a_N)$ is an $N \times N$ diagonal matrix whose i-th diagonal element is a_i.

Based on the statistical model discussed above, the proposed algorithm estimates the transmitted signal at iteration i as

$$\hat{\mathbf{x}}^{(i)} = \mathbf{D}_{\alpha}^{(i)}(\mathbf{H}^{(i)})^{\mathsf{T}}\left(\mathbf{H}^{(i)}\mathbf{D}_{\alpha}^{(i)}(\mathbf{H}^{(i)})^{\mathsf{T}} + \sigma^2\mathbf{I}_{2M}\right)^{-1}\mathbf{y}. \tag{20}$$

To further reduce the computational complexity, we rewrite the estimate in (20) as

$$\hat{\mathbf{x}}^{(i)} = \mathbf{D}_{\alpha}^{(i)}(\mathbf{H}^{(i)})^{\mathsf{T}}\boldsymbol{\Omega}_i\mathbf{y}, \tag{21}$$

where

$$\boldsymbol{\Omega}_i = \left(\mathbf{H}^{(i)}\mathbf{D}_{\alpha}^{(i)}(\mathbf{H}^{(i)})^{\mathsf{T}} + \sigma^2\mathbf{I}_{2M}\right)^{-1}. \tag{22}$$

Then $\boldsymbol{\Omega}_i$ in (22) can be computed in a recursive manner:

$$\begin{aligned} \boldsymbol{\Omega}_i &= \left(\boldsymbol{\Omega}_{i-1}^{-1} + \alpha_{k_i^\star}\mathbf{h}_{k_i^\star}\mathbf{h}_{k_i^\star}^{\mathsf{T}}\right)^{-1} \\ &= \boldsymbol{\Omega}_{i-1} - \frac{\alpha_{k_i^\star}\boldsymbol{\Omega}_{i-1}\mathbf{h}_{k_i^\star}\mathbf{h}_{k_i^\star}^{\mathsf{T}}\boldsymbol{\Omega}_{i-1}}{1 + \alpha_{k_i^\star}\mathbf{h}_{k_i^\star}^{\mathsf{T}}\boldsymbol{\Omega}_{i-1}\mathbf{h}_{k_i^\star}}, \end{aligned} \tag{23}$$

where the second equality holds from the matrix inversion lemma [32]. Therefore, the proposed algorithm reduces the computational complexity required for estimating the transmitted signal, by recursively updating $\boldsymbol{\Omega}_i$ at each iteration.

4) Stopping Criterion: An ideal stopping criterion for the proposed algorithm is very difficult to derive without assuming perfect information of the true support set and the true transmitted signal at the server. As an alternative approach, we design a stopping criterion that is expected to act optimally under an ideal scenario, in which 1) all elements of the true support set are correctly selected during the first $|\tilde{\mathcal{K}}|$ iterations; and 2) the transmitted signal associated with the true support set follows the statistical model assumed in the proposed algorithm, i.e., $\mathbb{E}[x_k] = 0$ and $\mathbb{E}[|x_k|^2] = \alpha_k$ for $k \in \tilde{\mathcal{K}}$. Although this scenario is ideal, it can also be realized in a massive MIMO system because both conditions hold from (13) and (19) when the number of antennas at the server is sufficiently large.

For the ideal scenario discussed above, we design a stopping criterion by deriving an analytical threshold for the norm on the residual vector. In this scenario, the support set at iteration i belongs to one of the following cases:

- **Case 1:** The support set at iteration i is a subset of the true support set, but not equal to the true set, i.e., $\mathcal{S}_i \subset \tilde{\mathcal{K}}$ and $\mathcal{S}_i \neq \tilde{\mathcal{K}}$.
- **Case 2:** The support set at iteration i is equal to the true support set, i.e., $\mathcal{S}_i = \tilde{\mathcal{K}}$.
- **Case 3:** The support set at iteration i includes the true support set and has one more element than the true set, i.e., $\mathcal{S}_i = \tilde{\mathcal{K}} \cup \{k_i^\star\}$.

Clearly, the optimal decision for the proposed algorithm is to stop if the current set belongs to **Case 2**. Motivated by this, we derive a condition that determines the case to which the current support set belongs. To achieve this goal, we characterize the expected value of the norm squared of the residual vector in these three cases. The result of this characterization is given in the following proposition:

Proposition 1. *If the support set at iteration i belongs to* **Case** *p, the expected value of the norm squared of the current residual vector is given by $\mathbb{E}[\|\mathbf{r}_i\|^2] = E_p^{(i)}$ for $p \in \{1, 2, 3\}$, where*

$$E_1^{(i)} = \sigma^4\left[\mathsf{Tr}(\boldsymbol{\Omega}_i) + \sum_{k \in \tilde{\mathcal{K}}\setminus\mathcal{S}_i}\alpha_k\|\boldsymbol{\Omega}_i\mathbf{h}_k\|^2\right], \tag{24}$$

$$E_2^{(i)} = \sigma^4\mathsf{Tr}(\boldsymbol{\Omega}_i), \tag{25}$$

$$E_3^{(i)} = \sigma^4\left[\mathsf{Tr}(\boldsymbol{\Omega}_i) - \frac{\mathsf{Tr}(\boldsymbol{\Omega}_{i-1}) - \mathsf{Tr}(\boldsymbol{\Omega}_i)}{1 + \alpha_{k_i^\star}\mathbf{h}_{k_i^\star}^{\mathsf{T}}\boldsymbol{\Omega}_{i-1}\mathbf{h}_{k_i^\star}}\right], \tag{26}$$

provided that $\mathbb{E}[x_k] = 0$ and $\mathbb{E}[|x_k|^2] = \alpha_k$ for $k \in \tilde{\mathcal{K}}$.

Proof: See Appendix A. ∎

Proposition 1 shows that $\mathbb{E}[\|\mathbf{r}_i\|^2]$ decreases as the algorithm proceeds, while $E_1^{(i)} \geq E_2^{(i)} \geq E_3^{(i)}$. Therefore, the current support set is expected to belong to **Case 3** if the norm squared of the residual vector is closer to $E_3^{(i)}$ than to $E_2^{(i)}$. Utilizing this observation, we set the stopping criterion of the proposed algorithm as $\|\mathbf{r}_i\|^2 \leq E_{\text{th}}^{(i)}$, where

$$\begin{aligned} E_{\text{th}}^{(i)} &= \frac{1}{2}\big(E_2^{(i)} + E_3^{(i)}\big) \\ &= \sigma^4\left[\mathsf{Tr}(\boldsymbol{\Omega}_i) - \frac{\mathsf{Tr}(\boldsymbol{\Omega}_{i-1}) - \mathsf{Tr}(\boldsymbol{\Omega}_i)}{2\big(1 + \alpha_{k_i^\star}\mathbf{h}_{k_i^\star}^{\mathsf{T}}\boldsymbol{\Omega}_{i-1}\mathbf{h}_{k_i^\star}\big)}\right]. \end{aligned} \tag{27}$$

The above criterion checks whether the support set at iteration i belongs to **Case 3** or not. This is equivalent to checking whether the support set at the previous iteration $i-1$ belongs to **Case 2** or not. Therefore, after the algorithm stops by satisfying $\|\mathbf{r}_i\|^2 \leq E_{\text{th}}^{(i)}$, the final estimate of the transmitted signal is set as the previous estimate obtained at iteration $i-1$, instead of the current estimate.

5) Asymptotic Performance: To provide insight into the convergence rate of federated learning with the proposed algorithm, we characterize the asymptotic behavior of the local gradient vector estimated by the proposed algorithm. The result is given in the following proposition:

Proposition 2. *When employing the proposed algorithm in Sec. III-C, the local gradient vector reconstructed at the server satisfies*

$$\hat{\mathbf{g}}_k[t] \to \mathbf{g}_k[t] \quad as \quad M, \rho_k[t,n] \to \infty, \tag{28}$$

$\forall k \in \mathcal{K}, n \in \{1,\ldots,N_{\mathrm{w}}\}$, *where* $\rho_k[t,n] = \frac{\|\mathbf{h}_k[t,n]\|^2 |x_k[t,n]|^2}{\sigma^2}$.

Proof: See Appendix B. ∎

Proposition 2 implies that the proposed algorithm allows the server to perfectly reconstruct all the local gradient vectors when the number of antennas at the server is sufficiently large and the effective SNR is infinite for all the devices. In this asymptotic regime, federated learning over massive MIMO communication systems achieves the same convergence rate as centralized learning, provided that the set of training data samples utilized at each iteration of the training algorithm is the same for both learning methods.

Algorithm 1 Federated learning over a massive MIMO system with the presented compressive sensing approach.

Initialize the weight vector $\mathbf{w}_1$.
for $t = 1$ to T **do**
  At the server:
    Transmit $\mathbf{w}_t$ to M wireless devices.
  At device $k \in \mathcal{K}$*:*
    Compute $\mathbf{g}_k[t]$ from (1).
    Compute $\mathbf{x}_k[t]$ from (9).
    Transmit $\mathbf{x}_k[t]$ to the server.
  At the server:
  for $n = 1$ to N_{w} **do**
    Set $\mathbf{h}_k = \mathbf{h}_k[t,n]$ and $\tilde{\mathbf{h}}_k = \frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}$ for $k \in \mathcal{K}$.
    Set $\mathcal{S}_0 = \emptyset$, $\mathbf{r}_0 = \mathbf{y}[t,n]$, and $\mathbf{\Omega}_0 = \frac{1}{\sigma^2}\mathbf{I}_{2M}$.
    for $i = 1$ to K **do**
      Find $k_i^\star = \mathrm{argmax}_{k \in \mathcal{S}_{i-1}^c} \left|\tilde{\mathbf{h}}_k^{\mathsf{T}} \mathbf{r}_{i-1}\right|^2$.
      Set $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{k_i^\star\}$ and $\alpha_{k_i^\star} = \frac{1}{\|\mathbf{h}_{k_i^\star}\|^2}\left|\tilde{\mathbf{h}}_{k_i^\star}^{\mathsf{T}} \mathbf{r}_{i-1}\right|^2$.
      Compute $\mathbf{\Omega}_i$ from (23).
      Set $\mathbf{r}_i = \sigma^2 \mathbf{\Omega}_i \mathbf{y}[t,n]$.
      Compute $E_{\mathrm{th}}^{(i)}$ from (27).
      if $\|\mathbf{r}_i\|^2 < E_{\mathrm{th}}^{(i)}$ **then**
        Update $I^\star = i - 1$.
        Break the loop.
      end if
    end for
    Compute $\hat{\mathbf{x}}^{(I^\star)} = \mathbf{D}_\alpha^{(I^\star)} (\mathbf{H}^{(I^\star)})^{\mathsf{T}} \mathbf{\Omega}_{I^\star} \mathbf{y}[t,n]$.
    Set $\hat{x}_{k_i^\star}[t,n] = \hat{x}_i^{(I^\star)}$ for $i \in \{1,\ldots,I^\star\}$.
    Set $\hat{x}_k[t,n] = 0$ for $m \notin \mathcal{S}_i$.
  end for
  Compute $\hat{\mathbf{g}}_k[t]$ from (10) for $k \in \mathcal{K}$.
  Compute $\bar{\mathbf{g}}[t]$ from (7).
  Update $\mathbf{w}_{t+1}$ based on $\bar{\mathbf{g}}[t]$.
end for

6) Summary: In Algorithm 1, we summarize the overall process of federated learning over the massive MIMO system when employing the presented compressive sensing approach. In this algorithm, Steps 3∼4 and Steps 5∼30 are associated with the downlink and uplink phases, respectively. Note that Steps 11∼26 may change according to the reception strategy adopted by the central server. In Step 17, the residual vector $\mathbf{r}_i$ is determined using the following equality:

$$\mathbf{r}_i = \mathbf{y} - \mathbf{H}^{(i)} \hat{\mathbf{x}}^{(i)} = \left(\mathbf{I}_{2M} - \mathbf{H}^{(i)} \mathbf{R}_{\mathbf{x}}^{(i)} (\mathbf{H}^{(i)})^{\mathsf{T}} \mathbf{\Omega}_i\right) \mathbf{y} = \sigma^2 \mathbf{\Omega}_i \mathbf{y}, \tag{29}$$

where the second equality is obtained from (21). By using the expression in (29), computing the estimated transmitted signal $\hat{\mathbf{x}}^{(i)}$ is not required at each iteration. Instead, $\hat{\mathbf{x}}^{(i)}$ is computed only once after the proposed algorithm ends (see Step 23). Using this strategy, the overall computational complexity of the proposed algorithm is further reduced. As can be seen in Algorithm 1, the proposed algorithm is a greedy algorithm whose optimality is not guaranteed in general. Nevertheless, the numerical results in Sec. V illustrate that the performance gap between federated learning with the proposed algorithm and centralized learning with perfect reconstruction is marginal under certain scenarios.

Remark (*Comparison to Conventional Orthogonal Matching Pursuit Algorithms*): The proposed algorithm can be regarded as a variant of the orthogonal matching pursuit (OMP) algorithm in compressive sensing. The common feature of the proposed algorithm and the conventional OMP algorithms in [33]–[36] is that a support set is iteratively updated by adding an index with the maximum correlation to the residual vector. Despite this fact, the proposed algorithm still differs from the algorithms in [33]–[36]. The major differences between the proposed and the conventional OMP algorithms are summarized as follows:

- The proposed algorithm utilizes the LMMSE estimate that generalizes the least-squares (LS) estimate considered in the conventional OMP algorithms in [33], [34]. It is well-known that the LMMSE estimate is superior to the LS estimate because the LMMSE estimate is optimal in terms of the MSE. Although the LMMSE estimate is also considered in [35], [36], the LMMSE estimate in the proposed algorithm utilizes a different statistical model for the transmitted signal, established based on both a large-scale approximation and a statistical feature obtained from the random permutation strategy.
- The stopping criterion of the proposed algorithm differs from those of the conventional OMP algorithms. Particularly, the criterion of the proposed algorithm is uniquely designed by deriving an analytical threshold for the residual error at each iteration, based on the statistical model established for the transmitted signal in federated learning.

IV. Comparison to A Linear Beamforming Approach

Another possible solution for reconstructing the local gradient vectors at the central server is to apply conventional linear beamforming methods developed to solve a MIMO data detection problem [37]. Motivated by this, we introduce two

linear beamforming methods as performance benchmarks for the proposed algorithm. We also discuss the limitation of each method for the use in federated learning over a massive MIMO system. We then compare the computational complexity of the proposed algorithm with those of the linear beamforming methods.

A. Limitation of Maximal Ratio Combining (MRC)

The simplest yet effective linear beamforming method is the MRC method which aims to maximize the power of the desired signal by aligning the direction of the receive beamforming into the channel direction. The estimate of the transmitted signal is given by

$$\hat{\mathbf{x}}_{\mathrm{MRC}}[t,n] = \mathsf{diag}\left(\frac{1}{\|\mathbf{h}_1[t,n]\|^2}, \ \cdots \ , \frac{1}{\|\mathbf{h}_K[t,n]\|^2}\right)\mathbf{H}^{\mathsf{T}}[t,n]\mathbf{y}[t,n]. \tag{30}$$

Consequently, the estimate of the local gradient vector is given by $\hat{\mathbf{g}}_k^{\mathrm{MRC}}[t] = \sqrt{\frac{\|\mathbf{g}_k[t]\|^2}{N_{\mathrm{w}}}}\hat{\mathbf{x}}_k^{\mathrm{MRC}}[t]$, where $\hat{\mathbf{x}}_k^{\mathrm{MRC}}[t] = \left[\hat{x}_k^{\mathrm{MRC}}[t,1], \cdots, \hat{x}_k^{\mathrm{MRC}}[t,N_{\mathrm{w}}]\right]^{\mathsf{T}}$ and $\hat{x}_k^{\mathrm{MRC}}[t,n]$ is the k-th element of $\hat{\mathbf{x}}_{\mathrm{MRC}}[t,n]$.

To highlight a limitation of the MRC method, we take a closer look at the k-th element of the estimated signal $\hat{\mathbf{x}}_{\mathrm{MRC}}[t,n]$ given by

$$\hat{x}_{\mathrm{MRC},k}[t,n] = x_k[t,n] + \sum_{j\neq k}\frac{\mathbf{h}_k^{\mathsf{T}}[t,n]\mathbf{h}_j[t,n]}{\|\mathbf{h}_k[t,n]\|^2}x_j[t,n] + \frac{\mathbf{h}_k^{\mathsf{T}}[t,n]\mathbf{z}[t,n]}{\|\mathbf{h}_k[t,n]\|^2}. \tag{31}$$

As can be seen in (31), the estimate in the MRC method consists of not only the desired signal $x_k[t,n]$, but also an inter-user interference (IUI) and an effective noise corresponding to the second and the third terms in the RHS of (31), respectively. Although both IUI and noise terms vanish as the number of antennas at the server goes to infinity by the central limit theorem [38], this does not hold for a general number of antennas at the server. Therefore, in most practical scenarios, the MRC method is suboptimal in terms of the estimation performance as will be shown in Sec. V.

B. Limitation of Linear Minimum Mean Square Error (LMMSE)

The LMMSE method is an optimal linear beamforming method that minimizes the MSE between the true transmitted signal and its estimate. The estimate of the transmitted signal in the LMMSE method is given by [37]

$$\hat{\mathbf{x}}_{\mathrm{LMMSE}}[t,n] = \mathbf{F}_{\mathrm{LMMSE}}[t,n]\left(\mathbf{y}[t,n] - \boldsymbol{\mu}_{\mathbf{y}}[t,n]\right) + \boldsymbol{\mu}_{\mathbf{x}}[t,n], \tag{32}$$

where

$$\mathbf{F}_{\mathrm{LMMSE}}[t,n] = \mathbf{R}_{\mathbf{x}}[t,n]\mathbf{H}^{\mathsf{T}}[t,n] \times \left(\mathbf{H}[t,n]\mathbf{R}_{\mathbf{x}}[t,n]\mathbf{H}^{\mathsf{T}}[t,n] + \sigma^2\mathbf{I}_{2M}\right)^{-1}, \tag{33}$$

where $\boldsymbol{\mu}_{\mathbf{x}}[t,n] = \mathbb{E}[\mathbf{x}[t,n]]$, $\mathbf{R}_{\mathbf{x}}[t,n] = \mathbb{E}[\mathbf{x}[t,n]\mathbf{x}^{\mathsf{T}}[t,n]]$, and $\boldsymbol{\mu}_{\mathbf{y}}[t,n] = \mathbb{E}[\mathbf{y}[t,n]]$. Consequently, the estimate of the local gradient vector is given by $\hat{\mathbf{g}}_k^{\mathrm{LMMSE}}[t] = \sqrt{\frac{\|\mathbf{g}_k[t]\|^2}{N_{\mathrm{w}}}}\hat{\mathbf{x}}_k^{\mathrm{LMMSE}}[t]$, where $\hat{\mathbf{x}}_k^{\mathrm{LMMSE}}[t] = \left[\hat{x}_k^{\mathrm{LMMSE}}[t,1], \cdots, \hat{x}_k^{\mathrm{LMMSE}}[t,N_{\mathrm{w}}]\right]^{\mathsf{T}}$ and $\hat{x}_k^{\mathrm{LMMSE}}[t,n]$ is the k-th element of $\hat{\mathbf{x}}_{\mathrm{LMMSE}}[t,n]$.

A major limitation of the LMMSE method is that it is applicable only when both the mean $\mathbb{E}[\mathbf{x}[t,n]]$ and the covariance $\mathbf{R}_{\mathbf{x}}[t,n]$ of the transmitted signal are known at the server. Characterizing the statistical behavior of the gradient vector is very challenging in the most learning tasks, which is due to the randomness and the heterogeneity of real-world data. For this reason, the statistics of the transmitted signal are generally unknown at the server, and consequently, the LMMSE method suffers from performance degradation caused by imperfect statistical information. Another limitation of the LMMSE method is that it requires the computation of a $2M \times 2M$ matrix inversion to determine the beamforming matrix in (33). Therefore, the computational complexity required in this method may not be affordable in the massive MIMO system with a large number of antennas at the server, as will be discussed in detail in the following subsection.

C. Comparison of Computational Complexity

We analyze and compare the computational complexity of the proposed algorithm in Sec. III-C, the MRC method in Sec. IV-A, and the LMMSE method in Sec. IV-B. To this end, we count the number of real multiplications required to compute Steps 13–24 in Algorithm 1 for the proposed algorithm, (30) for the MRC method, and (32) for the LMMSE method. Particularly for the LMMSE method, we present the minimum complexity by considering a recursive computation of matrix inversion, as done in (23), under the assumptions of $\boldsymbol{\mu}_{\mathbf{x}}[t,n] = \mathbf{0}_K$ and $\mathbf{R}_{\mathbf{x}}[t,n] = \mathbf{I}_K$. The complexity results for three methods are summarized in Table I, where $I^\star$ is the size of the support set determined by the proposed algorithm (see Algorithm 1). In Table I, we also present the complexity for a large-scale scenario (i.e., $M \gg 1$ and $K \gg 1$) which is the region of interest in massive MIMO systems.

Table I shows that the proposed algorithm has a significantly lower complexity compared to the LMMSE method, when the transmitted signal is very sparse. More precisely, the ratio of the complexity of the proposed algorithm to that of the LMMSE method is obtained as

$$\frac{C_{\mathrm{Pro}}^{\mathrm{large}}}{C_{\mathrm{LMMSE}}^{\mathrm{large}}} = \frac{(I^\star)^2}{8KM} + \frac{3I^\star}{2K} + \frac{I^\star}{4M}, \tag{34}$$

for $M \gg 1$ and $K \gg 1$. If the proposed algorithm properly stops with $I^\star = |\tilde{\mathcal{K}}|$, the complexity ratio in (34) becomes

$$\frac{C_{\mathrm{Pro}}^{\mathrm{large}}}{C_{\mathrm{LMMSE}}^{\mathrm{large}}} = \frac{|\tilde{\mathcal{K}}|}{K}\left(\frac{|\tilde{\mathcal{K}}|}{8M} + \frac{3}{2}\right) + \frac{|\tilde{\mathcal{K}}|}{4M} \to \frac{|\tilde{\mathcal{K}}|}{4M} \quad \text{as} \quad \frac{|\tilde{\mathcal{K}}|}{K} \to 0, \tag{35}$$

for a fixed M. The above result implies that if the size of the true support set is much smaller than the number of devices, the complexity reduction achieved by the proposed algorithm over the LMMSE method increases with the number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.3038407, IEEE Transactions on Wireless Communications

TABLE I
THE NUMBER OF REAL MULTIPLICATIONS REQUIRED FOR VARIOUS ESTIMATION METHODS.

Method	General case	Large-scale case ($M \gg 1$ and $K \gg 1$)
Proposed	$(I^\star)^2(M+\frac{1}{2})+I^\star(12M^2+2KM+11M+K+\frac{17}{2})$ $+12M^2+4KM+10M+2K+7$	$(I^\star)^2M+I^\star(12M^2+2KM)$ $+12M^2+4KM$
MRC	$4KM+K$	$4KM$
LMMSE	$8KM^2-4M^2+6KM+2M$	$8KM^2$

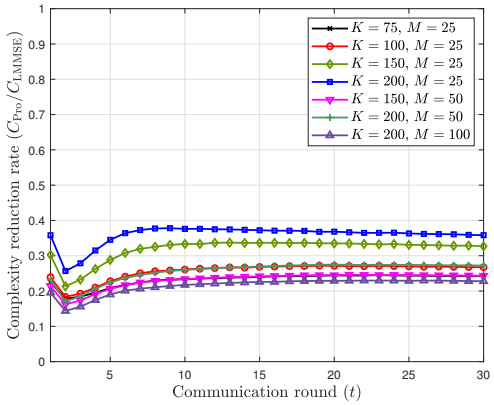


Fig. 4. The complexity ratio of the proposed algorithm to the LMMSE method for various K and M.

of antennas at the server. Therefore, in federated learning over the massive MIMO system, the proposed algorithm is significantly more beneficial than the LMMSE method in terms of the computational complexity. Note that although the MRC method achieves the lowest complexity among three methods, it suffers from a performance degradation as will be shown in Sec. V.

We also compare the computational complexity of the proposed algorithm and the LMMSE method using simulation. In this comparison, we numerically obtain the average size of the support set for the proposed algorithm, namely $I^\star_{\text{avg}}$. We then compute the ratio of the number of real multiplications of the proposed algorithm to that of the LMMSE method by using the results in Table I with $I^\star = I^\star_{\text{avg}}$. Fig. 4 illustrates the complexity ratio of the proposed algorithm to the LMMSE method for different K and M under simulation setting described in Sec. V-A, where we consider the stochastic setting ($|\mathcal{B}_k[t]| = 1$). Fig. 4 shows that the complexity reduction achieved by the proposed algorithm over the LMMSE method is more than 60% for all the cases. Furthermore, this complexity reduction is shown to be even higher for the case of $K = 200$ and $M = 100$, corresponding to a large-scale scenario. Therefore, these results demonstrate that the proposed algorithm has a significantly lower complexity compared to LMMSE for the scenarios under consideration.

V. SIMULATION RESULTS

In this section, using simulations, we evaluate the classification accuracy of federated learning over a massive MIMO system with various local gradient reconstruction approaches. The wireless channel of the communication system is modeled by 10-tap CIR that follows uniform power delay profile, in which each CIR tap is distributed as $\mathcal{CN}(0, 0.1)$. The number of subcarriers for OFDM signaling is set as $N_{\text{sub}} = 1024$, and the noise power is set as $\sigma_{\text{c}}^2 = 1$ (i.e., $\sigma^2 = 0.5$). For the implementation of the LMMSE method in Sec. IV-B, we assume[4] that $\boldsymbol{\mu}_{\mathbf{x}}[t,n] = \mathbf{0}_K$ and $\mathbf{R}_{\mathbf{x}}[t,n] = \mathbf{I}_K$.

A. Simulation Setting

In this simulation, we consider an image classification task in which a neural network is used to classify a 28×28 grayscale image of a handwritten digit (from 0 to 9) in the MNIST dataset that consists of 60000 training and 10000 test data samples [39]. We assume that a central server utilizes a neural network[5] that consists of 784 input nodes, a single hidden layer with 20 hidden nodes, and 10 output nodes. The activation functions of the hidden layer and the output layer are set as the ReLU and the softmax functions, respectively. The weights of the neural network at communication round t are mapped into a parameter vector $\mathbf{w}_t$ with length $N_{\text{w}} = 15910$. To train $\mathbf{w}_t$, we adopt the ADAM optimizer in [30] whose update rule is given by

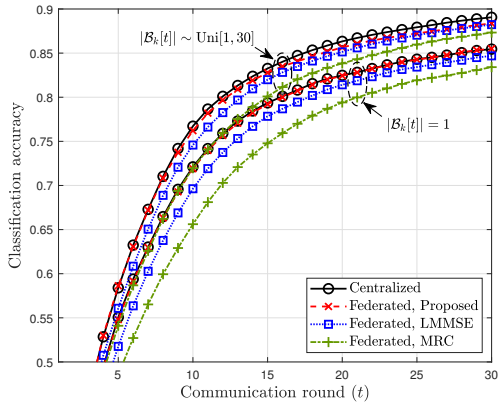
$$\begin{aligned}\mathbf{m}_{t+1} &\leftarrow \beta_1\mathbf{m}_t + (1-\beta_1)\bar{\mathbf{g}}[t],\\ \mathbf{v}_{t+1} &\leftarrow \beta_2\mathbf{v}_t + (1-\beta_2)(\bar{\mathbf{g}}[t])^2,\\ \mathbf{w}_{t+1} &\leftarrow \mathbf{w}_t - \alpha\hat{\mathbf{m}}_{t+1}/(\sqrt{\hat{\mathbf{v}}_{t+1}}+\epsilon),\end{aligned}$$

where all the operations are element-wise, $\hat{\mathbf{m}}_t = \mathbf{m}_t/(1-\beta_1^t)$, and $\hat{\mathbf{v}}_t = \mathbf{v}_t/(1-\beta_2^t)$. For this, we set $\mathbf{m}_1 = \mathbf{0}_{N_{\text{w}}}$, $\mathbf{v}_1 = \mathbf{0}_{N_{\text{w}}}$, $\alpha = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The global gradient vector $\bar{\mathbf{g}}[t]$ required to compute the update rule is the function of the local gradient vectors computed-and-sent by the wireless devices, as can be seen in (7). For this, we assume that the central server adopts one of the local gradient reconstruction approaches introduced in our work. The local gradient vector at each wireless device is computed according to (1) using the cross-entropy loss function. For the local data set $\mathcal{B}_k$ of device

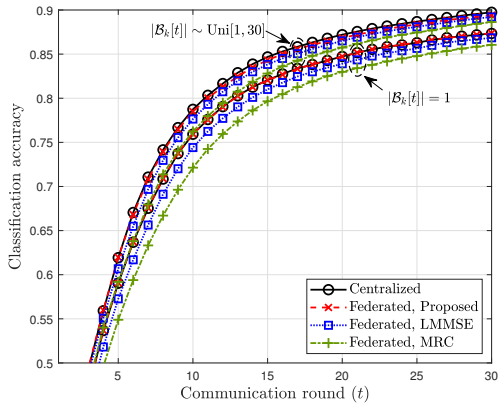
[4]The assumption of $\boldsymbol{\mu}_{\mathbf{x}}[t,n] = \mathbf{0}_K$ is employed because we do not have any prior information about the true mean. Also, the assumption of $\mathbb{E}[\mathbf{x}[t,n]\mathbf{x}^{\text{T}}[t,n]] = \mathbf{I}_K$ is employed because the definition of $\mathbf{x}_k[t]$ in (2) implies that $\mathbb{E}[|x_k[t,n]|^2] = 1$ if the power of the local gradient vector $\mathbf{g}_k$ is uniformly distributed across all its elements.

[5]Recall that the amount of radio resources required to convey the model parameter vector is proportional to the number of the weights of the neural network. Considering this fact, we adopt a relatively simple neural-network structure as it is more suitable for the use in practical communication systems whose radio resources are scarce. In particular, the number of hidden nodes is chosen as 20 which provides a fair performance-complexity tradeoff for the considered classification task. Nevertheless, the proposed algorithm can be applied to a more complex neural network (e.g., a convolutional neural network) to further improve the classification accuracy.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.3038407, IEEE Transactions on Wireless Communications



(a) $K = 75$ and $M = 25$



(b) $K = 150$ and $M = 50$

Fig. 5. Classification accuracies of federated learning with various local gradient reconstruction approaches for different K and M.

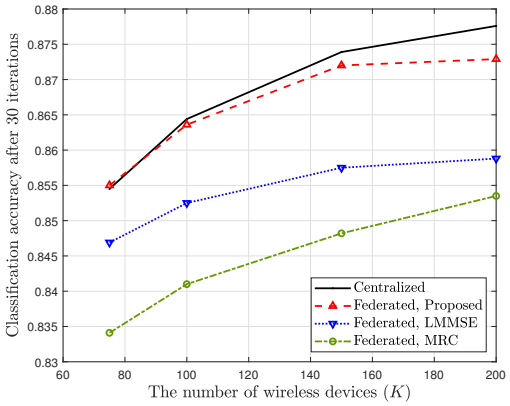


Fig. 6. The impact of the number of wireless devices K on the classification accuracies after $T = 30$ iterations for federated learning with various local gradient reconstruction approaches when $M = 25$ and $|\mathcal{B}_k[t]| = 1$.

k, we select the set of 1000 training data samples at random among the training samples labeled with digit $d_k = \lfloor \frac{k-1}{K/10} \rfloor$ in the MNIST dataset. This corresponds to a *non-IID* setting because each device has the information of only one digit. Then device k randomly selects $|\mathcal{B}_k[t]|$ samples from $\mathcal{B}_k$ to compute the local gradient vector at communication round t. To determine the batch size at each device, we consider two settings: 1) a *stochastic* setting with $|\mathcal{B}_k[t]| = 1$, and 2) a *mini-batch* setting with $|\mathcal{B}_k[t]| \sim \text{Uni}[1, 30]$ where $|\mathcal{B}_k[t]|$ is randomly drawn from a uniform distribution over $[1, 30]$, for all $k \in \mathcal{K}$ and $t \in \mathcal{T}$.

B. Classification Accuracy Results

Fig. 5 compares the classification accuracies of federated learning with various local gradient reconstruction approaches for different K and M. As a performance benchmark, we also plot the accuracy achieved by centralized learning with perfect reconstruction of the local gradient vectors at the central server in which $\hat{\mathbf{g}}_k[t] = \mathbf{g}_k[t]$ for all $k \in \mathcal{K}$ and $t \in \mathcal{T}$. Both Figs. 5(a) and 5(b) show that the proposed compressive sensing approach outperforms all the linear beamforming methods in terms of the classification accuracy and the convergence rate regardless of the size of the mini-batch data samples used for computing the local gradient vector at each device. In other words, the proposed approach requires a less number of communication rounds than the linear beamforming methods, to achieve the same level of the classification accuracy. It is also shown that the performance gap between federated learning with the proposed approach and the centralized learning is marginal for the stochastic setting ($|\mathcal{B}_k[t]| = 1$). This result demonstrates that our compressive sensing approach effectively compensates for the performance-degrading factors in wireless communications including IUI, channel fading, and noise effects when the size of the mini-batch samples employed at each device is small. When referring to Figs. 4 and 5 together, it can also be observed that the proposed approach is superior to the LMMSE method in terms of both the classification accuracy and the computational complexity.

Fig. 6 evaluates the impact of the number of wireless devices, K, on the classification accuracies after $T = 30$ iterations for federated learning with various local gradient reconstruction approaches when $M = 25$ and $|\mathcal{B}_k[t]| = 1$. Fig. 6 shows that the classification accuracy of all the considered approaches improves with the number of the wireless devices. The intuition behind this result is that increasing the number of the devices leads to an increase in the number of training samples utilized to train the neural network at each communication round. It is also shown that the performance gap between centralized learning and federated learning increases with K. The major reason is that the number of unknown values that need to be estimated at the server increases with K while the number of the available observations (i.e., the number of antennas at the server) is fixed. Nevertheless, the proposed compressive sensing approach outperforms all the linear beamforming methods regardless of the number of the devices. This result demonstrates that the performance efficiency of the proposed approach is not degraded by the number of wireless devices participating in the training.

Fig. 7 evaluates the impact of the number of antennas at the server, M, on the classification accuracies after $T = 30$ iterations for federated learning with various local gradient reconstruction approaches when $K = 200$ and $|\mathcal{B}_k[t]| = 1$. Fig. 7 shows that the classification accuracies of all the considered approaches improves with the number of antennas at

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2020.3038407, IEEE Transactions on Wireless Communications

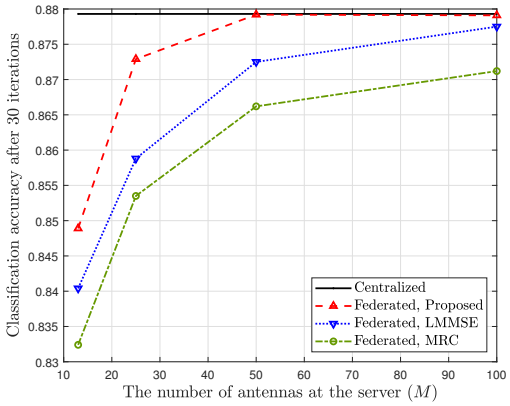


Fig. 7. The impact of the number of antennas at the server, M, on the classification accuracies after $T = 30$ iterations for federated learning with various local gradient reconstruction approaches when $K = 200$ and $|\mathcal{B}_k[t]| = 1$.

the server. This performance improvement is the consequence of exploiting the receive diversity that provides the robustness to the performance-degrading factors in wireless communications. Another important observation is that the proposed compressive sensing approach requires a significantly less number of antennas at the server than the linear beamforming methods, to achieve the same level of classification accuracy. For example, to achieve 87% classification accuracy, for the proposed approach, the central server requires only 23 antennas, while for LMMSE and MRC, the server should be equipped with 45 and 90 antennas, respectively. Therefore, it is demonstrated that the proposed approach also contributes to reduce the hardware requirement at the server to achieve the desired level of accuracy in federated learning.

VI. Conclusion

In this paper, we have presented a compressive sensing approach for federated learning over a massive MIMO system, which allows a central server to efficiently reconstruct local gradient vectors sent from wireless devices. In particular, motivated by the sparsity of the local gradient vectors, we have established a proper transmission strategy to construct a sparse transmitted signal that aggregates the signals sent from all the wireless devices at each radio resource. Based on this transmission strategy, we have proposed a compressive sensing algorithm enabling the central server to iteratively find the LMMSE estimate of the transmitted signal. We have also analyzed the computational complexity of the proposed algorithm and demonstrated that when the transmitted signal is sparse, the proposed algorithm requires a significantly lower complexity compared to the LMMSE method. Using simulations, we have demonstrated that the presented approach outperforms the linear beamforming approaches in terms of the accuracy, while reducing the performance gap between federated learning and centralized learning with perfect reconstruction.

An important direction for future research is to extend our compressive sensing approach by developing a proper device scheduling algorithm. In this extension, a joint optimization of device scheduling and local gradient reconstruction may further improve the performance of federated learning, particularly when the number of wireless devices is much larger than the number of antennas at the server. Another important research direction is to develop a proper downlink transmission strategy for broadcasting the parameter vector to the wireless devices and investigate its impact on the performance of federated learning over a massive MIMO system. It would also be important to analyze the convergence rate of federated learning with the proposed approach for a general number of antennas at the server operating at finite SNR. To this end, it should be possible to apply or extend the techniques introduced in [40].

Appendix A
Proof of Proposition 1

Suppose that $\mathcal{S}_{i-1} \subset \tilde{\mathcal{K}}$, $\mathbb{E}[x_k] = 0$, and $\mathbb{E}[|x_k|^2] = \alpha_k$ for $k \in \tilde{\mathcal{K}}$. Then the norm squared of the residual vector at iteration i is expressed as

$$\mathbb{E}[\|\mathbf{r}_i\|^2] = \mathsf{Tr}\big(\mathbb{E}[\mathbf{r}_i\mathbf{r}_i^{\mathsf{T}}]\big) = \sigma^4\mathsf{Tr}\big(\mathbf{\Omega}_i\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^{\mathsf{T}}]\mathbf{\Omega}_i\big), \tag{36}$$

where the second equality is obtained from (29). Thanks to the use of the permutation matrix in (9), we can ensure that $\mathbb{E}[x_k x_j] = 0$ for $k \neq j$ with $k, j \in \mathcal{K}$. Utilizing this fact, the covariance of the received signal is obtained as

$$\mathbb{E}[\mathbf{y}\mathbf{y}^{\mathsf{T}}] = \sum_{k\in\tilde{\mathcal{K}}} \alpha_k \mathbf{h}_k\mathbf{h}_k^{\mathsf{T}} + \sigma^2\mathbf{I}_{2M}. \tag{37}$$

In what follows, we characterize $\mathbf{\Omega}_i$ and then provide a closed-form expression for $\mathbb{E}[\|\mathbf{r}_i\|^2]$ for three cases discussed in Sec. III-C.

A. Case 1: $\mathcal{S}_i \subset \tilde{\mathcal{K}}$ and $\mathcal{S}_i \neq \tilde{\mathcal{K}}$

In this case, $\mathbf{\Omega}_i$ in (22) is expressed as

$$\mathbf{\Omega}_i = \left(\sum_{k\in\mathcal{S}_i} \alpha_k \mathbf{h}_k\mathbf{h}_k^{\mathsf{T}} + \sigma^2\mathbf{I}_{2M}\right)^{-1}. \tag{38}$$

From (38), the covariance of the received signal in (37) is rewritten as

$$\mathbb{E}[\mathbf{y}\mathbf{y}^{\mathsf{T}}] = \mathbf{\Omega}_i^{-1} + \sum_{k\in\tilde{\mathcal{K}}\setminus\mathcal{S}_i} \alpha_k \mathbf{h}_k\mathbf{h}_k^{\mathsf{T}}. \tag{39}$$

Then the norm squared of the residual vector when the support set belongs to **Case 1** is obtained by applying (39) into (36):

$$E_1^{(i)} = \sigma^4\left[\mathsf{Tr}(\mathbf{\Omega}_i) + \sum_{k\in\tilde{\mathcal{K}}\setminus\mathcal{S}_i} \alpha_k \|\mathbf{\Omega}_i\mathbf{h}_k\|^2\right]. \tag{40}$$

B. Case 2: $\mathcal{S}_i = \tilde{\mathcal{K}}$

In this case, $\mathbf{\Omega}_i$ in (22) is expressed as

$$\mathbf{\Omega}_i = \left(\sum_{k\in\tilde{\mathcal{K}}} \alpha_k \mathbf{h}_k\mathbf{h}_k^{\mathsf{T}} + \sigma^2\mathbf{I}_{2M}\right)^{-1}. \tag{41}$$

From (41), the covariance of the received signal in (37) is rewritten as $\mathbb{E}[\mathbf{y}\mathbf{y}^{\mathsf{T}}] = \mathbf{\Omega}_i^{-1}$. By applying this result into (36), the norm squared of the residual vector when the support set belongs to **Case 2** is given by $E_2^{(i)} = \sigma^4\mathsf{Tr}(\mathbf{\Omega}_i)$.

C. Case 3: $\mathcal{S}_i = \tilde{\mathcal{K}} \cup \{k_i^\star\}$

In this case, $\mathbf{\Omega}_i$ in (22) is expressed as

$$\mathbf{\Omega}_i = \left(\sum_{k \in \tilde{\mathcal{K}}} \alpha_k \mathbf{h}_k \mathbf{h}_k^\mathsf{T} + \alpha_{k_i^\star} \mathbf{h}_{k_i^\star} \mathbf{h}_{k_i^\star}^\mathsf{T} + \sigma^2 \mathbf{I}_{2M} \right)^{-1}. \tag{42}$$

From (42), the covariance of the received signal in (37) is rewritten as

$$\mathbb{E}[\mathbf{y}\mathbf{y}^\mathsf{T}] = \mathbf{\Omega}_i^{-1} - \alpha_{k_i^\star} \mathbf{h}_{k_i^\star} \mathbf{h}_{k_i^\star}^\mathsf{T}. \tag{43}$$

Applying (43) into (36) yields

$$E_3^{(i)} = \sigma^4 \left[\mathsf{Tr}(\mathbf{\Omega}_i) - \alpha_{k_i^\star} \|\mathbf{\Omega}_i \mathbf{h}_{k_i^\star}\|^2 \right]. \tag{44}$$

From (23), the second term in the RHS of (44) is expressed as

$$\alpha_{k_i^\star} \|\mathbf{\Omega}_i \mathbf{h}_{k_i^\star}\|^2 = \frac{\alpha_{k_i^\star} \|\mathbf{\Omega}_{i-1} \mathbf{h}_{k_i^\star}\|^2}{(1 + \alpha_{k_i^\star} \mathbf{h}_{k_i^\star}^\mathsf{T} \mathbf{\Omega}_{i-1} \mathbf{h}_{k_i^\star})^2}. \tag{45}$$

Also, applying the trace function to (23) yields

$$\mathsf{Tr}(\mathbf{\Omega}_i) = \mathsf{Tr}(\mathbf{\Omega}_{i-1}) - \frac{\alpha_{k_i^\star} \|\mathbf{\Omega}_{i-1} \mathbf{h}_{k_i^\star}\|^2}{1 + \alpha_{k_i^\star} \mathbf{h}_{k_i^\star}^\mathsf{T} \mathbf{\Omega}_{i-1} \mathbf{h}_{k_i^\star}}, \tag{46}$$

so we have

$$\alpha_{k_i^\star} \|\mathbf{\Omega}_{i-1} \mathbf{h}_{k_i^\star}\|^2 = \frac{\mathsf{Tr}(\mathbf{\Omega}_{i-1}) - \mathsf{Tr}(\mathbf{\Omega}_i)}{1 + \alpha_{k_i^\star} \mathbf{h}_{k_i^\star}^\mathsf{T} \mathbf{\Omega}_{i-1} \mathbf{h}_{k_i^\star}}. \tag{47}$$

By applying (47) into (44), the norm squared of the residual vector when the support set belongs to **Case 3** is given by

$$E_3^{(i)} = \sigma^4 \left[\mathsf{Tr}(\mathbf{\Omega}_i) - \frac{\mathsf{Tr}(\mathbf{\Omega}_{i-1}) - \mathsf{Tr}(\mathbf{\Omega}_i)}{1 + \alpha_{k_i^\star} \mathbf{h}_{k_i^\star}^\mathsf{T} \mathbf{\Omega}_{i-1} \mathbf{h}_{k_i^\star}} \right]. \tag{48}$$

Appendix B
Proof of Proposition 2

From (9) and (10), it can be easily shown that if $\hat{x}_k[t,n] \to x_k[t,n]$, $\forall k \in \tilde{\mathcal{K}}[t,n] \triangleq \{x_k[t,n] \neq 0 \mid k \in \mathcal{K}\}, n \in \{1,\ldots,N_\mathrm{w}\}$, we have $\hat{\mathbf{g}}_k[t] \to \mathbf{g}_k[t]$, $\forall k \in \mathcal{K}$. Considering this fact, we derive the result in (28) by proving that

$$\hat{x}_k[t,n] \to x_k[t,n] \quad \text{as} \quad M, \rho_k[t,n] \to \infty, \tag{49}$$

for $\forall k \in \tilde{\mathcal{K}}[t,n], n \in \{1,\ldots,N_\mathrm{w}\}$. In the rest part of the proof, we omit the indices t and n to simplify the notation.

Let $\tilde{k}_i \in \tilde{\mathcal{K}}$ be the index with the i-th largest effective SNR, i.e., $\rho_{\tilde{k}_1} \geq \ldots, \geq \rho_{\tilde{k}_{I^\star}}$, where $I^\star = |\tilde{\mathcal{K}}|$. Since the approximation in (13) becomes tighter as M increases, the expressions in (11) and (17) imply that $k_i^\star \to \tilde{k}_i$ and $\alpha_{k_i^\star} \to |x_{k_i^\star}|^2$ as $M \to \infty$, $\forall i \in \{1,\ldots,I^\star\}$. Therefore, for sufficiently large M, the proposed algorithm correctly finds all the elements in $\tilde{K}$ during the first $I^\star$ iterations. This fact also implies that

$$\mathbf{y} = \sum_{k \in \tilde{\mathcal{K}}} \mathbf{h}_k x_k + \mathbf{z} = \mathbf{H}^{(I^\star)} \mathbf{x}^{(I^\star)} + \mathbf{z}. \tag{50}$$

From (20), the estimate of the transmitted signal at iteration $I^\star$ can be expressed as

$$\begin{aligned} \hat{\mathbf{x}}^{(I^\star)} &= \mathbf{D}_\alpha^{(I^\star)} (\mathbf{H}^{(I^\star)})^\mathsf{T} \left(\mathbf{H}^{(I^\star)} \mathbf{D}_\alpha^{(I^\star)} (\mathbf{H}^{(I^\star)})^\mathsf{T} + \sigma^2 \mathbf{I}_{2M} \right)^{-1} \mathbf{y} \\ &= \left(\sigma^2 (\mathbf{D}_\alpha^{(I^\star)})^{-1} + (\mathbf{H}^{(I^\star)})^\mathsf{T} \mathbf{H}^{(I^\star)} \right)^{-1} (\mathbf{H}^{(I^\star)})^\mathsf{T} \mathbf{y}, \end{aligned} \tag{51}$$

where the second equality holds from the matrix inversion lemma [32]. From (50) and the definition of $\rho_k = \|\mathbf{h}_k\|^2 |x_k|^2 / \sigma^2$, we have

$$\hat{\mathbf{x}}^{(I^\star)} \to \left((\mathbf{H}^{(I^\star)})^\mathsf{T} \mathbf{H}^{(I^\star)} \right)^{-1} (\mathbf{H}^{(I^\star)})^\mathsf{T} \mathbf{H}^{(I^\star)} \mathbf{x}^{(I^\star)} = \mathbf{x}^{(I^\star)}, \tag{52}$$

as $\rho_k \to \infty$, $\forall k \in \tilde{\mathcal{K}}$, provided that $M \geq I^\star$. In this regime, the norm of the residual error is given by $\|\mathbf{r}_{I^\star}\|^2 = \|\mathbf{y} - \mathbf{H}^{(I^\star)} \hat{\mathbf{x}}^{(I^\star)}\|^2 = 0 \leq E_\mathrm{th}$ according to (52), so the proposed algorithm properly stops at iteration $I^\star$. This fact along with (52) leads to the result in (49).

References

[1] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.,* vol. 24, no. 2, pp. 98–105, Apr. 2017.

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

[3] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Commun.,* vol. 26, no. 5, pp. 77–83, Oct. 2019.

[4] Y.-S. Jeon, N. Lee, and H. V. Poor, "Robust data detection for MIMO systems with one-bit ADCs: A reinforcement learning approach," *IEEE Trans. Wireless Commun.,* vol. 19, no. 3, pp. 1663–1676, Mar. 2020.

[5] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cognitive Commun. Netw.,* vol. 3, no. 4, pp. 563–575, Dec. 2017.

[6] Y.-S. Jeon, S.-N. Hong, and N. Lee, "Supervised-learning-aided communication framework for MIMO systems with low-resolution ADCs," *IEEE Trans. Veh. Tech.,* vol. 67, no. 8, pp. 7299–7313, Aug. 2018.

[7] J. Konečný, B. H. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," arXiv:1511.03575v1 [cs.LG], Nov. 2015 [Online]. Available: http://arxiv.org/abs/1511.03575

[8] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency." arXiv:1610.05492v2 [cs.LG], Oct. 2017 [Online] Available: http://arxiv.org/abs/1610.05492

[9] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas "Communication-efficient learning of deep networks from decentralized data," arXiv:1602.05629v3 [cs.LG], Feb. 2017. [Online] Available: http://arxiv.org/abs/1602.05629

[10] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Represent. (ICLR),* Vancouver, BC, Canada, May 2018, pp. 1–13.

[11] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities and challenges," arXiv:1908.06847v3 [eess.SP], Sep. 2019. [Online]. Available: http://arxiv.org/abs/1908.06847

[12] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.,* vol. 68, no. 2, pp. 1146–1159, Feb. 2020.

[13] M. Chen, O. Semiari, W. Saad, X. Liu, and C. Yin, "Federated echo state learning for minimizing breaks in presence in wireless virtual reality networks," *IEEE Trans. Wireless Commun.,* vol. 19, no. 1, pp. 177–191, Jan. 2020.

[14] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun. (ICC),* Shanghai, China, May 2019, pp. 1–6.

[15] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Data-importance aware user scheduling for communication-efficient edge machine learning," arXiv:1910.02214v1 [cs.NI], Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.02214

[16] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Age-based policy for federated learning in mobile edge networks," in *Proc. IEEE International Conf. Acoustics, Speech and Signal Process. (ICASSP),* Barcelona, Spain, May 2020, pp. 8743–8747.

[17] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.,* vol. 68, no. 1, pp. 317–333, Jan. 2020.

[18] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Update aware device scheduling for federated learning at the wireless edge," in *Proc.IEEE International Symp. Inf. Theory (ISIT),* Los Angeles, CA, June 2020, pp. 2598–2603.

[19] 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," arXiv:1909.07972v1 [cs.NI], Sep. 2019. [Online]. Available: http://arxiv.org/abs/1909.07972

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

[21] 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.

[22] 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.

[23] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farhad, S. Jin, T. Q. S. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics and Security,* vol. 15, pp. 3454–3469, Apr. 2020.

[24] M. M. Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," in *Proc. IEEE Global Conf. Sig. Inf. Process. (GlobalSIP),* Ottawa, ON, Canada, Nov. 2019, pp. 1–6.

[25] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.,* vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

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

[27] 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, no. 11, pp. 5255–5268, Nov. 2019.

[28] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.,* vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[29] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO; Opportunities and challenges with very large arrays," *IEEE Sig. Process. Mag.,* vol. 30, no. 1, pp. 40–60, Jan. 2013.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980v9 [cs.LG], Jan. 2017. [Online]. Available: http://arxiv.org/abs/1412.6980

[31] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications,* Cambridge, U.K.: Cambridge University Press, 2012.

[32] M. A. Woodbury, "Inverting modified matrices," Statist. Res. Group, Princeton, NJ, USA, Memo Rep. 12, 1950.

[33] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory,* vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[34] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inf. Theory,* vol. 57, no. 7, pp. 4680–4688, Jul. 2011.

[35] M. Sundin, M. Jansson, and S. Chatterjee, "Conditional prior based LMMSE estimation of sparse signals," in *Proc. 21st European Signal Processing Conference (EUSIPCO),* Marrakech, Morocco, Sep. 2013, pp. 1–5.

[36] S. Sparrer and R. F. H. Fischer, "MMSE-based version of OMP for recovery of discrete-valued sparse signals," *IET Electronics Letts.,* vol. 52, no. 1, pp. 75–77, Jan. 2016.

[37] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication,* Cambridge, U.K.: Cambridge University Press, 2005.

[38] A. Papoulis and U. Pillai, *Probability, Random Variables and Stochastic Processes,* New York: McGraw-Hill, 2001.

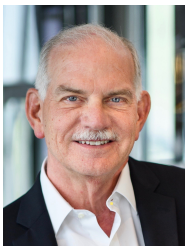
[39] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," [Online]. Available: http://yann.lecun.com/exdb/mnist/

[40] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," arXiv:2001.10402 [cs.IT], Jan. 2020. [Online]. Available: http://arxiv.org/abs/2001.10402

Yo-Seb Jeon (S'12–M'17) received the B.S. (Top Hons.) and Ph. D. degrees in the Department of Electrical Engineering from Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2012 and 2016, respectively. From September 2016 to August 2018, he was a Postdoctoral Research Associate at POSTECH. From September 2018 to January 2020, he was a Postdoctoral Research Fellow in the Department of Electrical Engineering at Princeton University. Since 2020, he has been on the faculty at POSTECH, where he is currently an Assistant Professor in the Department of Electrical Engineering. His research interests include the areas of wireless communications, machine learning, and signal processing. He was a recipient of the TJ PARK Graduate Fellowship from POSTECH (2012—2014).

Mohammad Mohammadi Amiri (S'16) received the B.Sc. and M.Sc. degrees in Electrical Engineering from Iran University of Science and Technology in 2011 and University of Tehran in 2014, respectively, both with highest rank in classes. He also obtained a Ph.D. degree at Imperial College London in 2019, and he is the recipient of the Best Ph.D. Thesis award from the IEEE Information Theory Chapter of UK and Ireland in 2019. He is currently a Postdoctoral Research Associate in the Department of Electrical Engineering at Princeton University. His research interests include information and coding theory, machine learning, wireless communications, and signal processing.

Jun Li (M'09-–SM'16) received Ph. D degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. From January 2009 to June 2009, he worked in the Department of Research and Innovation, Alcatel Lucent Shanghai Bell as a Research Scientist. From June 2009 to April 2012, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia. From April 2012 to June 2015, he was a Research Fellow at the School of Electrical Engineering, the University of Sydney, Australia. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He was a visiting professor at Princeton University from 2018 to 2019. His research interests include network information theory, game theory, distributed intelligence, multiple agent reinforcement learning, and their applications in ultra-dense wireless networks, mobile edge computing, network privacy and security, and industrial Internet of things. He has co-authored more than 200 papers in IEEE journals and conferences, and holds 1 US patents and more than 10 Chinese patents in these areas. He was serving as an editor of IEEE Communication Letters and TPC member for several flagship IEEE conferences. He received Exemplary Reviewer of IEEE Transactions on Communications in 2018, and best paper award from IEEE International Conference on 5G for Future Wireless Networks in 2017.



H. Vincent Poor (S'72–M'77–SM'82–F'87)received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor of Electrical Engineering. During 2006 to 2016, he served as Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the recent book *Multiple Access Techniques for 5G Wireless Networks and Beyond.* (Springer, 2019).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and a D.Eng. *honoris causa* from the University of Waterloo awarded in 2019.