Random Orthogonalization for Federated Learning in Massive MIMO Systems

Xizixiang Wei,Cong Shen, *Senior Member, IEEE,* Jing Yang, *Member, IEEE,* and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—We propose a novel communication design, termed *random orthogonalization*, for federated learning (FL) in a massive multiple-input and multiple-output (MIMO) wireless system. The key novelty of random orthogonalization comes from the tight coupling of FL and two unique characteristics of massive MIMO – channel hardening and favorable propagation. As a result, random orthogonalization can achieve natural over-the-air model aggregation without requiring transmitter side channel state information (CSI) for the uplink phase of FL, while significantly reducing the channel estimation overhead at the receiver. We extend this principle to the downlink communication phase and develop a simple but highly effective model broadcast method for FL. We also relax the massive MIMO assumption by proposing an enhanced random orthogonalization design for both uplink and downlink FL communications, that does not rely on channel hardening or favorable propagation. Theoretical analyses with respect to both communication and machine learning performance are carried out. In particular, an explicit relationship among the convergence rate, the number of clients, and the number of antennas is established. Experimental results validate the effectiveness and efficiency of random orthogonalization for FL in massive MIMO.**

***Index Terms*—Federated Learning; Convergence Analysis; Massive MIMO.**

I. Introduction

Machine learning (ML) model communication is widely considered as one of the primary bottlenecks for federated learning (FL) [2]–[4]. This is because an FL task consists of multiple learning rounds, each of which requires uplink and downlink model exchanges between clients and the server. The limited communication resources in both uplink and downlink, combined with the detrimental effects from channel fading, noise, and interference, severely impact the *scalability* (in terms of the number of participating clients) of FL in a wireless communication system.

One promising technique to tackle the scalability problem of FL over wireless communications is over-the-air computation (also known as AirComp); see [5] and the references therein. Instead of the standard approach of decoding the individual local models of each client and then aggregating, AirComp allows multiple clients to transmit uplink signals in a superpositioned fashion, and decodes the average global model directly at the FL server. In order to achieve this goal, a common approach is to "invert" the fading channel at each transmitter [6], [7], so that the sum model can be obtained at the server. AirComp has attracted considerable interest and a detailed literature review can be found in Section II.

However, much of the existing work on AirComp has several limitations. First, these methods often require channel state information at the transmitter (CSIT) for each individual client. The process of enabling individual CSIT is complicated – in a frequency division duplex (FDD) system, this involves the receiver estimating the channels and then sending back the estimates to the transmitters; in a time division duplex (TDD) system, one can benefit from channel reciprocity [8], [9], but there is still a need for an independent pilot for each client. In both cases, practical mechanisms to obtain individual CSIT do not scale with the number of clients. In addition, the precision of CSIT is often worse than that of channel state information at the receiver (CSIR). Second, most AirComp approaches in the literature require a channel inversion-type power control, which is well known to "blow up" when at least one of the channels is experiencing deep fading [8]. Third, AirComp approaches focus on improving the scalability and efficiency of the uplink communication phase in FL. How to address these challenges in the downlink communication phase remains underdeveloped.

Another important limitation is that the AirComp solution does not naturally extend to multiple-input and multiple-output (MIMO) systems where the uplink and downlink channels become vectors. Compared with the studies in scalar channels, there are only a few recent papers that explore the potential of MIMO for wireless FL. MIMO beamforming design to optimize FL has been studied in [10], [11]. Coding, quantization, and compressive sensing over MIMO channels for FL have been studied in [12], [13]. Nevertheless, none of these works tightly incorporates the unique properties of MIMO to the FL communication design. On the other hand, if we ignore the unique characteristics of FL, MIMO can also be utilized in a straightforward manner. In the uplink phase, we can use conventional MIMO estimators such as zero-forcing (ZF) or minimum mean square error (MMSE) to estimate each local model, and then compute the global model. In the downlink phase, we can design MIMO precoders to broadcast

A preliminary version of this work has been presented at the 2022 IEEE International Conference on Communications [1].

Xizixiang Wei and Cong Shen are with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, USA. (E-mail: {xw8cw,cong}@virginia.edu.)

Jing Yang is with the Department of Electrical Engineering, The Pennsylvania State University, USA. (E-mail: yangjing@psu.edu.)

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, USA. (E-mail: poor@princeton.edu.)

CS and XW was partially supported by the US National Science Foundation (NSF) under Grants CPS-2313110, ECCS-2143559, ECCS-2033671, and the Commonwealth Cyber Initiative (CCI) of Virginia under Award VV-1Q23-005; JY was partially supported by the US NSF under Grants CNS-1956276 and CNS-2114542; HVP was partially supported by the US NSF under Grants CNS-2128448.

the global model. However, these approaches incur a large channel estimation overhead, especially when the channels have high dimensions. Moreover, matrix inversions in the ZF or MMSE estimators and the optimization algorithms for the precoding design are computationally demanding, in particular for massive MIMO. This increases the complexity and latency of the overall system. In addition, decoding individual local models also makes it easier for the server to sketch the data distribution of the clients, leading to potential privacy leakage.

This paper aims at designing simple-yet-effective FL communication methods that can efficiently address the scalability challenge in FL for both uplink and downlink phases. The novelty comes from a tight integration of MIMO and FL – our design explicitly utilizes the characteristics of both components. The contributions of this paper are summarized as follows.

- We propose a novel *Random Orthogonalization* design for *massive* MIMO where the base station (BS) has a large number of antennas. In uplink communications, by leveraging the unique channel hardening and favorable propagation properties of massive MIMO, the proposed framework only requires the BS to estimate a *summation channel* and allows it to directly compute the global model via a simple linear projection, which significantly alleviates the burden on channel estimation[1] and achieves extremely low complexity and low latency. In downlink communications, the proposed method leads to a simple but highly effective model broadcast method for FL. Moreover, our approach is agnostic to the number of clients, and thus improves the scalability of FL.
- As the random orthogonalization designs rely on channel hardening and favorable propagation to eliminate the interference, which do not always hold in practice (e.g., when the number of antennas is small), we further propose an *enhanced random orthogonalization* design for both uplink and downlink FL communications, that leverages *channel echos* to compensate for the lack of channel hardening and favorable propagation. The enhanced random orthogonalization design thus can be applied to a general MIMO system.
- To analyze the performances of random orthogonalization, we derive the Cramer-Rao lower bounds (CRLBs) of the average model estimation errors as a theoretical benchmark. Moreover, taking both interference and noise into consideration, a novel convergence bound of FL is derived for the proposed methods over massive MIMO channels. Notably, we establish an explicit relationship among the convergence rate, the number of clients, and the number of antennas, which provides practical design guidance for wireless FL. Extensive numerical results validate the effectiveness and efficiency of the proposed random orthogonalization principle in a variety of FL and MIMO settings.

[1]For example, a single pilot can be used by all clients as long as it is sent synchronously, regardless of the number of clients that participate in the current FL round.

The remainder of this paper is organized as follows. Related works are surveyed in Section II. Section III introduces the FL pipeline and the wireless communication model. The proposed random orthogonalization principle is presented in Section IV, and then the enhanced design is proposed in Section V. Analyses of the CRLB as well as the FL model convergence are given in Section VI. Experimental results are reported in Section VII, followed by the conclusions in Section VIII.

II. Related Works

Improve FL communication efficiency. The original Federated Averaging (FedAvg) algorithm [2] reduces the communication overhead by only periodically averaging the local models. Theoretical understanding of the communication-computation tradeoff has been actively pursued and, depending on the underlying assumptions (e.g., independent and identically distributed (i.i.d.) or non-i.i.d. local datasets, convex or non-convex loss functions, gradient descent or stochastic gradient descent (SGD)), convergence analyses have been carried out [14], [15]. The approaches to reduce the payload size or communication frequency include sparsification [16], [17] and quantization [18]–[20]. There are also efforts to improve resource allocation [21]–[23].

AirComp for FL. As a special case of computing over multiple access channels [24], AirComp [6], [7], [10], [25] leverages the signal superposition properties in a wireless multiple access channel to efficiently compute the average ML model. This technique has attracted considerable interest, as it can reduce the uplink communication cost to be (nearly) agnostic to the number of participating clients. Client scheduling and various power and computation resource allocation methods have been investigated [26]–[31]. The assumption of full CSIT is relaxed in [32] by only using the phase information of each individual channel. Convergence guarantees of Aircomp under different constraints are reported in [33]–[37].

Communication design for FL in MIMO systems. There are some recent studies on optimizing the communication efficiency and learning performance in MIMO systems for FL, including transmit power control [38]–[40], data rate allocation [41], and compression [13], [42]. Several beamforming designs have been proposed to improve the performance of wireless FL [10], [43]–[46]. However, these methods require full CSIT and rely on complex optimization methods to design the beamformers, which becomes less attractive in massive MIMO due to the high communication and computation cost. Asymptotic analysis of the aggregation error in massive MIMO is provided in [43], [47], [48], which leads to beamformer designs that can relax the individual CSIT assumption in wireless FL. However, they only focus on the uplink communication phase.

III. System Model

A. FL Model

The FL problem studied in this paper mostly follows that in the original paper [2]. In particular, we consider an FL system with one central parameter server (e.g., base station) and a

set of at most N clients (e.g., mobile devices). Client $k \in [N] \triangleq \{1, 2, \cdots, N\}$ stores a local dataset $\mathcal{D}_k = \{\xi_i\}_{i=1}^{D_k}$, with its size denoted by D_k, that never leaves the client. Datasets across clients are assumed to be non-i.i.d. and disjoint. The maximum data size when all clients participate in FL is $D = \sum_{k=1}^{N} D_k$. Each data sample ξ is denoted by an input-output pair $\{\mathbf{x}, y\}$ for a supervised learning task. We use $f_k(\mathbf{w})$ to denote the local loss function at client k, which measures how well an ML model with parameter $\mathbf{w} \in \mathbb{R}^d$ fits its local dataset. The global objective function over all N clients is $f(\mathbf{w}) = \sum_{k\in[N]} p_k f_k(\mathbf{w})$, where $p_k = \frac{D_k}{D}$ is the weight of each local loss function, and the purpose of FL is to distributively find the optimal model parameter $\mathbf{w}^*$ that minimizes the global loss function: $\mathbf{w}^* \triangleq \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$. Let f^* and f_k^* be the minimum value of $f(\mathbf{w})$ and $f_k(\mathbf{w})$, respectively. Then, $\Gamma = f^* - \sum_{k=1}^{N} \frac{D_k}{D} f_k^*$ quantifies the degree of non-i.i.d. as defined in [15].

Specifically, the FL pipeline [2] iteratively executes the following steps at the t-th learning round.

1) **Downlink communication.** The BS broadcasts the current global model $\mathbf{w}_t$ to K randomly selected clients over the downlink wireless channel. We use $[K]$ to denote the selected client set to simplify the notation, but this should be interpreted as possibly different sets of clients at different round t.
2) **Local computation.** Each selected client uses its local dataset to train a local model improved upon the received global model $\mathbf{w}_t$. We assume that mini-batch SGD is used to minimize the local loss function. The parameter is updated iteratively (for E steps) at client k as: $\mathbf{w}_{t,0}^k = \mathbf{w}_t; \mathbf{w}_{t,\tau}^k = \mathbf{w}_{t,\tau-1}^k - \eta_t \nabla \tilde{f}_k(\mathbf{w}_{t,\tau-1}^k), \forall \tau = 1, \cdots, E; \mathbf{w}_{t+1}^k = \mathbf{w}_{t,E}^k$, where $\nabla \tilde{f}_k(\mathbf{w})$ denotes the mini-batch SGD operation at client k on model $\mathbf{w}$, and η_t is the learning rate (step size).
3) **Uplink communication.** Each selected client uploads its latest local model to the server synchronously over the uplink wireless channel.
4) **Server aggregation.** The BS aggregates the received noisy local models $\tilde{\mathbf{w}}_{t+1}^k$ to generate a new global model: $\mathbf{w}_{t+1} = \Sigma_{k\in[K]} \tilde{p}_k \tilde{\mathbf{w}}_{t+1}^k$, where $\tilde{p}_k \triangleq \frac{D_k}{\Sigma k\in[K] D_k}$. For simplicity, we assume that each local dataset has equal size, hence $\tilde{p}_k = 1/K$.

This work focuses on *both downlink and uplink* communication design in the FL pipeline. We next describe the communication models under consideration.

B. Communication Model

Consider a MIMO TDD communication system equipped with M antennas at the BS (server) where K randomly-selected single-antenna devices (clients) are involved in the t-th round of the aforementioned FL task. Let $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ denote the uplink wireless channel between the k-th client and the BS. During the uplink communication phase, each client transmits the difference between the received global model and the newly computed local model

$$\mathsf{x}_t^k = \mathbf{w}_t - \mathbf{w}_{t+1}^k, \quad \forall k \in [K] \tag{1}$$

to the BS, where $\mathsf{x}_t^k \triangleq [x_{1,t}^k, \cdots, x_{d,t}^k]^T \in \mathbb{R}^{d\times 1}$ denotes the d-dimensional model differential of client k at the t-th communication round. To simplify the notation, we omit index t by using $x_{k,i}$ instead of $x_{i,t}^k$ barring any confusion. Throughout this paper, we assume all active clients are synchronized. This can be achieved in practice by having the BS send a beacon signal to initialize uplink transmissions. For more details on synchronization, please refer to [49]. Therefore, each client can transmit every element of the differential model $\{x_{k,i}\}_{i=1}^d$ via d shared time slots[2]. For a given element $x_{k,i}$, the received signal at the BS is $\mathbf{y}_i^{\mathsf{UL}} = \sqrt{P_{\mathsf{Client}}} \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \mathbf{n}_i, \forall i = 1, \cdots, d$, where P_{Client} is the maximum transmit power of each client, and $\mathbf{n}_i \in \mathbb{C}^{M\times 1}$ represents the uplink noise. Denoting $\mathbf{H} \triangleq [\mathbf{h}_1, \cdots, \mathbf{h}_K] \in \mathbb{C}^{M\times K}$ as the channel vectors from all K clients and $\mathbf{x}_i \triangleq [x_{1,i}, \cdots, x_{K,i}]^T \in \mathbb{R}^{K\times 1}, \forall i = 1, \cdots, d$ as the i-th dimension model differential from all K clients at the t-th learning round, the received signal[3] can be written as

$$\mathbf{y}_i^{\mathsf{UL}} = \sqrt{P_{\mathsf{Client}}} \mathbf{H} \mathbf{x}_i + \mathbf{n}_i. \tag{2}$$

It is easy to see that (2) is a standard MIMO communication model and traditional MIMO estimators can be adopted to estimate $\hat{\mathbf{x}}_i = [\hat{x}_{1,i}, \cdots, \hat{x}_{K,i}]^T$. However, as discussed before, decoding $\{x_{k,i}\}_{i=1}^d$ individually and obtaining the aggregated parameter $\tilde{x}_i \triangleq \sum_{k\in[K]} \hat{x}_{k,i}$ by a summation is inefficient. After the BS decoding all aggregated parameter $\tilde{\mathbf{x}}_t \triangleq [\tilde{x}_1, \cdots, \tilde{x}_d]^T$ in d slots, it can compute the new global model as

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{K} \tilde{\mathbf{x}}_t. \tag{3}$$

In the downlink, after the computation of the global model $\mathbf{w}_{t+1} = [w_{1,t+1}, \cdots, w_{d,t+1}]^T$, the BS broadcasts the global model to all clients via a precoder $\mathbf{f} \in \mathbb{C}^{M\times 1}$, and the received signal at client k is given by

$$y_i^{\mathsf{DL}} = \sqrt{P_{\mathsf{BS}}} \mathbf{h}_{k,t+1}^H \mathbf{f} w_{i,t+1} + z_i^k, \quad \forall i = 1, \cdots, d, \tag{4}$$

where P_{BS} is the maximum transmit power of the BS and z_i^k denotes the downlink noise. We note that channel $\mathbf{h}_{k,t+1}^H \in \mathbb{C}^{1\times M}$ denotes the downlink vector channel that is reciprocal of the uplink channel in round $t+1$. Each client then computes an estimated global model and uses it as a new initial point for the next learning round after all d elements are received via (4). Traditionally, the precoder design of $\mathbf{f}$ belongs to broadcasting common messages (see [50] and the references therein). However, existing methods become impractical due to the difficulty in obtaining full CSI in massive MIMO systems, which motivates us to design $\mathbf{f}$ with only partial CSI. For mathematical simplicity, we assume a normalized symbol power[4], i.e., $\mathbb{E}\|x_{k,i}\|^2 = 1$ and $\mathbb{E}\|w_{i,t+1}\|^2 = 1$;

[2]In general, differential model parameters can be transmitted over any d shared orthogonal communication resources (e.g., time or frequency). For simplicity, we use d time slots here.

[3]For simplicity, we assume real signals $\{x_{k,i}\}_{i=1}^d$ are transmitted in this paper. It can be easily extended to complex signals by stacking two real model parameters into a complex signal, so that the full degree of freedom (d.o.f.) is utilized.

[4]The parameter normalization and de-normalization procedure in wireless FL follows the same as that in the Appendix of [6].

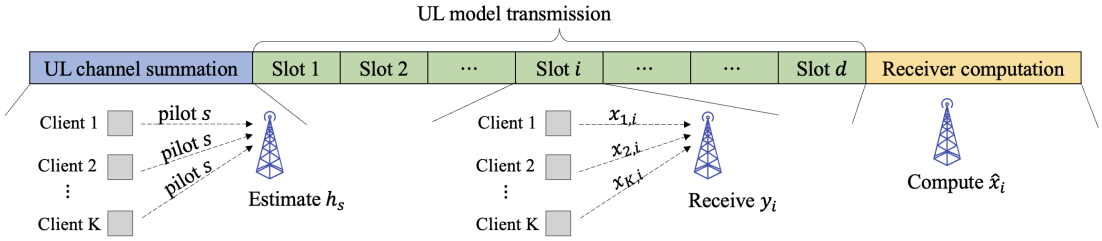


Fig. 1. An illustration of the proposed uplink FL design with massive MIMO.

normalized Rayleigh block fading channels[5] $\mathbf{h}_k \sim \mathcal{CN}(0, \frac{1}{M}\mathbf{I})$ in d slots; and i.i.d. Gaussian noise $\mathbf{n}_i \sim \mathcal{CN}(0, \frac{\sigma_{\mathsf{UL}}^2}{M}\mathbf{I})$ and $z_i^k \sim \mathcal{CN}(0, \sigma_{\mathsf{DL}}^2)$. We define the signal-to-noise ratio (SNR) as $\mathsf{SNR}_{\mathsf{UL}} \triangleq P_{\mathsf{Client}}/\sigma_{\mathsf{UL}}^2$ for uplink communications and $\mathsf{SNR}_{\mathsf{DL}} \triangleq P_{\mathsf{BS}}/\sigma_{\mathsf{DL}}^2$ for downlink communications, and without loss of generality (w.l.o.g.) we set $P_{\mathsf{Client}} = 1$ and $P_{\mathsf{BS}} = 1$.

IV. Random Orthogonalization

In this section, we present the key ideas of random orthogonalization. With this principle, the global model can be directly obtained at the BS via a simple operation in the uplink communications, and the global model can be broadcast to clients efficiently in the downlink communications. By exploring favorable propagation and channel hardening in massive MIMO, our proposed methods only require *partial* CSI, which significantly reduces the channel estimation overhead.

A. Uplink Communication Design

The designed framework contains the following three main steps in the uplink communications.

(U1) Uplink channel summation. The BS first schedules all clients participating in the current learning round to transmit a *common* pilot signal s synchronously. The received signal at the BS is

$$\mathbf{y}_s = \sum_{k\in[K]} \mathbf{h}_k s + \mathbf{n}_s, \tag{5}$$

and the BS can estimate the *summation* of channel vectors $\mathbf{h}_s \triangleq \sum_{k\in[K]} \mathbf{h}_k$ from the received signal $\mathbf{y}_s$. Given the pilot s, the estimates can be obtained via a maximum likelihood estimator $\arg\min_{\mathbf{h}_s} \|\mathbf{y}_s - \mathbf{h}_s s\|^2$ [52]. We can also adopt multiple pilots to improve the accuracy of channel estimation. We note that the complexity of this sum channel estimation does not scale with K. For the purpose of illustrating our key ideas, we assume perfect summation channel estimation at the BS for now. The channel estimation error of $\mathbf{h}_s$ will affect the effective SNR of the decoded model, and we will evaluate this impact in numerical experiments. We also note that when the pilot SNR is sufficiently high, one can directly scale the received signal $\mathbf{y}_s$ (by $1/s$) to obtain the estimated summation channel.

(U2) Uplink model transmission. All selected clients transmit model differential parameters $\{x_{k,i}\}_{i=1}^d$ to the BS in d shared time slots. The received signal for each differential model element is $\mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \mathbf{n}_i, \forall i = 1, \cdots, d$.

(U3) Receiver computation. The BS estimates each aggregated model element via the following simple *linear projection* operation:

$$\begin{aligned}\tilde{x}_i = \mathbf{h}_s^H \mathbf{y}_i &= \sum_{k\in[K]} \mathbf{h}_k^H \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i \\ &\overset{(a)}{=} \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i}}_{\text{Signal}} + \underbrace{\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i}}_{\text{Interference}} \\ &+ \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i}_{\text{Noise}} \overset{(b)}{\approx} \sum_{k\in[K]} x_{k,i}, \quad \forall i = 1, \cdots, d. \end{aligned} \tag{6}$$

The above three-step uplink communication procedure is illustrated in Fig. 1. Based on Eqn. (6), the BS then computes the global model via Eqn. (3) and begins the downlink global model broadcast.

As shown in (a) of Eqn. (6), inner product $\mathbf{h}_s^H \mathbf{y}_i$ can be viewed as the combination of three parts: signal, interference, and noise. We next show that, taking advantage of two fundamental properties of massive MIMO, the error-free approximation (b) in (6) is asymptotically accurate (as the number of BS antennas M goes to infinity).

Channel hardening. Since each element of $\mathbf{h}_k$ is i.i.d. complex Gaussian, by the law of large numbers, massive MIMO enjoys channel hardening [53]: $\mathbf{h}_k^H \mathbf{h}_k \to 1$, as $M \to \infty$. In practical systems, when M is large but finite, for the signal part of (6), we have

$$\mathbb{E}_{\mathbf{h}} \left[\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i} \right] = \sum_{k\in[K]} x_{k,i}, \tag{7}$$

and

$$\mathbb{V}\mathrm{ar}_{\mathbf{h}} \left[\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i} \right] = \frac{\sum_{k\in[K]} x_{k,i}^2}{M}. \tag{8}$$

Favorable propagation. Since channels between different users are independent random vectors, massive MIMO also offers favorable propagation [53]: $\mathbf{h}_k^H \mathbf{h}_j \to 0$, as $M \to \infty$, $\forall k \neq j$. Similarly, when M is finite, we have

$$\mathbb{E}_{\mathbf{h}} \left[\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} \right] = 0, \tag{9}$$

[5]The large-scale pathloss and shadowing effect is assumed to be taken care of by, e.g., open loop power control [51].

and

$$\mathbb{V}\text{ar}_{\mathbf{h}}\left[\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j x_{j,i}\right]=\frac{(K-1)\sum_{k\in[K]}x_{k,i}^2}{M}. \tag{10}$$

Furthermore, the expectation of the noise part in (6) is zero. Therefore, $\tilde{x}_i$ in (6) is an unbiased estimate of the average model. For a given K, the variances of both signal and interference decrease in the order of $\mathcal{O}(1/M)$, which shows that *massive MIMO offers* **random orthogonality** *for analog aggregation over wireless channels*. In particular, the asymptotic element-wise orthogonality of channel vector ensures channel hardening, and the asymptotic vector-wise orthogonality among different wireless channel vectors provides favorable propagation. Both properties render the linear projection operation $\mathbf{h}_s^H\mathbf{y}_i$ an ideal fit for the server model aggregation in FL.

To gain some insight of random orthogonality, we approximate the average signal-to-interference-plus-noise-ratio (SINR) after the operation in (6) as

$$\begin{aligned}\mathbb{E}[\text{SINR}_i]\approx&\\ &\frac{\mathbb{E}_{\mathbf{h},x}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k x_{k,i}\right\|^2}{\mathbb{E}_{\mathbf{h},\mathbf{n},x}\left\|\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j x_{j,i}+\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{n}_i\right\|^2}\\ &=\frac{M}{K-1+1/\mathsf{SNR}},\end{aligned} \tag{11}$$

which grows linearly with M for a fixed K. On the other hand, for a given number of antennas M, Eqn. (11) can be used to guide the choice of K in each communication round to satisfy an SINR requirement. We will provide more details on the scalability of clients via the convergence analysis of FL with random orthogonalization in Section VI-B. We note that Eqn. (11) is an approximate expression for SINR but it sheds light into the relationship between K and M. This approximation, however, is not used in the convergence analysis of FL with random orthogonalization in Section VI-B.

We note that the uplink random orthognalization design presented above is similar to that in [47], which also relies on orthogonality to directly compute the summation ML model at the server. Our work, however, builds a more complete framework that has both uplink and downlink designs, for both massive MIMO and general MIMO. This will be elaborated in the following sections.

B. Downlink Communication Design

Inspired by the uplink communication design, the downlink design contains the following two steps.

(D1) Uplink channel summation. This step remains the same as **U1** in the uplink design. We similarly assume perfect sum channel estimation $\mathbf{h}_s=\sum_{k\in[K]}\mathbf{h}_k$ at the BS.

(D2) Downlink global model broadcast. The BS broadcasts global model $\{w_i\}$ to all users, using the estimated summation channel $\mathbf{h}_s$ as the precoder. Hence the received signal at the k-th user is

$$\begin{aligned}y_k=\mathbf{h}_k^H\mathbf{h}_s w_i+z_i^k&\overset{(a)}{=}\underbrace{\mathbf{h}_k^H\mathbf{h}_k w_i}_{\text{Signal}}+\underbrace{\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j w_i}_{\text{Interference}}\\ &+\underbrace{z_i^k}_{\text{Noise}}\overset{(b)}{\approx}w_i\ \ \forall i=1,\cdots,d.\end{aligned} \tag{12}$$

The above two-step downlink communication procedure is illustrated in Fig. 2. Similar to the uplink case, the global model signal obtained at each client can also be regarded as the combination of three parts: signal, interference, and noise as shown in (12). Leveraging channel hardening and favorable propagation of massive MIMO channels as mentioned before, we have

$$\mathbb{E}_{\mathbf{h}}\left[\mathbf{h}_k^H\mathbf{h}_k w_i\right]=w_i\quad\text{and}\quad\mathbb{V}\text{ar}_{\mathbf{h}}\left[\mathbf{h}_k^H\mathbf{h}_k w_i\right]=\frac{w_i^2}{M}, \tag{13}$$

for the signal part of (12). Besides, we have

$$\mathbb{E}_{\mathbf{h}}\left[\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j w_i\right]=0 \tag{14}$$

and

$$\mathbb{V}\text{ar}_{\mathbf{h}}\left[\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j w_i\right]=\frac{(K-1)w_i^2}{M}, \tag{15}$$

for the interference part. The above derivation demonstrates that, similar to the uplink design, received signals obtained via (12) are unbiased estimates of global model parameters whose variances decrease in the order of $\mathcal{O}(1/M)$ with the increase of BS antennas. We next give a few remarks about the proposed uplink and downlink communication designs of FL with random orthogonalization.

Remark 1. *In uplink communications, unlike the analog aggregation method in [6], the proposed random orthogonalization does not require any individual CSIT. On the contrary, it only requires partial CSIR, i.e., the estimation of a summation channel $\mathbf{h}_s$, which is $1/K$ of the channel estimation overhead compared with the AirComp method in [10] or the traditional MIMO estimators. In downlink communications, the traditional precoder design for common message broadcast requires CSIT for each client. By using the summation channel $\mathbf{h}_s$ as the precoder for global model broadcast, only partial CSIT is needed. Since we assume a TDD system configuration, the downlink summation channel $\mathbf{h}_s$ can be estimated at a low cost utilizing channel reciprocity as shown in Step D1. Therefore, the proposed method is attractive in wireless FL due to its mild requirement of partial CSI. Moreover, the server obtains global models directly after a series of simple linear projections, which improves the privacy and reduces the system latency as a result of the extremely low computational complexity of random orthogonalization. The same applies to the downlink phase.*

Remark 2. *Note that although we assume i.i.d. Rayleigh fading channels across different clients, the proposed random orthogonalization method is still valid for other channel mod-*

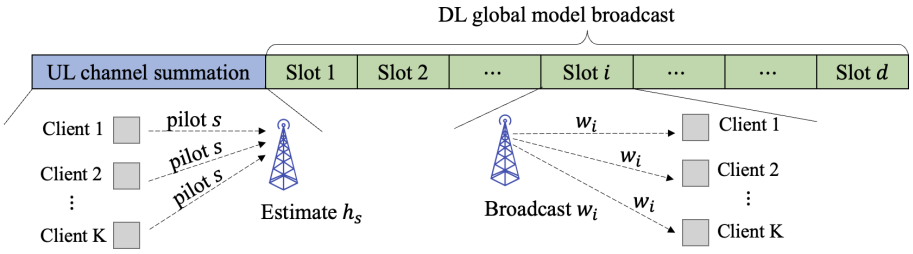


Fig. 2. An illustration of the proposed downlink FL design with massive MIMO.

els as long as channel hardening and favorable propagation are offered. In massive MIMO millimeter-wave (mmWave) communications, Rayleigh fading channels and light-of-sight (LOS) channels represent two extreme cases: rich scattering and no scattering. It is shown in [53] that both channel models offer asymptotic channel hardening and favorable propagation. In practice, we are likely to have a scenario which lies in between these two cases. Therefore, even under some channel correlations, the MIMO channels can still provide certain level of channel hardening and favorable propagation. Generally speaking, the random orthogonalization method is still valid in MIMO channels with low to moderate correlations, albeit with increased interference in the decoded models.

V. Enhanced Random Orthogonalization Design

The proposed random orthogonalization principle in Section IV requires channel hardening and favorable propagation. Although these two properties are quite common in massive MIMO systems, in the case that they are not available (e.g., the number of BS antennas is small), our design philosophy can still be applied by introducing a novel **channel echo mechanism**. In this section, we present an enhanced design to the methods in Section IV by taking advantage of channel echos.

Channel echo refers to that the receiver sends whatever it receives back to the original transmitter as the data payload. Similar techniques have been proposed before, such as "Echo-MIMO" and "two-way training" in [54] and [55]. However, they are developed for cooperative beamforming and optimal power allocation, respectively, and only focus on the single-user case. The main purpose of channel echo in our setting, however, is to "cancel" channel fading for each of the involved clients. The enhanced design for the uplink communications contains the following four main steps, which is demonstrated in Fig. 3.

(EU1) Uplink channel summation. The first step of the enhanced design follows the same as the random orthogonalization method (U1 and D1), so that the BS has the estimate sum channel vector $\mathbf{h}_s = \sum_{j\in[K]} \mathbf{h}_k$.

(EU2) Downlink channel echo. The BS sends the previously estimated $\mathbf{h}_s$ (after normalization to satisfy the power constraint) to all clients. For the k-th client, the received signal is $\mathbf{y}_k = \frac{\mathbf{h}_k^H \mathbf{h}_s}{\sqrt{K}} + \mathbf{n}_k$, by which client k can estimate $g_k = \mathbf{h}_k^H \mathbf{h}_s = \mathbf{h}_k^H \sum_{j\in[K]} \mathbf{h}_k = \|\mathbf{h}_k\|^2 + \sum_{j\in[K], j\neq k}^{K} \mathbf{h}_k^H \mathbf{h}_j$. Note again that we assume a perfect estimation of $\mathbf{h}_s$. An additional error term will appear in the estimation of g_k when the summation channel estimation is imperfect, which will be discussed later.

(EU3) Uplink model transmission. All involved clients transmit local parameter $\{x_{k,i}/\mathrm{Re}(g_k)\}_{k\in[K]}$ to the BS synchronously in d shared time slots: $\mathbf{y}_i = \sum_{k\in[K]}^{K} \mathbf{h}_k \frac{x_{k,i}}{\mathrm{Re}(g_k)} + \mathbf{n}_i, \ \forall i = 1, \cdots, d$.

(EU4) Server computation. The BS obtains $\sum_{k\in[K]} x_{k,i}$ via the following operation:

$$\begin{aligned}\tilde{x}_i &= \mathrm{Re}(\mathbf{y}_i^H \mathbf{h}_s) = \mathrm{Re}\left[\sum_{k\in[K]} \mathbf{h}_k^H \frac{x_{k,i}}{\mathrm{Re}(g_k)} \sum_{j\in[K]}^{K} \mathbf{h}_j + \mathbf{n}_i^H \sum_{j\in[K]} \mathbf{h}_j\right] \\ &= \sum_{k\in[K]} \frac{x_{k,i}}{\mathrm{Re}(g_k)} \mathrm{Re}\left[\mathbf{h}_k^H \sum_{j\in[K]} \mathbf{h}_j\right] + \mathrm{Re}\left[\mathbf{n}_i^H \sum_{j\in[K]} \mathbf{h}_j\right] \\ &= \sum_{k\in[K]} x_{k,i} + \mathrm{Re}\left[\sum_{j\in[K]} \mathbf{h}_j^H \mathbf{n}_i\right].\end{aligned} \tag{16}$$

Similarly, as shown in Fig. 4, the enhanced design for the downlink communication contains the following four main steps.

(ED1-2) Uplink channel summation and downlink channel echo. The first two steps in the downlink design remain the same as Steps EU1 and EU2 in the uplink design, so that the BS can estimate channel vector summation $\mathbf{h}_s = \sum_{j\in[K]} \mathbf{h}_k$ and each client can estimate the parameter g_k.

(ED3) Downlink global model broadcast. The BS broadcasts global model $\{w_i\}$ to all clients using the estimated sum channel $\frac{\mathbf{h}_s}{\sqrt{K}}$ as the precoder. The received signal at the k-th client is $y_k = \mathbf{h}_k^H \frac{\mathbf{h}_s}{\sqrt{K}} w_i + \mathbf{n}_i = \frac{1}{\sqrt{K}} g_k w_i + z_i^k, \forall i = 1, \cdots, d$.

(ED4) Model parameter computation. Each user obtains the global model $\{w_i\}$ via the following calculation:

$$\mathrm{Re}\left[\frac{\sqrt{K} y_k}{g_k}\right] = w_i + \mathrm{Re}\left(\frac{\sqrt{K} z_i^k}{g_k}\right) \ \forall i = 1, \cdots, d. \tag{17}$$

Note that the estimations of the aggregated signal and the global model in (16) and (17) are both *unbiased*, since $\mathbf{n}_j$, $\mathbf{h}_j$ and z_i^k are independent random variables with zero mean, and $\mathbb{E}[g_k] \neq 0$. Compared with the random orthogonalization method that offers *asymptotic* interference-free global model estimation, the received FL parameters obtained by the enhanced method are *completely interference-free* at both the

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3302335

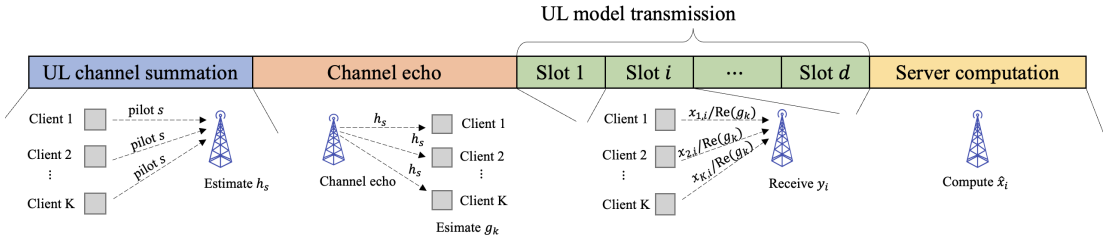


Fig. 3. An illustration of the proposed enhanced uplink FL design with massive MIMO.

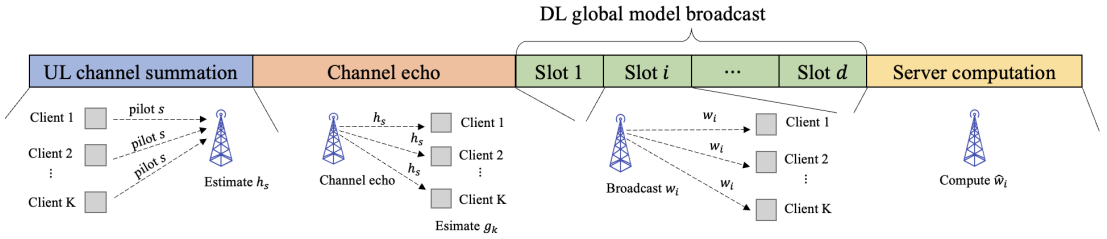


Fig. 4. An illustration of the proposed enhanced downlink FL design with massive MIMO.

server and the clients, as shown in (16) and (17). The extra channel echo steps (Step EU2 in uplink and Step ED1 in downlink) allow clients to obtain *partial* CSI g_k, so that they can pre-cancel and post-cancel channel interference among different user channels in uplink and downlink communications, respectively. Therefore, **this enhancement is valid even if channel hardening and favorable propagation are not present in wireless channels**, at a low cost of using one extra slot for the channel echo operation, and preserves all the other advantages of random orthogonalization.

Remark 3. *We note that so far, both random orthogonalization and enhanced methods assume a perfect estimation of* $\mathbf{h}_s$. *In practical systems, to improve the accuracy of the estimate* $\hat{\mathbf{h}}_s$, *BS can use multiple pilots or multiple time slots for improved channel estimation, but summation channel estimation error will inevitably exist. In the following, we use uplink random orthogonalization as an example to analytically evaluate the impact of imperfect summation channel estimation. Denote the imperfect summation channel as* $\hat{\mathbf{h}}_s = \mathbf{h}_s + \boldsymbol{\epsilon}$, *where* $\boldsymbol{\epsilon} \sim \mathcal{CN}(0, \frac{\sigma_\epsilon^2}{M}\mathbf{I}_M)$ *is the summation channel estimation error that is modeled as a Gaussian random vector with i.i.d. elements. The decoded signal in* (6) *becomes*

$$\tilde{x}_i = \hat{\mathbf{h}}_s^H \mathbf{y}_i = \left[\sum_{k\in[K]} \mathbf{h}_k^H + \boldsymbol{\epsilon}^H\right] \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i + \boldsymbol{\epsilon}^H \mathbf{n}_i = \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i}}_{\textit{Signal}} + \underbrace{\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} + \sum_{k\in[K]} \boldsymbol{\epsilon}^H \mathbf{h}_k x_{k,i}}_{\textit{Effective interference}}$$

$$+ \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i + \boldsymbol{\epsilon}^H \mathbf{n}_i}_{\textit{Effective noise}} \approx \sum_{k\in[K]} x_{k,i}, \quad \forall i = 1, \cdots, d. \tag{18}$$

Note that estimation in (18) *is still unbiased, since* $\mathbb{E}\left[\sum_{k\in[K]} \boldsymbol{\epsilon}^H \mathbf{h}_k x_{k,i}\right] = 0$ *and* $\mathbb{E}\left[\boldsymbol{\epsilon}^H \mathbf{n}_i\right] = 0$. *Moreover, we have*

$$\mathbb{V}ar_{\mathbf{h},\boldsymbol{\epsilon}}\left[\sum_{k\in[K]} \boldsymbol{\epsilon}^H \mathbf{h}_k x_{j,i}\right] = \frac{K\sigma_{\boldsymbol{\epsilon}}^2 \sum_{k\in[K]} x_{k,i}^2}{M}, \tag{19}$$

and

$$\mathbb{V}ar\left[\boldsymbol{\epsilon}^H \mathbf{n}_i\right] = \frac{\sigma_{\mathsf{UL}}^2 \sigma_{\boldsymbol{\epsilon}}^2}{M}. \tag{20}$$

Therefore, it is equivalent to consider the presence of channel estimation error as a perfect estimation case with a larger effective interference and noise, which depend on the channel estimation quality $\sigma_{\boldsymbol{\epsilon}}^2$. *In addition, for the enhanced method, the estimation error of* $\mathbf{h}_s$ *itself will not affect the performance, since the imperfect estimated summation channel will cancel out in Step EU/ED4. Only the imperfect estimation of* g_k *will influence the results. We provide more details on the robustness of the proposed schemes over imperfect* $\hat{\mathbf{h}}_s$ *and* g_k *in the experiment results.*

Remark 4. *In the enhanced uplink design, each client pre-cancels the channel fading effect so that the global model can be directly obtained at the BS after simple operations. Note that the analog aggregation method in [6] also uses "channel inversion" to pre-cancel channel fading. However, our design outperforms the method in [6] because the latter requires full CSIT, which leads to a large channel estimation overhead even with channel reciprocity in a TDD system. On the contrary, our method only requires partial CSI, which can be efficiently obtained via channel echos. Moreover, channel-inversion-based methods do not naturally extend to MIMO systems when the uplink channels become vectors, which*

makes the inversion operations at the transmitters nontrivial.

VI. Performance Analyses

We analyze the performance of the proposed methods from two aspects. On the communication performance side, we derive CRLBs of the estimates of model parameters in both uplink and downlink phases as the theoretical benchmarks. On the machine learning side, we present the convergence analysis of FL when the proposed communication designs are applied.

A. Cramer-Rao Lower Bounds

In the uplink communication, recall that the received signal is $\mathbf{y}_i = \mathbf{H}\mathbf{x}_i + \mathbf{n}_i$. Denoting $\boldsymbol{\mu}_{\mathsf{UL}} = \mathbf{H}\mathbf{x}_i$, we have $\mathbf{y}_i \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathsf{UL}}, \frac{1}{\mathsf{SNR}}\mathbf{I})$. To leverage CRLBs to evaluate the parameter estimation, we first need to derive the Fisher information of $\mathbf{x}_i$. Based on Example 3.9 in [52], we can write the Fisher information matrix (FIM) of the estimation of $\mathbf{x}_i$ as:

$$\mathbf{F}_{\mathsf{UL}} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}\left[\frac{\partial^H \boldsymbol{\mu}_{\mathsf{UL}}(\mathbf{x}_i)}{\partial \mathbf{x}_i}\frac{\partial \boldsymbol{\mu}_{\mathsf{UL}}(\mathbf{x}_i)}{\partial \mathbf{x}_i}\right]. \tag{21}$$

After inserting $\frac{\partial \boldsymbol{\mu}_{\mathsf{UL}}(\mathbf{x}_i)}{\partial \mathbf{x}_i} = \mathbf{H}$ into the FIM, we have $\mathbf{F}_{\mathsf{UL}} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}(\mathbf{H}^H\mathbf{H})$. Note that for the enhanced uplink design, we can absorb $\mathrm{Re}(g_k)$ into the effective channel as $\tilde{\mathbf{H}} \triangleq [\mathbf{h}_1/\mathrm{Re}(g_1), \cdots, \mathbf{h}_K/\mathrm{Re}(g_K)]$, and calculate FIM via $\mathbf{F}_{\mathsf{UL}} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}(\tilde{\mathbf{H}}^H\tilde{\mathbf{H}})$.

In the downlink communication, since $y_k = \mathbf{h}_k^H\mathbf{h}_s w_i + \mathbf{n}_k$, by the definition of $\mu_{\mathsf{DL}} = \mathbf{h}_k^H\mathbf{h}_s w_i$, we have that $y_k \sim \mathcal{CN}(\mu_{\mathsf{DL}}, \frac{1}{\mathsf{SNR}})$. The Fisher information of global model parameters is

$$\begin{aligned} F_{\mathsf{DL}} &= 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}\left[\frac{\partial^H \mu_{\mathsf{DL}}(w_i)}{\partial w_i}\frac{\partial \mu_{\mathsf{DL}}(w_i)}{\partial w_i}\right] \\ &= 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}(\mathbf{h}_k^H\mathbf{h}_s\mathbf{h}_s^H\mathbf{h}_k). \end{aligned} \tag{22}$$

The CRLBs of estimates are then given by the inverse of the Fisher information (matrix): $\mathbf{C}_{\hat{\mathbf{x}}_i} = \mathbf{F}_{\mathsf{UL}}^{-1}$ and $C_{\hat{w}_i} = 1/F_{\mathsf{DL}}$, respectively. CRLBs are the lower bounds on the variances of unbiased estimators, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information (matrix). We have shown that the proposed methods lead to unbiased estimations of the global model in both uplink and downlink communications. Hence, we can use the sum of all diagonal elements of $\mathbf{C}_{\hat{\mathbf{x}}}$ as the lower bound of the mean squared error (MSE) $\mathbb{E}\left\|\mathbf{x}_i - \hat{\mathbf{x}}_i\right\|^2$, and use $C_{\hat{w}_i}$ as the lower bound of MSE $\mathbb{E}\left\|w_i - \hat{w}_i\right\|^2$, to evaluate the performance of model estimation in both uplink and downlink communications. These bounds will be validated in the experiment results.

B. ML Model Convergence Analysis

We now analyze the ML model convergence performances of the proposed methods. Note that as we have proposed two different designs (basic and enhanced) for the uplink and downlink communications, respectively, there would be four cases of convergence analysis. Since these convergence analyses are quite similar, we only report one of these results.

We first make the following standard assumptions that are commonly adopted in the convergence analysis of FedAvg and its variants [15], [56], [57]. In particular, Assumption 1 indicates that the gradient of f_k is Lipschitz continuous. The strongly convex loss function in Assumption 2 is a category of loss functions that are widely studied in the literature (see [15] and its follow-up works). Assumptions 3 and 4 imply that the mini-batch stochastic gradient and its variance are bounded [14].

Assumption 1. *L-smooth:* $\forall\ \mathbf{v}$ *and* $\mathbf{w}$, $\|f_k(\mathbf{v}) - f_k(\mathbf{w})\| \leq L\|\mathbf{v} - \mathbf{w}\|$;

Assumption 2. *μ-strongly convex:* $\forall\ \mathbf{v}$ *and* $\mathbf{w}$, $\langle f_k(\mathbf{v}) - f_k(\mathbf{w}), \mathbf{v} - \mathbf{w}\rangle \geq \mu\|\mathbf{v} - \mathbf{w}\|^2$;

Assumption 3. ***Bounded variance for unbiased mini-batch SGD:*** $\forall k \in [N]$, $\mathbb{E}[\nabla \tilde{f}_k(\mathbf{w})] = \nabla f_k(\mathbf{w})$ *and* $\mathbb{E}\left\|\nabla f_k(\mathbf{w}) - \nabla \tilde{f}_k(\mathbf{w})\right\|^2 \leq H_k^2$;

Assumption 4. ***Uniformly bounded gradient:*** $\forall k \in [N]$, $\mathbb{E}\left\|\nabla \tilde{f}_k(\mathbf{w})\right\|^2 \leq H^2$ *for all mini-batch data.*

We next provide a convergence analysis of FL when the uplink communication utilizes random orthogonalization and the enhanced design is applied to the downlink communication. Note that unlike uplink communications, we cannot use model differential for downlink FL communications because of partial clients selection. To guarantee the convergence of FL, we need to borrow the necessary condition for noisy FL downlink communication from our previous work [58], i.e., downlink transmit power should scale in the order of $\mathcal{O}(t^2)$.

Theorem 1 (***Convergence for random orthogonalization in the uplink and enhanced method in the downlink***)**.** *Consider a wireless FL task that applies random orthogonalization for the uplink communications and the enhanced method for the downlink communications. With Assumptions 1-4, for some $\gamma \geq 0$, if we set the learning rate as $\eta_t = \frac{2}{\mu(t+\gamma)}$ and downlink SNR scales as $\mathsf{SNR}_{\mathsf{DL}} \geq \frac{1-\mu\eta_t}{\eta_t^2}$ in round t, we have*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)}\left[\frac{4B}{\mu^2} + (1+\gamma)\|\mathbf{w}_0 - \mathbf{w}^*\|^2\right], \tag{23}$$

for any $t \geq 1$, where

$$\begin{aligned} B \triangleq &\sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2 \\ &+ \frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2 + \frac{MK}{N^2(K+M)}. \end{aligned} \tag{24}$$

Proof. Proof of Theorem 1 is given in Appendix C. □

Theorem 1 shows that applying random orthogonalization in the uplink communications and enhanced method in the downlink communications preserves the $\mathcal{O}(1/T)$ convergence rate of vanilla SGD in FL tasks with perfect communications in both uplink and downlink phases. The factors that impact the convergence rate are captured entirely in the constant B, which come from multiple sources as explained below: $\frac{\sum_{k\in[N]}H_k^2}{N^2}$ comes from the variances of stochastic gradients; $6L\Gamma$ is introduced by the non-i.i.d. of local datasets; the

choice of local computation steps and the fraction of partial client participation lead to $8(E-1)^2H^2$ and $\frac{N-K}{N-1}\frac{4}{K}E^2H^2$, respectively; and the interference and noise in uplink and downlink communications result in $\frac{4}{K}\left(\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2$ and $\left(d+\frac{dK}{M}\right)$, respectively. Note that the impact of the downlink noise, i.e., $\mathsf{SNR}_{\mathsf{DL}}$, is not explicit in B due to the requirement of $\mathsf{SNR}_{\mathsf{DL}} \geq \frac{1-\mu\eta_t}{\eta_t^2}$ to guarantee the convergence.

Remark 5. *We note that Theorem 1 considers random orthogonalization in the uplink and enhanced method in the downlink. When random orthogonalization is adopted in the downlink, the convergence bound in* (23) *will suffer from an additive constant term. This is because the interference cannot be effectively reduced when downlink power scales in the order of $\mathcal{O}(t^2)$, as required for direct model transmission [58]. This gap is also empirically observed in the experiments (see Section VII-B). However, we also note that this gap is inversely proportional to the number of antennas M. Hence, as M becomes large, it reduces to zero asymptotically*[6].

We next analyze the relationship between the number of selected clients K and the number of BS antennas M to understand the scalability of multi-user MIMO for FL, which provides more insight for practical system design. To this end, we consider a simplified case where the system only configures random orthogonalization in the uplink communications, assuming that the downlink communications are error-free. Note that this configuration is reasonable when the BS has large transmit power. We further assume full client participation ($N = K$), one-step SGD at each device ($E = 1$), and i.i.d datasets across all clients ($\Gamma = 0$). For this special case, we establish Corollary 2 as follows.

Corollary 2 (***Convergence for the simplified case***)**.** *Consider a MIMO system that applies random orthogonalization for the uplink communications of FL with full client participation, one-step SGD at each device, and i.i.d datasets across all clients. Based on Assumptions 1-4 and choosing learning rate as $\eta_t = \frac{2}{\mu(t+\gamma)}$, $\forall t \in [T]$, the following inequality holds:*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)}\left[\frac{4\tilde{B}}{\mu^2} + (1+\gamma)\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2\right] \quad (25)$$

for any $t \geq 1$, where

$$\tilde{B} \triangleq \left[1 + \frac{K}{M} + \frac{1}{\mathsf{SNR}}\right]\frac{H^2}{K}. \quad (26)$$

Proof. Corollary 2 comes naturally from Theorem 1 by setting $N = K$, $\Gamma = 0$, $E = 1$, omitting the $\left(d+\frac{dK}{M}\right)$ term due to the perfect downlink communications, and the fact that $\mathbb{E}\left\|\nabla f_k(\mathbf{w}) - \nabla\tilde{f}_k(\mathbf{w})\right\|^2 \leq \mathbb{E}\left\|\nabla\tilde{f}_k(\mathbf{w})\right\|^2 \leq H^2$. □

Corollary 2 shows that there are two main factors that impact the convergence rate of FL with MIMO: **variance reduction** and **channel interference and noise**. In particular, the definition of $\tilde{B}$ in (26), which appears in Corollary 2, captures the joint impact of both factors. The nature of distributed SGD suggests that, for a fixed mini-batch size at each client, involving K devices enjoys a $\frac{1}{K}$ variance reduction of stochastic gradient at each SGD iteration [60], which is captured by the $\frac{H^2}{K}$ term in (26). However, due to the existence of interference and noise, the convergence rate is determined by both factors, shown as $\frac{H^2}{K}$ and $\frac{(K/M+1/\mathsf{SNR})H^2}{K} \approx \frac{H^2}{M}$ terms in (26). This suggests that the desired variance reduction may be adversely impacted if channel interference and noise dominate the convergence bound. In particular, when $M \gg K$, we have $\frac{1}{K} \gg \frac{1}{M}$, and the system enjoys almost the same variance reduction as the interference-free and noise-free case. However, in the case of $K \gg M$, we have $\frac{(K/M+1/\mathsf{SNR})}{K} \approx \frac{1}{M} \gg \frac{1}{K}$, and $\frac{H^2}{M}$ dominates the convergence bound. In this case, it is unwise to blindly increase the number of clients, as it does not have the advantage of variance reduction.

Remark 6. *In massive MIMO, a BS is usually equipped with many (up to hundreds) antennas. Although there may be large number of users participating in FL, only a small number of them are simultaneously active [10]. Both factors indicate that $K \ll M$ often holds in typical massive MIMO systems. The analysis reveals that our proposed framework enjoys nearly the same interference-free and noise-free convergence rate with low communication and computation overhead in massive MIMO systems.*

VII. Experiment Results

We evaluate the performances of random orthogonalization and the enhanced method for uplink and downlink FL communications through numerical experiments. From a communication performance perspective, we compare the proposed methods with the classic MIMO estimators and precoders with respect to the MSE. We provide the computation time comparison as a measure of the complexity of various methods. We also discuss the robustness of the proposed methods when the properties of channel hardening and favorable propagation are not fully offered and the channel estimation is imperfect. We further verify the effectiveness of the proposed methods via FL tasks using real-world datasets.

A. Communication Performance

We consider a massive MIMO BS with $M = 64$, 128, 256, 512, or 1024 antennas, with $K = 8$ active users participating in an FL task. We assume a Rayleigh fading channel model, i.e., $\mathbf{h}_k \sim \mathcal{CN}(0, \frac{1}{M}\mathbf{I})$, for each user, and use the MSE of the computed global model parameters in uplink and downlink communications to evaluate the system performance. All MSE results are obtained from 2000 Monte Carlo experiments. We use CRLBs derived in Section VI-A as the benchmark of the computed MSEs. More specifically, the benchmark corresponds to the mean CRLBs calculated via (21) and (22) using the channel realizations in the Monte Carlo simulation. In addition, we adopt the traditional MIMO MMSE estimator and the semidefinite relaxation based (SDR-based) precoder design method in [50] for performance comparisons of uplink and downlink communications, respectively.

[6]Due to the space limitation, the technical details for this remark are deferred to our technical report [59].

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3302335

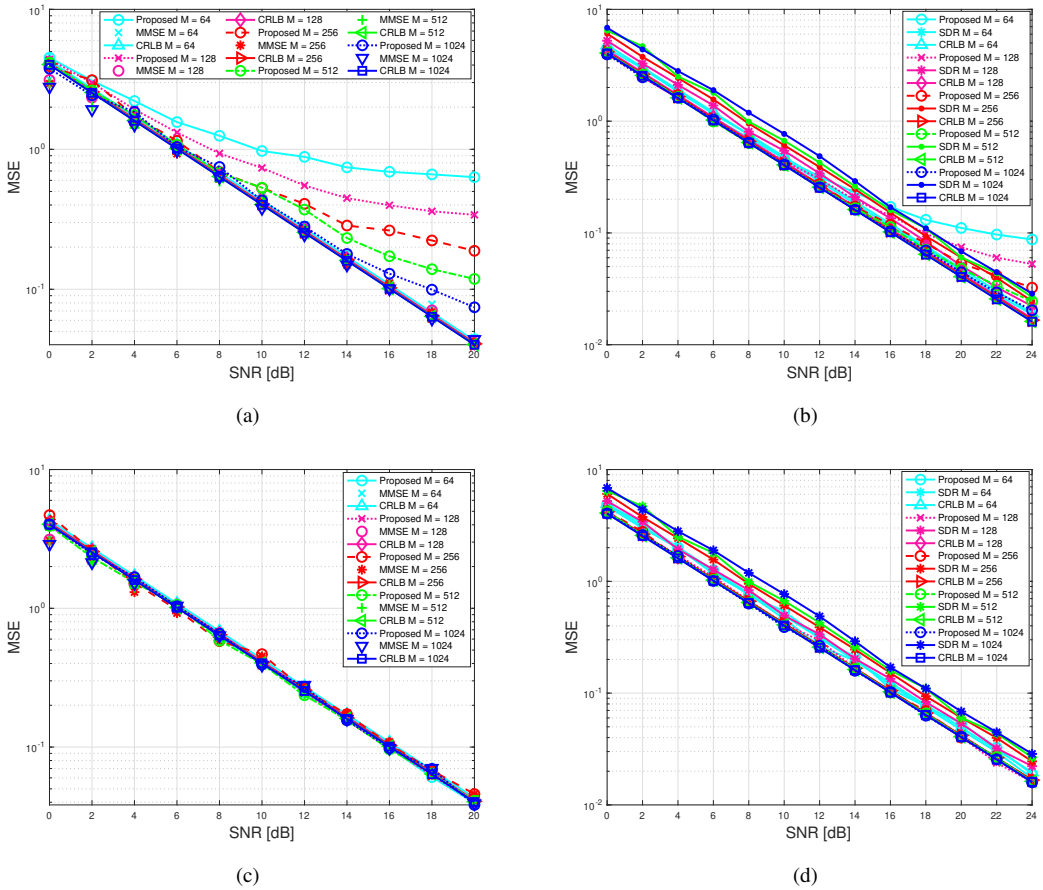


Fig. 5. MSE of the received global ML model parameters versus SNR of random orthogonalization in uplink (a) and downlink (b) communications and of the enhanced method in uplink (c) and downlink (d) communications.

Effectiveness. Fig. 5(a) and Fig. 5(b) compare the MSE performance of the random orthogonalization method in uplink and downlink communications with the traditional MIMO estimator/precoder as well as the CRLB under different system SNRs. As illustrated in these two plots, the proposed method performs nearly identically to the CRLB in low and moderate SNRs under different antenna configurations (see $\mathsf{SNR} \leq 12$ dB for uplink and $\mathsf{SNR} \leq 18$ dB for downlink, when $M = 128$, 256 and 1024). As the SNR increases, the dominant factor affecting system performances becomes the interference among different users. In the uplink communications, when $K \leq M$ and at high SNR, Eqn. (11) shows that for a given K and M, the proposed method has a fixed (approximate) $\mathsf{SIR} = \frac{K-1}{M}$ as $\mathsf{SNR} \to \infty$, which explains why the performance of the proposed scheme deteriorates compared with MMSE at high SNR. This phenomenon is more prominent when the number of antennas at the BS is relatively small ($M = 64$ and 128). However, this issue disappears naturally as the number of BS antennas increases. It can be seen in Fig. 5(a) that the performance gap between the proposed method and the CRLB reduces, from about 12 dB when $M = 64$ to about 2 dB when $M = 1024$ at $\mathsf{SNR} = 20$ dB, in uplink communications. We note that, although random orthogonalization produces higher MSEs than the MMSE estimator, the FL tasks have the same convergence rate under a constant SINR in uplink communications as indicated by the convergence analysis. This will be further validated in Section VII-B by showing that random orthogonalization hardly slows the convergence of FL. Similar to uplink, random orthogonalization performs nearly identically to the CRLB in low and moderate SNRs in downlink, and only loses about $0.5 \sim 6$ dB under different antenna configurations at $\mathsf{SNR} = 24$ dB. Moreover, random orthogonalization outperforms the SDR-based method at almost all SNRs and antenna configurations. Due to its sub-optimality, the performance of SDR-based method deteriorates as the number of antennas increases. In particular, it has about 3 dB loss compared with the CRLB when $M = 1024$, which further highlights the strengths of our approach for large arrays in massive MIMO. We should emphasize that our method only requires $1/K$ of the channel estimation overheard (partial CSI) compared with both MMSE and the SDR-based method (full CSI), and this advantage is more pronounced when the BS is equipped with a larger number of antennas.

Similarly, Fig. 5(c) and Fig. 5(d) compare the MSE performance of the enhanced method in uplink and downlink communications with the MMSE estimator / SDR-based precoder. It is clear from both plots that the enhanced method achieves MSEs that are very close to the CRLBs. Furthermore, it performs nearly identically as the MMSE estimator in uplink, and outperforms the SDR-based method by about

TABLE I
COMPUTATION TIME COMPARISON BETWEEN THE PROPOSED METHODS AND THE MMSE/SDR METHOD

# antennas (M)	Total CPU time (second) RO-UL	MMSE	Ratio RO-UL/MMSE	Total CPU time (second) Enhanced-UL	MMSE	Ratio Enhanced-UL/MMSE
256	0.0186	2.7141	0.68%	0.0203	2.9228	0.69%
512	0.0303	12.4155	0.24%	0.0469	16.3938	0.30%
1024	0.0448	82.3530	0.05%	0.0711	91.4117	0.07%
# antennas (M)	Total CPU time (second) RO-DL	SDR	Ratio RO-DL/SDR	Total CPU time (second) Enhanced-DL	SDR	Ratio Enhanced-DL/SDR
256	0.0157	25.1492	0.062%	0.0163	28.8593	0.056%
512	0.0415	324.7349	0.012%	0.0592	492.9539	0.012%
1024	0.0571	4819.6221	0.0012%	0.0695	5925.9250	0.0011%

$0.5-3$ dB in downlink for different antenna configurations. Therefore, by introducing channel echos, the enhanced method achieves excellent performance while consuming relatively low additional resource. Unlike random orthogonalization, the enhanced method cancels out all the interference in the decoded signal. Therefore, it is more attractive for smaller antenna arrays when the MIMO channels are not sufficiently orthogonal, despite at an additional channel echo cost. Random orthogonalization and the enhancement hence supplement each other, and they jointly serve as an efficient framework for both uplink and downlink communications of FL under different array configurations. The machine learning model parameters we estimate in FL are *real* signals. This constraint leads to a biased estimator at very low SNR, which leads to MSE saturation. Therefore, the MMSE estimator achieves a lower MSE than the CRLB when SNR ≤ 4 dB. It is worth noting that we can obtain an unbiased estimator by using *complex* signals and keeping the imaginary part of the estimates. For more details on MSE saturation in low SNRs, please refer to Section VI-D in [61].

Efficiency. We next focus on the low-latency advantage of the proposed methods, which originate from the low computational complexity. The complexity of both MMSE and SDR-based methods scale as $\tilde{O}(M^3)$, which is considerably higher compared with the $\tilde{O}(M)$ complexity of random orthogonalization and the enhanced method. To illustrate the benefit of low latency, we report the CPU time as a reference for an intuitive demonstration. Table I compares the computational time of the proposed schemes with the MMSE estimator and the SDR-based precoder when $\mathsf{SNR} = 10$ dB in the uplink and downlink communications, respectively. The total CPU time is the *cumulative time* of each algorithm over 2000 Monte Carlo experiments. We see that the time consumption of random orthogonalization and the enhanced method is much less than that of the MMSE estimator and the SDR-based precoder. Especially, when $M = 1024$, despite the 0.3 dB normalized MSE (NMSE) performance loss of random orthogonalization compared with the MMSE estimator in the uplink communications (as shown in Fig. 5(a)), the computation time of the former is only 0.05% of the latter. The proposed methods are even more computationally efficient for the downlink communications, as the total CPU time is less than 0.1% of the SDR-based method in all settings. All these results suggest that both random orthogonalization and its enhancement are attractive in massive MIMO systems, because they have promising MSE performances but require much less channel estimation overhead and achieve extremely lower system latency than the classic MIMO estimators and precoders.

Robustness. We now focus on the robustness of the proposed methods, and evaluate the MSEs of the global model parameters obtained at $\mathsf{SNR} = 10$ dB through 2000 Monte Carlo experiments. Fig. 6 reports the achieved MSEs of the random orthogonalization method when the (approximate) channel hardening and favorable propagation are not strictly offered, i.e., the wireless channels are correlated. We consider two channel correlation models with covariance matrix elements equal to 1 on the diagonal and equal to 0.01 or 0.05 off the diagonal, respectively. It is observed that when the off-diagonal elements are 0.01, random orthogonalization performs nearly identically as that in the ideal i.i.d. Rayleigh fading channel case. Even when the off-diagonal elements equal to 0.05, the achieved MSEs only increase by less than 1 dB in the worst case (when $M = 256$). The MSEs become closer to those of the i.i.d. Rayleigh channel cases when M increases, as larger antenna arrays offer higher orthogonality.

We next evaluate the performance when the estimation of the summation channel $\mathbf{h}_s$ (and g_k in the enhanced method) is imperfect. The right sub-figure of Fig. 6 compares the MSEs of both proposed methods when the channel estimation is obtained under $\mathsf{SNR} = 20$ dB. It reveals that the downlink communication is more robust than the uplink – the former achieves nearly identical MSEs as the ideal case even when the channel estimation is inaccurate. For the uplink, an imperfect channel estimation increases the MSEs by $1 \sim 3$ dB depending on the antenna configurations. However, we emphasize again that the FL tasks have the same convergence rate under a constant SINR in the uplink communications (thanks to the model differential transmission).

B. Learning Performance

To evaluate the learning performance, we carry out experiments of FL classification tasks using two widely adopted real-world datasets: MNIST and CIFAR-10, via a support vector machine (SVM) model and a convolution neural network (CNN) model, respectively. In the MNIST-SVM experiment, we evaluate the proposed uplink and downlink design separately, to identify their individual influence on the learning

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3302335

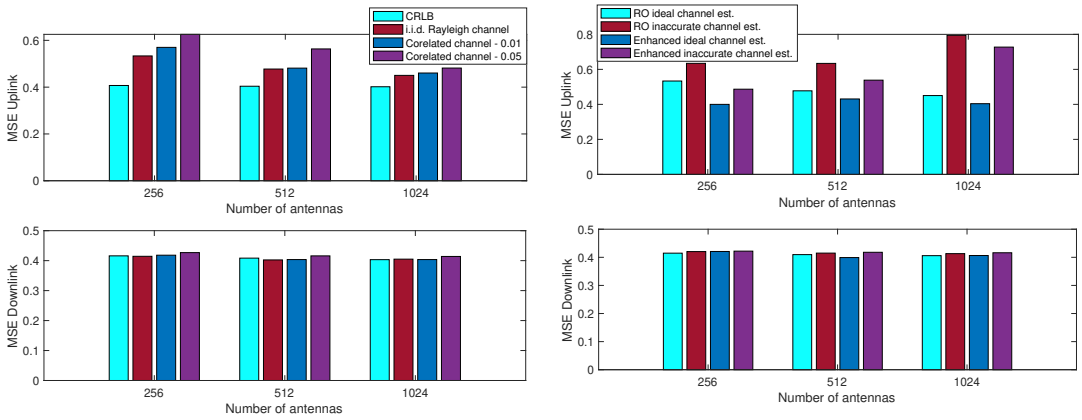
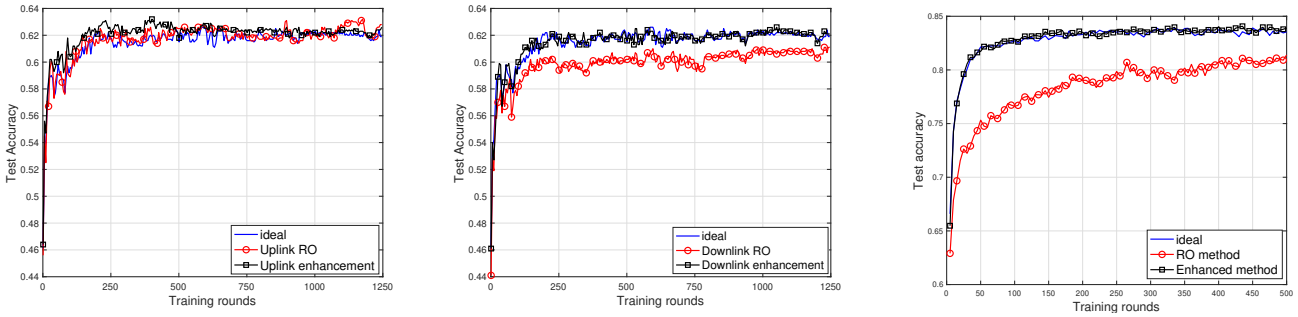


Fig. 6. MSE comparison of the received global ML model parameters when channel hardening and favorable propagation are not fully offered (left) and channel estimation is imperfect (right).



(a) MNIST uplink (b) MNIST downlink (c) CIFAR-10 uplink+downlink

Fig. 7. Comparison of test accuracy. (a): a SVM FL task with an ideal uplink communication (interference and noise free), random orthogonalization, and the enhanced method; (b): a SVM FL task with an ideal downlink communication (interference and noise free), random orthogonalization, and the enhanced method; (c): a CIFAR classification FL task with ideal uplink and downlink communications (interference and noise free), random orthogonalization, and the enhanced method.

performance. Then, in the CIFAR-CNN experiment, we jointly consider the uplink and downlink designs to evaluate the overall system performance of the proposed framework.

MNIST-SVM. The MNIST dataset collates small square 28×28 pixel gray-scale images of handwritten single digits between 0 and 9 [62]. We implement a SVM to classify even and odd numbers in the MNIST dataset, with $d = 784$. Total clients are set as $N = 20$, the size of each local dataset is 500, the size of the test set is 2000, and $E = 1$. The local dataset can be regarded as non-i.i.d. since we only allocate data of one label to each client. We consider a massive MIMO cell with $M = 256$ antennas at the BS and $K = 8$ (out of 20) randomly selected clients are involved in each learning round. The channels between each client and the BS are assumed to be i.i.d. Rayleigh fading.

Fig. 7(a) reports the test accuracy when the uplink adopts the proposed method, and the downlink is assumed to be noise-free. The uplink SNR is set as 10 dB. We can see that both random orthogonalization and the enhanced method behave almost identically as the ideal case where both uplink and downlink communications are perfect. Note that although the global model received at the BS has noise and interference components, the actual learning performances of the two methods do not deteriorate. Due to the model differential transmission in the uplink communications, the effective SINR of the received global model gradually increases as the model converges, despite the presence of channel interference and noise. Fig. 7(b) demonstrates the learning performance when the proposed designs are applied to the downlink communications. Since the model differential transmission is infeasible, we set the initial downlink SNR as 0 dB and scale at a rate of $\mathcal{O}(t^2)$ as the learning progresses (see [58]). We notice that the learning performance of the enhanced method is almost identical to that of the ideal case, while there is about 2% test accuracy loss for random orthogonalization. Note that, for downlink communications, because the BS only applies the normalized summation channel as the precoder, large-scale fading will result in different received SNRs for different clients. In this case, the above downlink SNR can be considered as the "worst" SNR among the involved clients (as the BS power control will need to target the worst-case user). Therefore, the reported result can be viewed as a lower bound of the actual performance.

CIFAR-CNN. The CIFAR-10 dataset consists of 32×32 color images in 10 classes, and we train a CNN model for the classification task. The CNN model consists of two 5×5 convolution layers (both with 64 channels), two fully connected layers (384 and 192 units respectively) with the

ReLU activation, and a final output layer with the softmax activation. The two convolution layers are both followed by 2×2 max pooling and a local response norm layer. In the FL tasks, we set $N = K = 10$, and the size of each local dataset is 1000, with mini-batch size 50 and $E = 5$. The initial learning rate is $\eta = 0.15$ and decays every 10 rounds with rate 0.99. We consider a massive MIMO cell with $M = 1024$ antennas at the BS and the channels between each client and the BS are assumed to be i.i.d. Rayleigh fading. The uplink SNR is set as 10 dB and the initial downlink SNR is set as 0 dB, and scales at the rate of $\mathcal{O}(t^2)$.

Fig. 7(c) illustrates the training loss and test accuracy versus the learning rounds when *both* the uplink and downlink communications adopt the random orthogonalization method or the enhanced method, respectively. It is observed that the enhanced method achieves similar training loss and test accuracy as the ideal case. Due to the constant interference in the downlink communications, random orthogonalization incurs a test accuracy loss of about 3%.

To summarize, experiments on both datasets demonstrate that random orthogonalization suffers a slight performance degradation over the ideal case when it is applied to the downlink communications. As we have stated in Remark 5, unlike the enhanced method that cancels all interference in the received global model, the interference is constant in the global model obtained via random orthogonalization despite the increased SNR. Note that this gap can be reduced by increasing M. Therefore, downlink random orthogonalization is more attractive in systems with large number of antennas or severely limited resources.

VIII. Conclusions

Leveraging the unique characteristics of channel hardening and favorable propagation in a massive MIMO system, we have proposed a novel uplink communication method, termed *random orthogonalization*, that significantly reduces the channel estimation overhead while achieving natural over-the-air model aggregation without requiring transmitter side channel state information. We have extended this principle to the downlink communication phase and developed a simple but highly effective model broadcast method for FL. We also relaxed the massive MIMO assumption by proposing an enhanced random orthogonalization design that utilizes channel echos. Theoretical performance analyses, from both communication (CRLB) and machine learning (model convergence rate) perspectives, have been carried out. The theoretical results suggested that random orthogonalization achieves the same convergence rate as vanilla FL with perfect communications asymptotically, and were further validated with numerical experiments. We will extend our work in the scenario where each client is equipped with multiple antennas in the future research.

Appendix A
Preliminaries

We first change the timeline to be with respect to the overall SGD iteration time steps instead of the communication rounds, i.e.,

$$t = \underbrace{1, \cdots, E}_{\text{round } 1}, \underbrace{E+1, \cdots, 2E}_{\text{round } 2}, \cdots, \underbrace{(T-1)E+1, \cdots, TE}_{\text{round } T}.$$

Note that the global model $\mathbf{w}_t$ is only accessible at the clients for specific $t \in \mathcal{I}_E$, where $\mathcal{I}_E = \{nE \mid n = 1, 2, \dots\}$, i.e., the time steps for communication. The notation for η_t is similarly adjusted to this extended timeline, but their values remain constant within the same round. The key technique in the proof is the *perturbed iterate framework* in [63]. In particular, we first define the following local training variables for client k: $\mathbf{v}_{t+1}^k \triangleq \mathbf{p}_t^k - \eta_t \nabla \tilde{f}_k(\mathbf{p}_t^k)$; when $t+1 \notin \mathcal{I}_E$, we have $\mathbf{v}_{t+1}^k = \mathbf{u}_{t+1}^k = \mathbf{w}_{t+1}^k = \mathbf{p}_{t+1}^k$; when $t+1 \in \mathcal{I}_E$, we have: $\mathbf{u}_{t+1}^k = \frac{1}{K}\sum_{i\in[K]} \mathbf{v}_{t+1}^i$, $\mathbf{w}_{t+1}^k = \frac{1}{K}\sum_{i\in[K]} \mathbf{h}_s^H \mathbf{h}_k (\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E}) + \frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E}$ and

$$\mathbf{p}_{t+1}^k \triangleq \begin{cases} \mathbf{w}_{t+1}^k + \tilde{\mathbf{z}}_{t+1}^k & \text{if } k \in [K], \\ \mathbf{w}_{t+1}^k & \text{if } k \notin [K]; \end{cases}$$

where $\mathbf{N}_{t+1} \triangleq [\mathbf{n}_1, \cdots, \mathbf{n}_i, \cdots, \mathbf{n}_d]^H \in \mathbb{C}^{d\times M}$ is the stack of uplink noise in (5), and

$$\tilde{\mathbf{z}}_{t+1}^k \triangleq \begin{cases} \sqrt{K}\left[\mathrm{Re}(z_1^k/g_1), \cdots, \mathrm{Re}(z_d^k/g_d)\right]^H \in \mathbb{C}^{d\times 1} & \text{if } k \in [K], \\ 0 & \text{otherwise,} \end{cases}$$

are the downlink noise in (12), respectively. Then, we construct the following *virtual sequences*: $\overline{\mathbf{v}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{u}_t^k$, $\overline{\mathbf{u}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{v}_t^k$, $\overline{\mathbf{w}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{w}_t^k$, and $\overline{\mathbf{p}}_t = \frac{1}{N}\sum_{k=1}^N \mathbf{p}_t^k$. We also define $\overline{\mathbf{g}}_t = \frac{1}{N}\sum_{k=1}^N \nabla f_k(\mathbf{w}_t^k)$ and $\mathbf{g}_t = \frac{1}{N}\sum_{k=1}^N \nabla \tilde{f}_k(\mathbf{w}_t^k)$ for convenience. Therefore, $\overline{\mathbf{v}}_{t+1} = \overline{\mathbf{w}}_t - \eta_t \mathbf{g}_t$ and $\mathbb{E}[\mathbf{g}_t] = \overline{\mathbf{g}}_t$. Note that the global model $\mathbf{w}_{t+1}$ is only meaningful when $t+1 \in \mathcal{I}_E$, hence we have $\mathbf{w}_{t+1} \triangleq \frac{1}{K}\sum_{k\in[K]} \mathbf{w}_{t+1}^k = \frac{1}{N}\sum_{k=1}^N \mathbf{w}_{t+1}^k = \overline{\mathbf{w}}_{t+1}$. Thus it is sufficient to analyze the convergence of $\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2$ to evaluate random orthogonalization.

Appendix B
Lemmas

We first establish the following lemmas that are useful in the proof of Theorem 1.

Lemma 1. *Let Assumptions 1-4 hold, η_t is non-increasing, and $\eta_t \le 2\eta_{t+E}$ for all $t \ge 0$. If $\eta_t \le 1/(4L)$, we have $\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \le (1 - \eta_t\mu)\mathbb{E}\|\overline{\mathbf{p}}_t - \mathbf{w}^*\|^2 + \eta_t^2\left(\sum_{k\in[N]} H_k^2/N^2 + 6L\Gamma + 8(E-1)^2H^2\right)$.*

Lemma 2. *Let Assumptions 1-4 hold. With $\eta_t \le 2\eta_{t+E}$ for all $t \ge 0$ and $\forall t+1 \in \mathcal{I}_E$, we have $\mathbb{E}[\overline{\mathbf{u}}_{t+1}] = \overline{\mathbf{v}}_{t+1}$, and $\mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t+1}\|^2 \le \frac{N-K}{N-1}\frac{4}{K}\eta_t^2E^2H^2$.*

Lemmas 1 and 2 establish bounds for the one-step SGD and random client sampling, respectively. These results only concern the local model update and user selection, and are not impacted by the noisy communication. The proofs are similar to the technique in [14], and are omitted due to space limitation.

Lemma 3. *Let Assumptions 1-4 hold. With $\eta_t \le 2\eta_{t+E}$ for all $t \ge 0$ and $\forall t+1 \in \mathcal{I}_E$, we have $\mathbb{E}[\overline{\mathbf{w}}_{t+1}] = \overline{\mathbf{u}}_{t+1}$, and $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\|^2 \le \frac{4}{K}\left[\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2E^2H^2$.*

Proof. We take expectation over randomness of fading channel and channel noise. As mentioned in Section IV, leveraging channel hardening and favorable propagation properties, we have

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{w}}_{t+1}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_{t+1}^{k}\right] = \mathbb{E}[\mathbf{w}_{t+1}^{k}] \\
&= \mathbb{E}\left[\frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_i(\mathbf{v}_{t+1}^{i}-\mathbf{w}_{t+1-E}) + \frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E}\right] \\
&= \mathbb{E}\left[\frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_i(\mathbf{v}_{t+1}^{i}-\mathbf{w}_{t+1-E})\right] + \mathbb{E}\left[\frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s\right] \\
&+ \mathbb{E}\left[\mathbf{w}_{t+1-E}\right] \\
&= \frac{1}{K}\sum_{i\in[K]}\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_i(\mathbf{v}_{t+1}^{i}-\mathbf{w}_{t+1-E})\right] + \mathbf{w}_{t+1-E} \\
&= \frac{1}{K}\sum_{i\in[K]}\mathbb{E}\left[\mathbf{h}_i^H\mathbf{h}_i(\mathbf{v}_{t+1}^{i}-\mathbf{w}_{t+1-E})\right] \\
&+ \frac{1}{K}\sum_{i\in[K]}\mathbb{E}\left[\sum_{k\in[K],k\neq i}\mathbf{h}_k^H\mathbf{h}_i(\mathbf{v}_{t+1}^{i}-\mathbf{w}_{t+1-E})\right] + \mathbf{w}_{t+1-E} \\
&= \frac{1}{K}\sum_{i\in[K]}(\mathbf{v}_{t+1}^{i}-\mathbf{w}_{t+1-E}) + \mathbf{w}_{t+1-E} = \frac{1}{K}\sum_{i\in[K]}\mathbf{v}_{t+1}^{i} = \overline{\mathbf{u}}_{t+1}.
\end{aligned}$$

We next evaluate the variance of $\overline{\mathbf{w}}_{t+1}$. Based on the facts that $\mathbb{E}[\mathbf{h}_i^H\mathbf{h}_i] = 1$, and $\forall i \neq j$, we have $\mathbb{E}[\mathbf{h}_i^H\mathbf{h}_j] = 0$, $\mathbb{V}\text{ar}[\mathbf{h}_i^H\mathbf{h}_j] = \frac{1}{M}$, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent, we have

$$\begin{aligned}
\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}\right\|^2 &= \mathbb{E}\left\|\frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_i(\mathbf{v}_{t+1}^{i}-\mathbf{w}_{t+1-E})\right. \\
&\left.+\frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E} - \frac{1}{K}\sum_{i\in[K]}\mathbf{v}_{t+1}^{i}\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{K}\sum_{k\in[K]}\hat{\mathbf{x}}_k - \frac{1}{K}\sum_{k\in[K]}\mathbf{x}_k\right\|^2 = \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\right. \\
&\left.+\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j + \mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k - \sum_{k\in[K]}\mathbf{x}_k\right\|^2 \\
&= \frac{1}{K^2}\left[\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\right\|^2\right. \\
&+\mathbb{E}\left\|\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\mathbf{x}_k\right\|^2 \\
&+ 2\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\right] \\
&+2\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\right] \\
&- 2\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\sum_{k\in[K]}\mathbf{x}_k\right] \\
&+2\mathbb{E}\left[\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\right] \\
&\left.- 2\mathbb{E}\left[\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\sum_{k\in[K]}\mathbf{x}_k\right]\right. \\
&\left.-2\mathbb{E}\left[\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\sum_{k\in[K]}\mathbf{x}_k\right]\right] \\
&= \frac{1}{K^2}\left[\left(1+\frac{1}{M}\right)\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 + \frac{K-1}{M}\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2\right. \\
&\left.+\frac{dK}{\mathsf{SNR}_{\mathsf{UL}}} + \sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 - 2\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2\right] \\
&= \frac{1}{K^2}\left[\frac{K}{M}\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 + \frac{\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2}{\mathsf{SNR}_{\mathsf{UL}}}\right] \\
&= \frac{1}{K^2}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\sum_{k\in[K]}\mathbb{E}\left\|\mathbf{x}_k\right\|^2 \\
&\leq \frac{1}{K^2}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\sum_{k\in[K]}E\sum_{i=t+1-E}^{t}\left\|\eta_i\nabla\tilde{f}_k(\mathbf{w}_i^k)\right\| \\
&\leq \frac{1}{K}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_{t+1-E}^2E^2H^2 \\
&\leq \frac{4}{K}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2E^2H^2,
\end{aligned}$$

where in the last inequality we use the fact that η_t is non-increasing and $\eta_{t+1-E} \leq 2\eta_t$. □

Lemma 4. *Let Assumptions 1-4 hold and downlink SNR scales* $\mathsf{SNR}_{\mathsf{DL}} \geq \frac{1-\mu\eta_t}{\eta_t^2}$ *as learning round* t. $\forall t+1 \in \mathcal{I}_E$, *we have* $\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \overline{\mathbf{w}}_{t+1}$, *and* $\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1}-\overline{\mathbf{w}}_{t+1}\right\|^2 \leq \left(\frac{dMK}{N^2(K+M)}\right)\frac{\eta_t^2}{1-\mu\eta_t}$.

Proof. We first show that

$$\mathbb{E}\left[\text{Re}\left(\frac{z_i^k}{g_k}\right)\right] = \text{Re}\left(\mathbb{E}\left[z_i^k\right]\frac{1}{\mathbb{E}\left[g_k\right]}\right) = 0,$$

and

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\text{Re}\left(\frac{z_i^k}{g_k}\right)\right] &= \mathbb{E}\left[\text{Re}\left(\frac{z_i^k}{g_k}\right)\text{Re}\left(\frac{z_i^{k^*}}{g_{k^*}}\right)\right] \\
&= \mathbb{E}\left[\text{Re}\left(\frac{z_i^k z_i^{k^*}}{g_k g_{k^*}}\right)\right] \leq \frac{\mathbb{E}\left[\text{Re}(z_i^{k^*}z_i^k)\right]}{\mathbb{E}\left[\text{Re}(g_k^*g_k)\right]} \\
&= \frac{1/(2\mathsf{SNR}_{DL})}{1/2(1+K/M)} = \left(\frac{M}{K+M}\right)\frac{1}{\mathsf{SNR}_{\mathsf{DL}}},
\end{aligned}$$

from which we can easily obtain $\mathbb{E}\left[\tilde{\mathbf{z}}_{t+1}^k\right] = \mathbf{0}$ and $\mathbb{V}\text{ar}\left[\tilde{\mathbf{z}}_{t+1}^k\right] = \left(\frac{M}{K+M}\right)\frac{d}{\mathsf{SNR}_{\mathsf{DL}}}$. Therefore, we have $\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_{t+1}^k + \frac{1}{N}\sum_{k\in[K]}\mathbb{E}\left[\tilde{\mathbf{z}}_{t+1}^k\right] = \overline{\mathbf{w}}_{t+1}$,

and $\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1}-\overline{\mathbf{w}}_{t+1}\right\|^2 = \mathbb{E}\left\|\frac{1}{N}\sum_{k\in[K]}\tilde{\mathbf{z}}_{t+1}^k\right\|^2 = \frac{1}{N^2}\sum_{k\in[K]}\mathbb{E}\left\|\tilde{\mathbf{z}}_{t+1}^k\right\|^2 = \left(\frac{MK}{N^2(K+M)}\right)\frac{d}{\mathsf{SNR}_{\mathsf{DL}}} \leq \left(\frac{dMK}{N^2(K+M)}\right)\frac{\eta_t^2}{1-\mu\eta_t}$.

□

APPENDIX C
PROOF OF THEOREM 1

We need to consider four cases for the analysis of the convergence of $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2$.

1) If $t\notin\mathcal{I}_E$ and $t+1\notin\mathcal{I}_E$, $\overline{\mathbf{v}}_{t+1}=\overline{\mathbf{w}}_{t+1}$ and $\overline{\mathbf{p}}_t=\overline{\mathbf{w}}_t$. Using Lemma 1, we have:

$$\begin{aligned}&\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1}-\mathbf{w}^*\right\|^2=\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2\\&\leq(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2+\eta_t^2\mathbb{E}\left\|\mathbf{g}_t-\overline{\mathbf{g}}_t\right\|^2+6L\eta_t^2\Gamma\end{aligned} \tag{27}$$

$$\begin{aligned}&+2\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^N\left\|\overline{\mathbf{w}}_t-\mathbf{w}_t^k\right\|^2\right]\leq(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{p}}_t-\mathbf{w}^*\right\|^2\\&+\eta_t^2\left[\sum_{k=1}^N\frac{\delta_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2\right].\end{aligned} \tag{28}$$

2) If $t\in\mathcal{I}_E$ and $t+1\notin\mathcal{I}_E$, we still have $\overline{\mathbf{v}}_{t+1}=\overline{\mathbf{w}}_{t+1}$. With $\overline{\mathbf{p}}_t=\overline{\mathbf{w}}_t+\frac{1}{N}\sum_{k=1}^N\tilde{\mathbf{z}}_t^k$, we have:

$$\begin{aligned}&\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2=\left\|\overline{\mathbf{p}}_t-\overline{\mathbf{w}}_t+\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2\\&=\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2+\underbrace{\left\|\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t\right\|^2}_{A_1}+\underbrace{2\left\langle\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t,\overline{\mathbf{p}}_t-\mathbf{w}^*\right\rangle}_{A_2}.\end{aligned}$$

We first note that the expectation of A_2 over the noise and fading channel randomness is zero since we have $\mathbb{E}\left[\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t\right]=\mathbf{0}$. Second, the expectation of A_1 can be bounded using Lemma 4. We then have

$$\begin{aligned}&\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2=\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2\\&\leq(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2+(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\overline{\mathbf{p}}_t\right\|^2\\&+\eta_t^2\left[\sum_{k=1}^N\frac{H_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2\right]\\&\leq(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2+\eta_t^2\left[\sum_{k=1}^N\frac{H_k^2}{N^2}+6L\Gamma\right.\\&\left.+8(E-1)^2H^2+\frac{MK}{N^2(K+M)}\right].\end{aligned} \tag{29}$$

3) If $t\notin\mathcal{I}_E$ and $t+1\in\mathcal{I}_E$, then we still have $\overline{\mathbf{p}}_t=\overline{\mathbf{w}}_t$. For $t+1$, we need to evaluate the convergence of $\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2$. We have

$$\begin{aligned}&\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2=\left\|\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}+\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2\\&=\underbrace{\left\|\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1}\right\|^2}_{B_1}+\underbrace{\left\|\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2}_{B_2}+\underbrace{2\left\langle\overline{\mathbf{w}}_{t+1}-\overline{\mathbf{u}}_{t+1},\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\rangle}_{B_3}.\end{aligned} \tag{30}$$

We first note that the expectation of B_3 over the noise is zero since we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{w}}_{t+1}\right]=\mathbf{0}$ and the expectation of B_1 can be bounded using Lemma 3. We next write B_2 into

$$\begin{aligned}&\left\|\overline{\mathbf{u}}_{t+1}-\mathbf{w}^*\right\|^2=\left\|\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1}+\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2\\&=\underbrace{\left\|\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1}\right\|^2}_{C_1}+\underbrace{\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2}_{C_2}+\underbrace{2\left\langle\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1},\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\rangle}_{C_3}.\end{aligned} \tag{31}$$

Similarly, the expectation of C_3 over the noise is zero since we have $\mathbb{E}\left[\overline{\mathbf{u}}_{t+1}-\overline{\mathbf{v}}_{t+1}\right]=\mathbf{0}$ and the expectation of C_1 can be bounded using Lemma 2. Therefore, we have

$$\begin{aligned}&\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2\leq\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2+\frac{4}{K}\left[\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2E^2H^2\\&+\frac{N-K}{N-1}\frac{4}{K}\eta_t^2E^2H^2\leq(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2\\&+\eta_t^2\left[\sum_{k=1}^N\frac{H_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2+\right.\\&\left.\frac{4}{K}\left(\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2+\frac{N-K}{N-1}\frac{4}{K}E^2H^2\right].\end{aligned} \tag{32}$$

4) If $t\in\mathcal{I}_E$ and $t+1\in\mathcal{I}_E$, $\overline{\mathbf{v}}_{t+1}\neq\overline{\mathbf{w}}_{t+1}$ and $\overline{\mathbf{p}}_t\neq\overline{\mathbf{w}}_t$. (Note that this is possible only for $E=1$.) Combining the results from the previous two cases, we have

$$\begin{aligned}&\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2\leq(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2\\&+\left[\sum_{k=1}^N\frac{H_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2\right.\\&+\frac{4}{K}\left(\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2+\frac{N-K}{N-1}\frac{4}{K}E^2H^2\\&\left.+\frac{MK}{N^2(K+M)}\right].\end{aligned} \tag{33}$$

Let $\Delta_t=\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2$. From (28), (29), (32) and (33), it is clear that no matter whether $t+1\in\mathcal{I}_E$ or $t+1\notin\mathcal{I}_E$, we always have $\Delta_{t+1}\leq(1-\eta_t\mu)\Delta_t+\eta_t^2B$, where $B=\sum_{k=1}^N\frac{H_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2+\frac{4}{K}\left(\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2+\frac{N-K}{N-1}\frac{4}{K}E^2H^2+\frac{MK}{N^2(K+M)}$. Define $v\triangleq\max\{\frac{4B}{\mu^2},(1+\gamma)\Delta_1\}$, by choosing $\eta_t=\frac{2}{\mu(t+\gamma)}$, we can prove $\Delta_t\leq\frac{v}{t+\gamma}$ by induction: $\Delta_{t+1}\leq\left(1-\frac{2}{t+\gamma}\right)\Delta_t+\frac{4B}{\mu^2(t+\gamma)^2}=\frac{t+\gamma-2}{(t+\gamma)^2}v+\frac{4B}{\mu^2(t+\gamma)^2}=\frac{t+\gamma-1}{(t+\gamma)^2}v+\left(\frac{4B}{\mu^2(t+\gamma)^2}-\frac{v}{(t+\gamma)^2}\right)\leq\frac{v}{t+\gamma+1}$. By the L-smoothness of f and $v\leq\frac{4B}{\mu^2}+(1+\gamma)\Delta_1$, we can prove the result in (23).

REFERENCES

[1] X. Wei, C. Shen, J. Yang, and H. V. Poor, "Random orthogonalization for federated learning in massive MIMO systems," in *Proc. IEEE International Conference on Communications (ICC)*, May 2022, pp. 1–6.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[3] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[4] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, p. e2024789118, 2021.

[5] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 46–51, 2020.

[6] 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, 2020.

[7] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7498–7513, 2020.

[8] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[9] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[10] 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, 2020.

[11] A. M. Elbir and S. Coleri, "Federated learning for hybrid beamforming in mm-Wave massive MIMO," *IEEE Commun. Letter*, vol. 24, no. 12, pp. 2795–2799, 2020.

[12] T. Huang, B. Ye, Z. Qu, B. Tang, L. Xie, and S. Lu, "Physical-layer arithmetic for federated learning in uplink MU-MIMO enabled wireless networks," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, 2020, pp. 1221–1230.

[13] Y.-S. Jeon, M. M. Amiri, J. Li, and H. V. Poor, "A compressive sensing approach for federated learning over massive MIMO communication systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1990–2004, 2020.

[14] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. International Conference on Learning Representations*, 2018.

[15] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. International Conference on Learning Representations*, 2020.

[16] K. Thonglek, K. Takahashi, K. Ichikawa, C. Nakasan, P. Leelaprute, and H. Iida, "Sparse communication for federated learning," in *Proc. IEEE 6th International Conference on Fog and Edge Computing (ICFEC)*, 2022, pp. 1–8.

[17] Y. Oh, Y.-S. Jeon, M. Chen, and W. Saad, "FedVQCS: Federated learning via vector quantized compressed sensing," *arXiv preprint arXiv:2204.07692*, 2022.

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

[19] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," *arXiv preprint arXiv:2006.10672*, 2020.

[20] Y. Du, S. Yang, and K. Huang, "High-dimensional stochastic gradient quantization for communication-efficient edge learning," *IEEE Trans. Signal Processing*, vol. 68, pp. 2128–2142, 2020.

[21] D. Wen, K.-J. Jeon, M. Bennis, and K. Huang, "Adaptive subcarrier, parameter, and power allocation for partitioned edge learning over broadband channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8348–8361, 2021.

[22] H. Chen, S. Huang, D. Zhang, M. Xiao, M. Skoglund, and H. V. Poor, "Federated learning over wireless IoT networks with optimized communication and resources," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16 592–16 605, 2022.

[23] S. Wang, M. Chen, C. Yin, W. Saad, C. S. Hong, S. Cui, and H. V. Poor, "Federated learning for task and resource allocation in wireless high-altitude balloon networks," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17 460–17 475, 2021.

[24] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

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

[26] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, 2020.

[27] J. Xu and H. Wang, "Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1188–1200, 2021.

[28] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Select. Areas Commun.*, vol. 40, no. 1, pp. 227–242, 2021.

[29] X. Ma, H. Sun, Q. Wang, and R. Q. Hu, "User scheduling for federated learning through over-the-air computation," in *Proc. IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, 2021, pp. 1–5.

[30] H.-S. Lee and J.-W. Lee, "Adaptive transmission scheduling in wireless networks for asynchronous federated learning," *IEEE J. Select. Areas Commun.*, vol. 39, no. 12, pp. 3673–3687, 2021.

[31] M. M. Wadu, S. Samarakoon, and M. Bennis, "Joint client scheduling and resource allocation under channel uncertainty in federated learning," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5962–5974, 2021.

[32] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Processing*, vol. 68, pp. 2897–2911, 2020.

[33] Z. Lin, X. Li, V. K. Lau, Y. Gong, and K. Huang, "Deploying federated learning in large-scale cellular networks: Spatial convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1542–1556, 2021.

[34] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Over-the-air federated learning with energy harvesting devices," in *Proc. IEEE Globecom*. IEEE, 2022, pp. 1942–1947.

[35] S. Wan, J. Lu, P. Fan, Y. Shao, C. Peng, and K. B. Letaief, "Convergence analysis and system design for federated learning over wireless networks," *IEEE J. Select. Areas Commun.*, vol. 39, no. 12, pp. 3622–3639, 2021.

[36] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Trans. Signal Processing*, vol. 69, pp. 3796–3811, 2021.

[37] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Time-correlated sparsification for efficient over-the-air model aggregation in wireless federated learning," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 1–6.

[38] T. T. Vu, H. Q. Ngo, M. N. Dao, D. T. Ngo, E. G. Larsson, and T. Le-Ngoc, "Energy-efficient massive MIMO for federated learning: Transmission designs and resource allocations," *arXiv preprint arXiv:2112.11723*, 2021.

[39] R. Hamdi, M. Chen, A. B. Said, M. Qaraqe, and H. V. Poor, "Federated learning over energy harvesting wireless networks," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 92–103, 2021.

[40] T. T. Vu, D. T. Ngo, H. Q. Ngo, M. N. Dao, N. H. Tran, and R. H. Middleton, "Joint resource allocation to minimize execution time of federated learning in cell-free massive MIMO," *IEEE Internet Things J.*, vol. Early access, 2022.

[41] 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, 2020.

[42] Y. Mu, N. Garg, and T. Ratnarajah, "Communication-efficient federated learning for massive MIMO systems," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 578–583.

[43] C. Xu, S. Liu, Z. Yang, Y. Huang, and K.-K. Wong, "Learning rate optimization for federated learning exploiting over-the-air computation," *IEEE J. Select. Areas Commun.*, vol. 39, no. 12, pp. 3742–3756, 2021.

[44] Z. Lin, Y. Gong, and K. Huang, "Distributed over-the-air computing for fast distributed optimization: Beamforming design and convergence analysis," *arXiv preprint arXiv:2204.06876*, 2022.

[45] C. Zhong, H. Yang, and X. Yuan, "Over-the-air federated multi-task learning over MIMO multiple access channels," *arXiv preprint arXiv:2112.13603*, 2021.

[46] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, "Fast convergence algorithm for analog federated learning," in *Proc. IEEE Int. Conf. Commun.*, June 2021, pp. 1–6.

[47] M. M. Amiri, T. M. Duman, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5129–5143, 2021.

[48] B. Tegin and T. M. Duman, "Blind federated learning at the wireless edge with low-resolution ADC and DAC," *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 7786–7798, 2021.

[49] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti, G. Vivier, E. De Carvalho, Y. Ji, Č. Stefanović *et al.*, "Towards massive connectivity support for scalable mmtc communications in 5g networks," *IEEE access*, vol. 6, pp. 28 969–28 992, 2018.

[50] N. Sidiropoulos and T. Davidson, "Broadcasting with channel state information," in *Proc. of 2004 Sensor Array and Multichannel Signal*, 2004, pp. 489–493.

[51] S. Sesia, I. Toufik, and M. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. Wiley, 2011.

[52] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[53] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Aspects of favorable propagation in massive MIMO," in *Proc. IEEE 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 76–80.

[54] L. P. Withers, R. M. Taylor, and D. M. Warme, "Echo-MIMO: A two-way channel training method for matched cooperative beamforming," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4419–4432, 2008.

[55] X. Zhou, T. A. Lamahewa, P. Sadeghi, and S. Durrani, "Two-way training: Optimal power allocation for pilot and data transmission," *IEEE transactions on wireless communications*, vol. 9, no. 2, pp. 564–569, 2010.

[56] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," in *Advances in Neural Information Processing Systems*, 2018, pp. 2525–2536.

[57] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE J. Select. Areas Commun.*, vol. 39, no. 7, pp. 2150–2167, July 2021.

[58] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1253–1268, 2022.

[59] X. Wei, C. Shen, J. Yang, and H. V. Poor, "Technical report: Random orthogonalization for federated learning in massive MIMO systems," http://www.ece.virginia.edu/~cs7dt/tech.pdf, University of Virginia, Tech. Rep., Aug. 2022.

[60] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in Neural Information Processing Systems*, vol. 26, pp. 315–323, 2013.

[61] X. Jiang, A. Decurninge, K. Gopala, F. Kaltenberger, M. Guillaud, D. Slock, and L. Deneire, "A framework for over-the-air reciprocity calibration for tdd massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 5975–5990, 2018.

[62] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, 2012.

[63] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan, "Perturbed iterate analysis for asynchronous stochastic optimization," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2202–2229, 2017.