Random Orthogonalization for Federated Learning in Massive MIMO Systems

Xizixiang Wei Cong Shen Jing Yang H. Vincent Poor

Abstract

We propose a novel communication design, coined *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 performances 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], [3]. This is because a FL task consists of multiple learning rounds,

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

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

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 [4] 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 [5], [6], so that the sum model can be directly obtained at the server. AirComp has attracted a lot of interest and a detailed literature review can be found in Section II.

However, majority of the existing works on AirComp have several limitations. First, these methods often require channel state information at 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 the channel reciprocity [7], [8] but still need independent pilot for every 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 the channel state information at receiver (CSIR). Second, majority of the AirComp literature essentially requires a channel inversion power control, which is well known to “blow up” when the channel is in deep fade [7]. Third, existing AirComp solutions almost exclusively apply only to the uplink communication phase in FL. The underlying philosophy has not been realized in the downlink communication phase.

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 investigations that explore the potential of MIMO for wireless FL. MIMO beamforming design to optimize FL has been studied in [9], [10]. Coding, quantization, and compressive sensing over a MIMO channel for FL has been studied in [11]–[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, e.g., in the uplink phase, one can use conventional MIMO decoders such as zero-forcing (ZF) or minimum mean-square-error (MMSE) to estimate each local model, and then compute the global model. However, this heuristic approach requires large channel estimation overhead, especially when the channel dimension is high. Decoding individual local models also makes it easier for the server to sketch the data distribution of a client, leading to potential privacy leakage. Moreover, matrix inversions in ZF or MMSE detectors are computationally demanding, in particular for massive MIMO. This increases the complexity and latency of the system.

This paper aims at designing simple-yet-effective FL communication methods that enable over-the-air computation for both uplink and downlink phases, to address the scalability challenge in FL. The novelty comes from a tight integration of MIMO and FL – our design explicitly utilizes the characteristics of both components. To illustrate the key idea, we start with *massive* MIMO where base station (BS) has a large number of antennas. The proposed framework only requires the BS to estimate a *summation channel*, which significantly alleviates the burden on channel estimation in FL[1]. Moreover, our approach is agnostic to the number of clients, and thus improves the scalability of FL. By leveraging the unique channel hardening and favorable propagation properties of massive MIMO, the proposed principle, coined *random orthogonalization*, allows the BS to directly compute the global model via a simple linear projection operation, thus achieving extremely low complexity and low latency for the uplink communication phase of FL. We then extend the random orthogonalization design to the downlink communication phase, which leads to a simple but highly effective model broadcast method for FL. As the random orthogonalization designs rely on channel hardening and favorable propagation to elimiate the interference, which does 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 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 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.

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 designs are proposed

[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.

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 conclusion of our work 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)), rigorous analyses of the convergence behavior have been carried out [14]–[16]. The approaches to reduce the payload size or frequency of FL communications include sparsification [17]–[19] and quantization [20]–[23]. There are also recent efforts to improve resource allocation [24]–[26]. Nevertheless, mostly of these works consider interference-free rate-limited communication links and often ignore the physical characteristics of wireless communication channels.

AirComp for FL. As a special case of computing over multiple access channels [27], AirComp [5], [6], [9], [28] leverages the signal superposition properties in a wireless multiple access channel to efficiently compute the function value. This technique has attracted a lot of interest as it could 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 [29]–[34]. Full CSIT is relaxed in [35] by only using the phase information of the channel. Several studies have provided convergence guarantees of Aircomp under different constraints and different types of heterogeneity [36]–[40]. However, most designs require individual CSIT for each client and focus on scalar channels.

Communication design for FL in MIMO systems. Recent years have also seen increased effort on the communication algorithm and system design for FL in MIMO systems. There are studies to optimize the communication efficiency and learning performance in MIMO systems for FL, including transmission power allocation [41]–[43], data rate [44], quantization and compression level [12], [45], and learning rate optimization [46]. Multiple beamforming designs have been proposed to improve the performance of FL [9], [47]–[49], yet these methods require full CSIT and rely on complex optimization methods to calculate beamformers, which is impractical in massive MIMO systems due to the high communication and computation cost. There is very limited study that relaxes the individual CSIT assumption in wireless FL over MIMO channels, with a notable exception of [50]. However, it only focuses on the uplink communication phase.

III. System Model

A. FL Model

The FL problem setting studied in this paper mostly follows the standard model in the original paper [2]. In particular, we consider a 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 given as 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 a 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 [16]. A typical wireless FL pipeline is illustrated in Fig. 1, which 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 over rounds.
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 $\tilde{p}_k = \frac{1}{K}$.

This work focuses on communication design of *both downlink and uplink* in the FL pipeline. In particular, we take advantages of the unique properties of (massive) MIMO to design efficient communication scheme and enable server aggregation over the air.

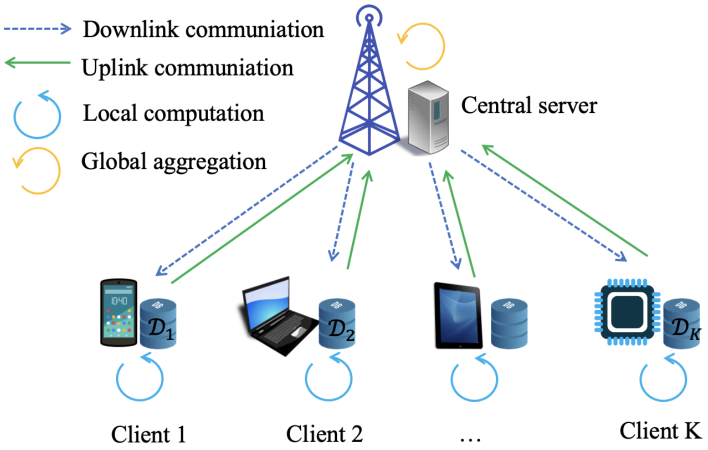


Fig. 1. The wireless FL pipeline.

B. Communication Model

Consider a MIMO time-division duplexing (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 step, each client transmits the differential between the received global model and the computed new local model

$$\mathrm{x}_t^k = \mathbf{w}_t - \mathbf{w}_{t+1}^k \in \mathbb{R}^d, \ \ \forall k \in [K] \tag{1}$$

to the BS, where $\mathrm{x}_t^k \triangleq [x_{1,t}^k, \cdots, x_{i,t}^k, \cdots, x_{d,t}^k]^T$. To simplify the notation, we omit index t by using $x_{k,i}$ instead of $x_{i,t}^k$ barring any confusion. We assume that each client transmits 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 $\mathbf{H} \triangleq [\mathbf{h}_1, \cdots, \mathbf{h}_K] \in \mathbb{C}^{M\times K}$ and $\mathbf{x}_i \triangleq [x_{1,i}, \cdots, x_{K,i}]^T \in \mathbb{R}^{K\times 1}, \forall i = 1, \cdots, d$, 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 decoders can be adopted to estimate $\hat{\mathbf{x}}_i = [\hat{x}_{1,i}, \cdots, \hat{x}_{K,i}]^T$. However, as discussed before, decoding $\{x_{k,i}\}_{i=1}^d$

[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 d.o.f. is utilized.

individually and obtaining the aggregated parameter $\tilde{x}_i \triangleq \sum_{k\in[K]} \hat{x}_{k,i}$ by a summation is inefficient. Note that after 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

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

After computing the global model $\mathbf{w}_{t+1} = [w_{1,t+1}, \cdots, w_{d,t+1}]^T$, the BS next 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

$$y_i^{\mathsf{DL}} = \sqrt{P_{\mathsf{BS}}}\mathbf{h}_{k,t+1}^H \mathbf{f} w_{i,t+1} + z_i^k, \ \ \forall i = 1, 2, \cdots, d, \tag{4}$$

where P_{BS} is the maximum transmission power of the BS and $\mathbf{z}_i$ is the downlink noise. Each client constructs the estimated global model $\hat{\mathbf{w}}_{t+1}$ and uses it as a new initial point for the next learning round after all d elements are received via (4). Traditionally, the design of the precoder $\mathbf{f}$ belongs to broadcasting common message with channel state information (CSI) (see [51] and the reference therein). However, these methods become impractical due to the difficulty in obtaining full CSI in massive MIMO systems, which motivates us to design $\mathbf{f}$ using only partial CSI. For mathematical simplicity, we assume normalized symbol power[4], i.e., $\mathbb{E}\left\|x_{k,i}\right\|^2 = 1$ and $\mathbb{E}\left\|w_{i,t+1}\right\|^2 = 1$; normalized Rayleigh block fading channel[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 communication and $\mathsf{SNR}_{\mathsf{DL}} \triangleq P_{\mathsf{BS}}/\sigma_{\mathsf{DL}}^2$ for downlink communication, and without loss of generality (w.l.o.g.) we set $P_{\mathsf{Client}} = 1$ and $P_{\mathsf{BS}} = 1$.

IV. Random Orthogonalization

We propose a new wireless FL framework by developing random orthogonalization with massive MIMO. With this principle, the global model can be directly obtained at the BS via a simple operation in uplink communications, and the global model can be broadcasted to clients efficiently in downlink communications. By exploring favorable propagation and channel hardening in massive MIMO, our proposed FL framework only requires *partial* CSI, which significantly reduces channel estimation overhead.

A. Uplink Communication Design

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

[4] The parameter normalization and de-normalization procedure in wireless FL follows the same as that in the Appendix of [5].

[5] Large-scale pathloss and shadowing effect is assumed to be taken care of by, e.g., open loop power control [52].

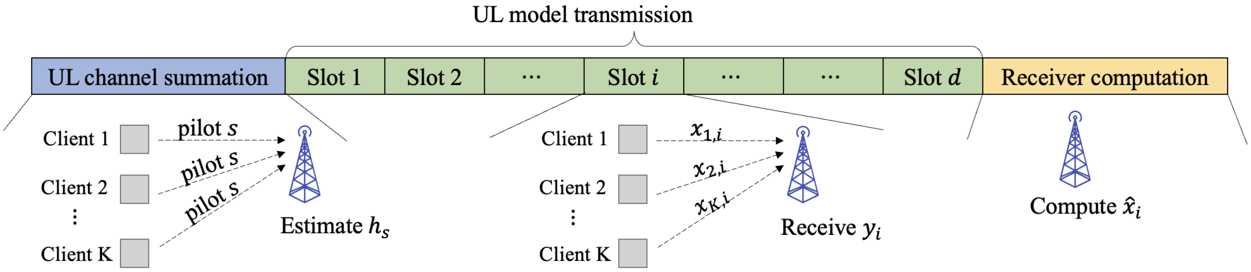


Fig. 2. An illustration of the proposed uplink FL design with massive MIMO.

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

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

so that the BS can estimate the *summation* of channel vectors $\mathbf{h}_s \triangleq \sum_{k\in[K]} \mathbf{h}_k$ from the received signal $\mathbf{y}_s$ (e.g., via a maximum likelihood estimator). 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.

(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. 2. 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]:

$$\mathbf{h}_k^H \mathbf{h}_k \to 1, \text{ as } M \to \infty.$$

In practical systems, when M is large but finite, 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

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

for the signal part of (6).

Favorable propagation. Since channels between different users are independent random vectors, massive MIMO also offers favorable propagation [53]:

$$\mathbf{h}_k^H \mathbf{h}_j \to 0, \text{ as } M \to \infty, \ \forall k \neq j.$$

Similarly, when M is finite, we have

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

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, which make the linear projection operation $\mathbf{h}_s^H \mathbf{y}_i$ an ideal fit for server aggregation of FL.

To gain some insight of random orthogonality, we approximate the average signal-to-interference-plus-

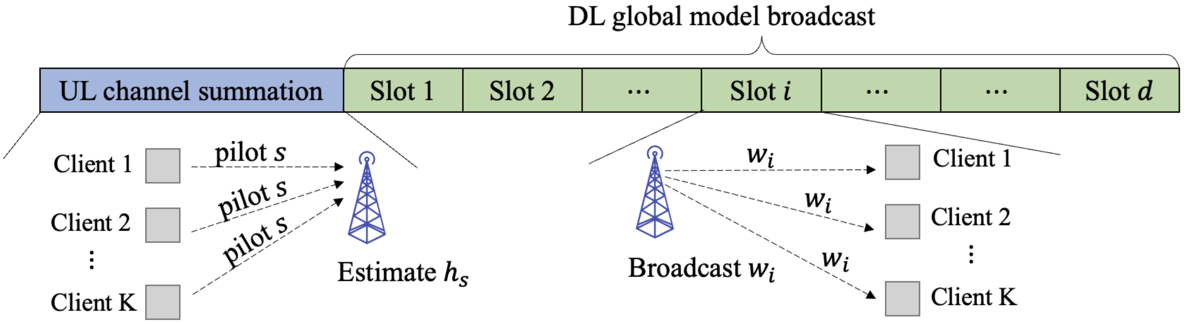


Fig. 3. An illustration of the proposed downlink FL design with massive MIMO.

noise-ratio (SINR) after the operation in (6) as

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

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.

B. Downlink Communication Design

Inspired by the uplink communication, 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 base station 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

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

The above two-step downlink communication procedure is illustrated in Fig. 3. 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

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

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

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

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 growth 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 [5], the proposed random orthogonalization does not require any individual CSIT. On the contrary, it only requires the receiver to estimate a summation channel* $\mathbf{h}_s$*, which is* $1/K$ *of the channel estimation overhead compared with the AirComp method in [9] or the traditional MIMO decoders. In downlink communications, traditional precoder design for common message broadcast requires CSIT for each client. By using summation channel* $\mathbf{h}_s$ *as precoder for global model broadcast, only partial CSIT is needed. Since we assume 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. Moreover, if the wireless channels are time-invariant during uplink and downlink communications, which happens when the channel coherence time is larger than the duration of one communication round, Steps U1 and D1 can be merged to further reduce the channel estimation overhead. Therefore, the proposed method is attractive in wireless FL design 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.*

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 models as long as channel*

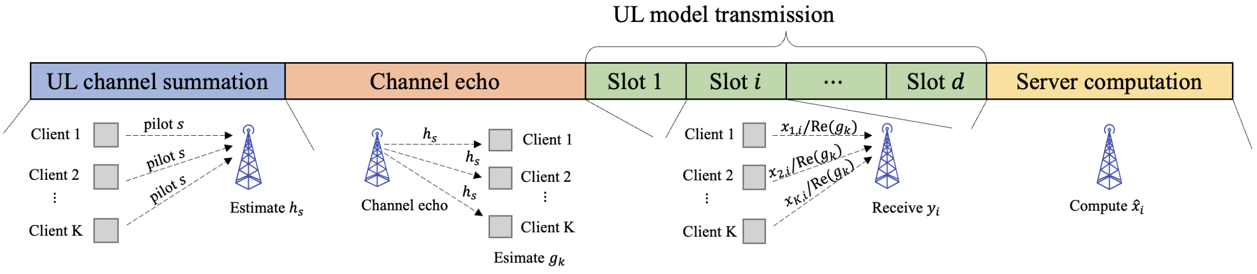


Fig. 4. An illustration of the proposed enhanced uplink FL design with massive MIMO.

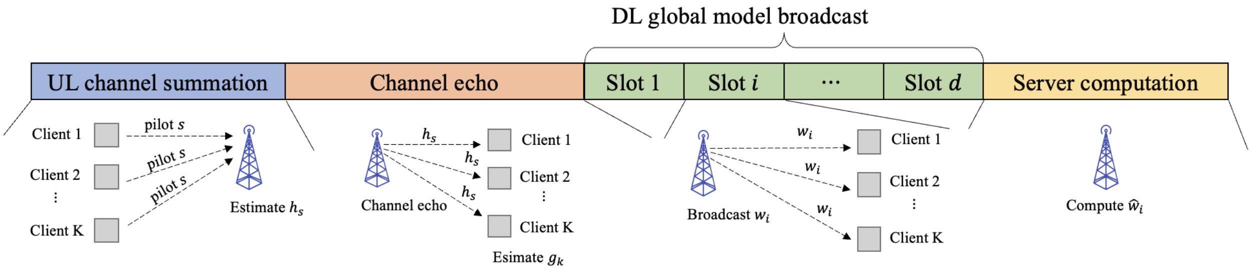


Fig. 5. An illustration of the proposed enhanced downlink FL design with massive MIMO.

hardening and favorable propagation are offered. In massive MIMO millimeter-wave (mmWave) communications, Rayleigh fading channel and light-of-sight (LOS) channel represent two extreme cases: rich scattering and no scattering. It is shown in [53] that both channel models offer asymptotically channel hardening and favorable propagation. We will discuss the general case that lies in between these two extremes in the next section as well as in the experiment results.

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 as discussed before, in case that channel hardening and favorable propagation 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 method in Section IV by taking advantage of channel echos.

Channel echo refers to that receiver sends whatever it receives back to the original transmitter directly as data payload. The main purpose of channel echo in uplink communication for FL is to “cancel” channel fading at each client. The enhanced design for uplink communication contains the following four main steps, which is demonstrated in Fig. 4.

(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 receives the estimate channel vector summation $\mathbf{h}_s = \sum_{j\in[K]} \mathbf{h}_k$.

(EU2) Downlink channel echo. The BS sends the previously received $\mathbf{h}_s$ (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 $g_k = \mathbf{h}_k^H \mathbf{h}_s = \mathbf{h}_k^H \sum_{j\in[K]} \mathbf{h}_k = \|\mathbf{h}_k\|^2 + \sum_{j\in[K],j\neq k}^{K} \mathbf{h}_k^H \mathbf{h}_j$.

(EU3) Uplink model transmission. All involved clients transmit local parameter $\{x_{k,i}/\text{Re}(g_k)\}_{k\in[K]}$ to the BS synchronously in d shared time slots:

$$\mathbf{y}_i = \sum_{k\in[K]}^{K} \mathbf{h}_k \frac{x_{k,i}}{\text{Re}(g_k)} + \mathbf{n}_i \;\; \forall i = 1, \cdots, d.$$

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

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

Similarly, as shown in Fig. 5, the enhanced design for downlink communication contains the following four main steps.

(ED1-2) Uplink channel summation and downlink channel echo. The first two steps in 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; hence the received signal at 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:

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

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 server and the clients, as shown in (16) and (17). The extra channel echo steps (Step EU2 in uplink and Step ED1 in downlink) enable 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 additional cost of using one extra slot for channel echo, and preserves all the other advantages of random orthogonalization.

Remark 3. *We note that 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 / time slots for channel estimation. Moreover, 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/ED*4*. 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 [5] also use "channel inversion" to pre-cancel channel fading. However, our design outperforms the method in [5] because the latter requires full CSIT, which leads to a large channel estimation overhead even with channel reciprocity in TDD systems. On the contrary, our method only requires partial CSI which is efficiently obtained via channel echos. Moreover, analog aggregation does not naturally extend to MIMO systems where the uplink channels become vectors, which makes channel inversions at the transmitters nontrivial.*

VI. Performance Analyses

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

A. Cramer-Rao Lower Bounds

In uplink communications, 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 that $\mathbf{y}_i \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathsf{UL}}, \frac{1}{\mathsf{SNR}}\mathbf{I})$. To leverage CRLBs to evaluate parameter estimation, we should first derive the Fisher information of $\mathbf{x}_i$. Fisher information of parameters in AWGN is given by Example 3.9 in [54]. Therefore, the Fisher information matrix (FIM) of the estimation of $\mathbf{x}_i$ is

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

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

$$F_{\mathsf{DL}} = 2 \cdot \mathsf{SNR} \cdot \text{Re}\left[\frac{\partial^H \mu_{\mathsf{DL}}(w_i)}{\partial w_i} \frac{\partial \mu_{\mathsf{DL}}(w_i)}{\partial w_i}\right] = 2 \cdot \mathsf{SNR} \cdot \text{Re}(\mathbf{h}_k^H \mathbf{h}_s \mathbf{h}_s^H \mathbf{h}_k).$$

The CRLBs of estimates are then given by the inverse of the Fisher information: $\mathbf{C}_{\hat{\mathbf{x}}_i} = \mathbf{F}_{\mathsf{UL}}^{-1}$ and $C_{\hat{w}_i} = 1/F_{\mathsf{DL}}$. CRLBs are the lower bounds on the variance 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 an unbiased estimation 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. This will also be validated in the expriment 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 uplink and downlink communications, respectively, there would be four cases of convergence analysis. Since these convergence analyses are quite similar, here 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; see [14], [16], [55], [56]. 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 [16] and its many 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* $\mathbf{w}$, $\left\|f_k(\mathbf{v}) - f_k(\mathbf{w})\right\| \leq L \left\|\mathbf{v} - \mathbf{w}\right\|$*;*

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

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 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 [57] – downlink transmission power should scale in the order of $\mathcal{O}(t^2)$.

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

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)}\left[\frac{4B}{\mu^2} + (1+\gamma)\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2\right], \tag{18}$$

for any $t \geq 1$, where

$$B \triangleq \sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2 + \frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2 + \frac{dMK}{N^2(K+M)}. \tag{19}$$

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

Theorem 1 shows that applying random othogonalization in uplink communication and enhanced method in downlink communication preserve 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 rounds and the fraction of partial 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 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* (18) *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 [57]. This gap is also empirically observed in the experiments; see Section VII-B. However, we also note that this gap term is inversely proportional to the number of antennas M. Hence, as M becomes large, it reduces to zero asymptotically. The technical details for this remark can be found in the additional convergence analysis Appendix-D.*

We next explore the relationship between the number of selected clients K and the number of BS antennas M to analyze 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 uplink communication, assuming that downlink communication is error-free. Note that this configuration is reasonable in case that 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 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 $\eta_t = \frac{2}{\mu(t+\gamma)}$, $\forall t \in [T]$, the following inequality holds:*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)} \left[\frac{4\tilde{B}}{\mu^2} + (1+\gamma) \left\| \mathbf{w}_0 - \mathbf{w}^* \right\|^2 \right] \tag{20}$$

for any $t \geq 1$, where

$$\tilde{B} \triangleq \left[1 + \frac{K}{M} + \frac{1}{\mathsf{SNR}} \right] \frac{H^2}{K}. \tag{21}$$

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 ideal 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 influence the convergence rate of FL with MIMO: **variance reduction** and **channel interference and noise**. In particular, the definition of $\tilde{B}$ in (21), 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 [58], which is captured by the $\frac{H^2}{K}$ term in (21). 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 (21). This suggests that the desired variance reduction may be adversely impacted if channel interference/noise dominates the convergence bound. In particular, when $M >> K$, we have $\frac{1}{K} >> \frac{1}{M}$, and the system enjoys almost the same variance reduction as the interference-free case. However, in the case of $K >> M$, we have $\frac{(K/M+1/\mathsf{SNR})}{K} \approx \frac{1}{M} >> \frac{1}{K}$, and $\frac{H^2}{M}$ dominates the convergence bound. In this case, blindly increasing the number of clients is unwise, as it does not have the advantage of variance reduction.

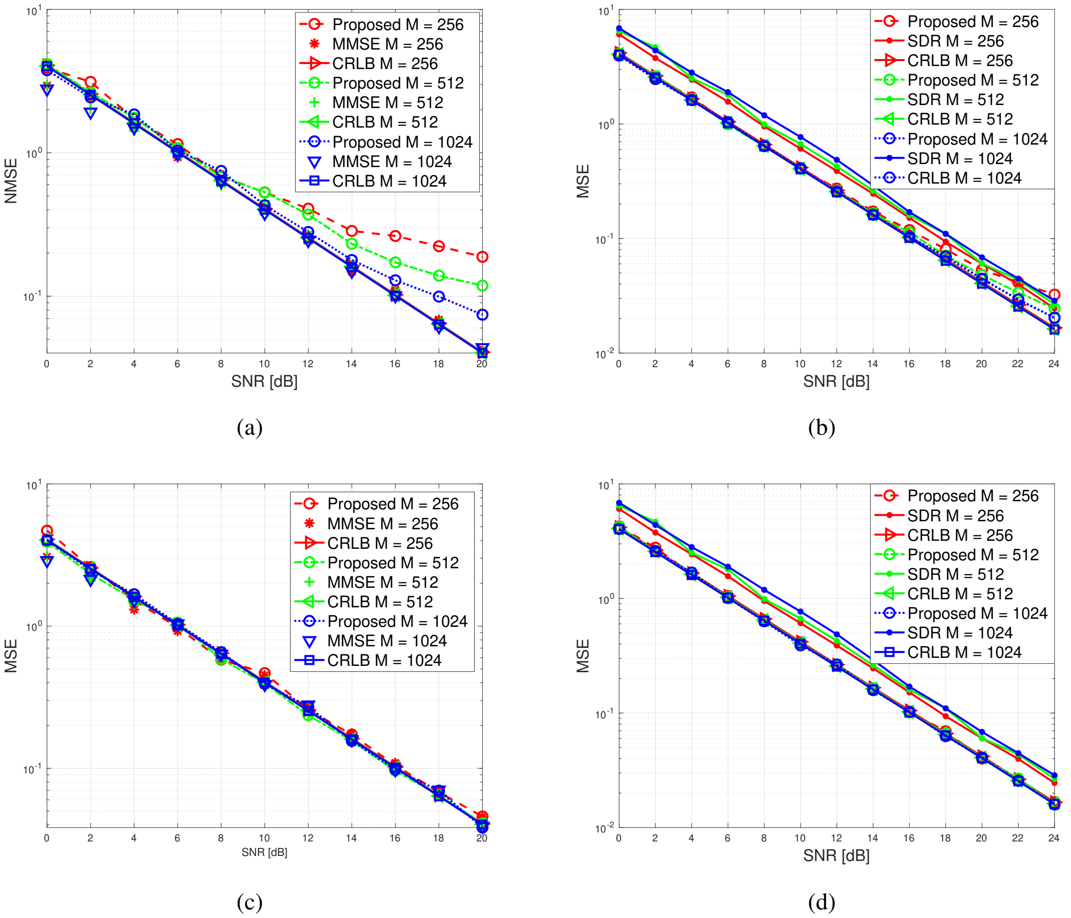
Remark 6. *In massive MIMO, a BS is usually equipped with many (up to hundreds) antennas. Although there may be large number of users participating in FL, only a small number of them are simultaneously active [9]. Both factors indicate that $K << M$ often holds in typical massive MIMO systems. The analysis reveals that our proposed framework enjoys nearly the same interference- 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 methods for uplink and downlink FL communications through numerical experiments. From a communication performance perspective, we compare the proposed methods with the classic MIMO detector and beamformer with respect to the mean squared error (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 approximate channel hardening and favorable propagation are not fully offered and channel estimation is imperfect. We further verify the effectiveness of the proposed methods via real-world FL tasks. The designed communication framework demonstrates convincing performance in both linear and non-linear FL tasks on two widely accepted datasets.

A. Communication Performance

We consider a massive MIMO BS with $M = 256, 512$, and 1024 antennas, with $K = 8$ active users participating in a 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 in simulations are obtained from $2,000$ Monte Carlo experiments. We use CRLB derived in Section VI-A as the benchmark of the computed MSEs. In



(a)

(b)

(c)

(d)

Fig. 6. MSE of the received global ML model parameters versus SNR of random orthogonalization method in uplink (a) and downlink (b) communications; and of the enhanced method in uplink (c) and downlink (d) communications.

addition, we adopt traditional MIMO MMSE decoder and the semidefinite relaxation based (SDR-based) beamformer design method in [51] for performance comparisons of uplink and downlink communications, respectively.

Communication effectiveness. Fig. 6(a) and Fig. 6(b) compare the MSE performance of the random othogonalization scheme in uplink (a) and downlink (b) communications with classic MIMO decoder/beamformer as well as CRLB under different system SNRs. As illustrated in these two plots, the proposed method performs nearly identically to CRLB in low and moderate SNRs under different antenna configurations (see SNR $\leq$ 12 dB for uplink and SNR $\leq$ 18 dB for downlink). As the SNR increases, the dominant factor affecting system performances becomes the interference among different users. In uplink communications, unlike the MMSE decoder that can cancel all interferences when $K \leq M$ at high SNR, Eqn. (11) shows that, for a given K and M, the proposed framework has a fixed (approximate)

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

# antennas	Total CPU time (second)		Ratio	Total CPU time (second)		Ratio
(M)	RO-UL	MMSE	RO-UL/MMSE	Enhanced-UL	MMSE	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	Total CPU time (second)		Ratio	Total CPU time (second)		Ratio
(M)	RO-DL	SDR	RO-DL/SDR	Enhanced-DL	SDR	Enhanced-DL/SDR
256	0.0157	25.1492	0.062%	0.0163	28.8593	0.68%
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%

$\mathsf{SIR} = \frac{K-1}{M}$ as $\mathsf{SNR} \to \infty$, which explains why the performance of the proposed scheme deteriorates compared with MMSE at high SNR. However, this issue disappears naturally as the number of BS antennas increases. It can be seen in Fig. 6(a) that the performance gap between the proposed method and CRLB reduces, from about 7 dB when $M = 256$ to about 2 dB when $M = 1024$ at $\mathsf{SNR} = 20$ dB, in uplink communications. We note that, although random othogonalization produces higher MSEs than MMSE (a lower equivalent SINR), FL tasks remain 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 othogonalization scheme hardly harms the convergence of FL in uplink communications. Similar to uplink, random othogonalization performs nearly identically to CRLB in low and moderate SNRs in downlink, and only loses about $0.5 \sim 3$ dB under different antenna configurations at $\mathsf{SNR} = 24$ dB. Moreover, random othogonalization outperforms SDR-based method nearly at all SNRs and antenna configurations. Due to the sub-optimality of SDR, SDR-based method has about 1.5 dB loss over CRLB. We need to emphasize here that our method only requires $1/K$ of channel estimation overheard (partial CSI) compared with MMSE or SDR-based method (full CSI), and this advantage is more significant when the BS is equipped with larger number of antennas.

Similarly, Fig. 6(c) and Fig. 6(d) compare the MSE performance of the enhanced method in uplink (c) and downlink (d) communications with MMSE decoder / SDR-based beamformer. It is clear in both plots that the enhanced method achieves MSEs very closed to the CRLBs. Furthermore, it performs nearly identically as the MMSE decoder in uplink and outperforms SDR-based method by about 1.5 dB in downlink. Therefore, by introducing channel echos, the enhanced method achieves excellent performance at relatively low additional resource cost.

Communication efficiency. We next focus on the low-latency benefits of the proposed methods. Table I compares the computational time of the proposed schemes with MMSE decoder and SDR-based beamformer when $\mathsf{SNR} = 10$ dB in uplink and downlink communications, respectively. The total CPU time is the *cumulative time* of each algorithm over $2,000$ Monte Carlo experiments. We see that the time consumption of random orthogonalization and enhanced method is much less than that of the MMSE decoder and SDR-based beamformer. Especially, when $M = 1024$, despite the 0.3 dB NMSE performance loss compared with the MMSE decoder in uplink of random orthogonalization method (as shown in Fig. 6(a)), the computation time of the proposed method is only 0.05% of the MMSE baseline. Since SDR is in general of $O(M^3)$ complexity, the proposed methods are even more computationally efficient for downlink communications, as the total CPU time is less than 0.1% of 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 nearly identical MSE performance to CRLB but require much less channel estimation overhead and achieve extremely lower system latency than classic MIMO decoders and beamformers.

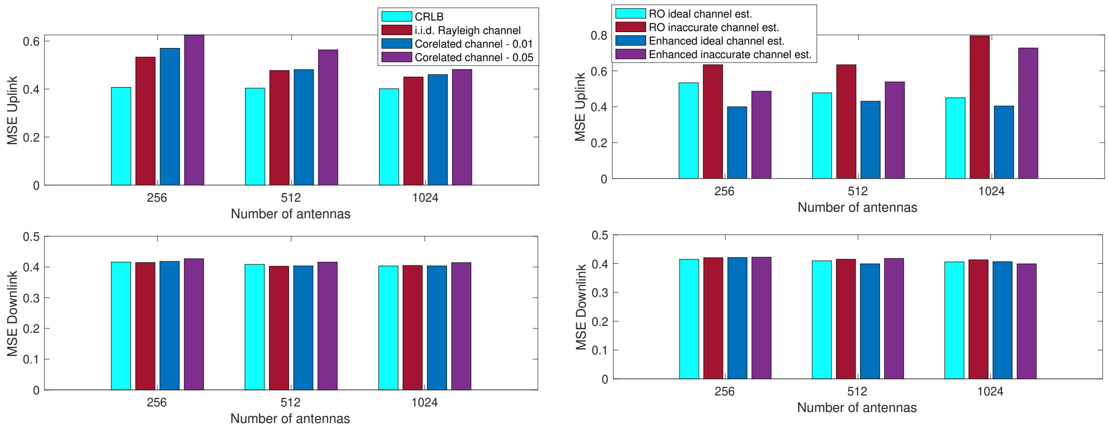
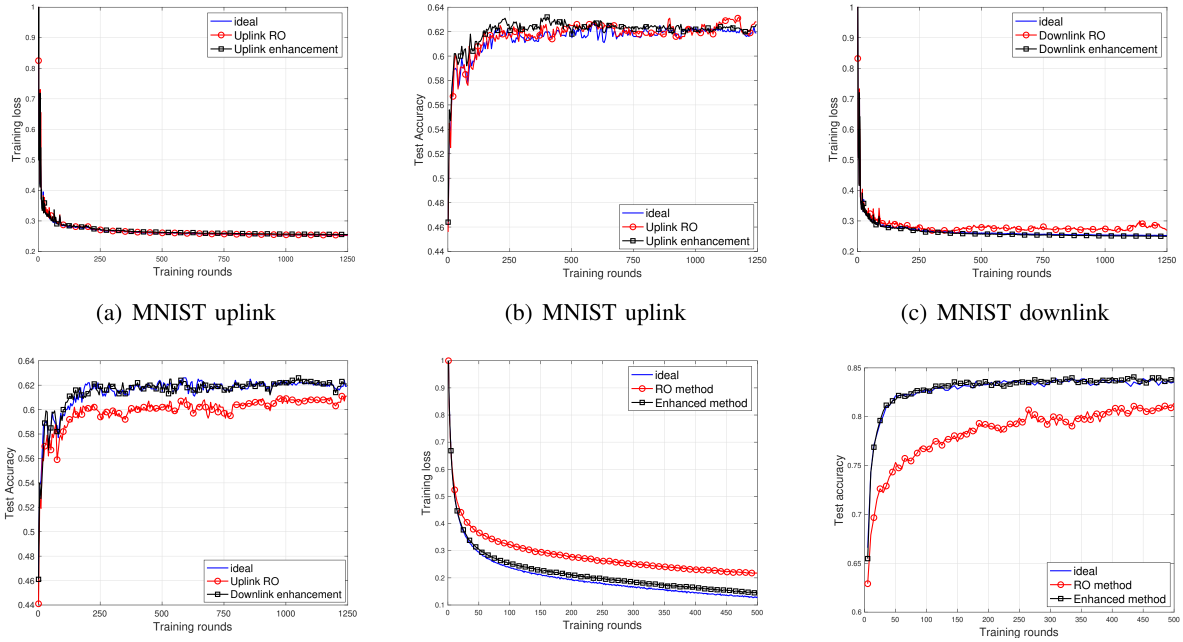


Fig. 7. MSE comparison of the received global ML model parameters when channel hardening and favorable propagation are not fully offered (left) and channel estimation is imperfect (right).

Robustness. We next focus on the robustness of both proposed schemes, and we evaluate them via the MSEs of the global model parameters obtained at $\mathsf{SNR} = 10$ dB through $2,000$ Monte Carlo experiments. Random orthogonalization assumes that the massive MIMO channels offer (approximate) channel hardening and favorable propagation. In case that this assumption is not fully valid, left sub-figure in Fig. 7 shows the achieved MSEs when wireless channels are correlated. We consider two correlated channel models with covariance matrix elements equal to 1 on the diagonal and equal to 0.01 or 0.05 off the diagonal, respectively. We see that when the off-diagonal elements are 0.01, random orthogonalization performs nearly identically as in the ideal i.i.d. Rayleigh fading channel case. Even when the off-diagonal elements equal to 0.05, the achieved MSE only increases by less than 1 dB in the worst case (when

$M = 256$). The MSE becomes more close to that of the i.i.d. Rayleigh channel case when M increases, as larger antenna arrays offer higher orthogonality.

Since the proposed method requires estimating the summation channel $\mathbf{h}_s$ (and g_k in the enhanced method), it is useful to evaluate the MSE performance when such channel estimation is imperfect. The right sub-figure of Fig. 7 compares the MSEs of both proposed methods when channel estimation is obtained with a pilot SNR of SNR $= 20$ dB. We see that downlink communication is more robust than uplink – the former achieves nearly identical MSE as the ideal case even when channel estimation is inaccurate. For uplink, imperfect channel estimation increases the MSEs by $1 \sim 3$ dB depending on different antenna configurations. However, we need to emphasize again that FL tasks have the same convergence rate under a constant SINR in uplink communications (thanks to model differential transmission).



(a) MNIST uplink

(b) MNIST uplink

(c) MNIST downlink

(d) MNIST downlink

(e) CIFAR-10 uplink+downlink

(f) CIFAR-10 uplink+downlink

Fig. 8. Comparison of training loss and test accuracy. (a) and (b): a SVM FL task with ideal uplink communication (interference and noise free), random orthogonalization and enhanced method; (c) and (d): a SVM FL task with ideal downlink communication (interference and noise free), random orthogonalization and enhanced method; (e) and (f): a CIFAR classification FL task with ideal uplink and downlink communications (interference and noise free), random orthogonalization and enhanced method.

B. Learning Performance

To evaluate ML performance, we carry out experiments of FL classification tasks using two well-adopted real-world datasets: MNIST and CIFAR-10, via support vector machine (SVM) model and convolution neural network (CNN) model, respectively. The experiment results reveal that both models have promising test accuracy and convergence rate with the proposed communication designs.

MNIST-SVM. We implement a SVM to classify even and odd numbers in the MNIST handwritten-digit dataset [59], with $d = 784$. Total participants are set as $N = 20$, the size of each local dataset is 500, the size of the test set is $2,000$, and $E = 1$. The local dataset can be regarded as non-i.i.d. since we only allocate data of one label to each participant. 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. 8(a) and Fig. 8(b) report the training loss and test accuracy when the uplink adopts the proposed methods, 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 enhanced method behave almost identically as the ideal case where both uplink and downlink are perfect. Note that although the global model received at the BS during the learning process has noise and interference, the actual learning performances of the two methods do not deteriorate. Due to the model differential transmission in uplink communications, the effective SINR of the received global model gradually increases as the model converges, despite constant system noise and interference. Fig. 8(c) and Fig. 8(d) demonstrate the learning performance when the proposed designs are applied to the downlink communication. Since model differential is not feasible, we set the initial downlink SNR as 0 dB and scale at the rate of $\mathcal{O}(t^2)$ as learning progresses (see [57]). We notice that the learning performance of the enhanced method is almost identical to the ideal case, while there exists a performance gap of about 2% test accuracy loss of random othogonalization method.

CIFAR-CNN. We train a CNN model with two 5×5 convolution layers (both with 64 channels), two fully connected layers (384 and 192 units respectively) with ReLU activation and a final output layer with softmax. The two convolution layers are both followed by 2×2 max pooling and a local response norm layer. In 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. Uplink SNR is set as 10 dB and the initial downlink SNR is set as 0 dB and scales at the rate of $\mathcal{O}(t^2)$ as learning progresses.

Fig. 8(e) and Fig. 8(f) illustrate the training loss and test accuracy versus learning rounds when *both* the uplink and downlink communication adopt the random orthogonalization methods or the enhanced methods, respectively. It is seen that the enhanced methods achieve similar training loss and test accuracy to the ideal case. Due to the constant interference in the downlink communication, The random orthogonalization method occurs a performance gap of about 3% test accuracy loss.

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

VIII. Conclusion

Leveraging the unique characteristics of channel hardening and favorable propagation in massive MIMO, we have proposed a novel uplink communication method, coined *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 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.

Appendix A
Preliminaries

With a slight abuse of notation, we change the timeline to be with respect to the overall SGD iteration time steps instead of the communication rounds, i.e.,

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

Note that the global model $\mathbf{w}_t$ is only accessible at the clients for specific $t \in \mathcal{I}_E$, where $\mathcal{I}_E = \{nE \mid n = 1, 2, \ldots\}$, 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 [60]. In particular, we first define the following local training variables for client k:

$$\mathbf{v}_{t+1}^k \triangleq \mathbf{p}_t^k - \eta_t \nabla \tilde{f}_k(\mathbf{p}_t^k);$$

$$\mathbf{u}_{t+1}^{k} \triangleq \begin{cases} \mathbf{v}_{t+1}^{k} & \text{if } t+1 \notin \mathcal{I}_E, \\ \frac{1}{K}\sum_{i\in[K]} \mathbf{v}_{t+1}^{i} & \text{if } t+1 \in \mathcal{I}_E; \end{cases}$$

$$\mathbf{w}_{t+1}^{k} \triangleq \begin{cases} \mathbf{v}_{t+1}^{k} & \text{if } t+1 \notin \mathcal{I}_E, \\ \frac{1}{K}\sum_{i\in[K]} \mathbf{h}_s^H \mathbf{h}_i(\mathbf{v}_{t+1}^{i} - \mathbf{w}_{t+1-E}) & \\ +\frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E} & \text{if } t+1 \in \mathcal{I}_E; \end{cases}$$

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

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

$$\tilde{\mathbf{z}}_{t+1}^{k} \triangleq \sqrt{K}\left[\mathrm{Re}(z_1^k/g_1), \cdots, \mathrm{Re}(z_i^k/g_k) \cdots, \mathrm{Re}(z_d^k/g_d)\right]^H \in \mathbb{C}^{d\times 1} \;\; \forall k \in [K] \quad \text{and} \quad \tilde{\mathbf{z}}_{t+1}^{k} = 0 \;\; \forall k \notin [K]$$

are the downlink noise in (12), respectively. Then, we construct the following *virtual sequences*:

$$\overline{\mathbf{v}}_t = \frac{1}{N}\sum_{k=1}^{N} \mathbf{u}_t^k, \;\; \overline{\mathbf{u}}_t = \frac{1}{N}\sum_{k=1}^{N} \mathbf{v}_t^k, \;\; \overline{\mathbf{w}}_t = \frac{1}{N}\sum_{k=1}^{N} \mathbf{w}_t^k, \;\; \text{and} \;\; \overline{\mathbf{p}}_t = \frac{1}{N}\sum_{k=1}^{N} \mathbf{p}_t^k.$$

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

$$\mathbf{w}_{t+1} \triangleq \frac{1}{K}\sum_{k\in[K]} \mathbf{w}_{t+1}^k = \mathbf{w}_{t+1}^k = \frac{1}{N}\sum_{k=1}^{N} \mathbf{w}_{t+1}^k = \overline{\mathbf{w}}_{t+1}. \tag{22}$$

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 ≥ 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*

$$\mathbb{E}[\overline{\mathbf{u}}_{t+1}] = \overline{\mathbf{v}}_{t+1}, \;\; \textit{and} \;\; \mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{u}}_{t+1}\|^2 \le \frac{N-K}{N-1}\frac{4}{K}\eta_t^2 E^2 H^2.$$

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 omited due to space limitation.

Lemma 3. *Let Assumptions 1-4 hold. With $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$ and $\forall t+1 \in \mathcal{I}_E$, we have*

$$\mathbb{E}\left[\overline{\mathbf{w}}_{t+1}\right] = \overline{\mathbf{u}}_{t+1}, \ \ \textit{and} \ \ \mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 \leq \frac{4}{K}\left[\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2 E^2 H^2.$$

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 $\mathbb{E}[\mathbf{h}_i^H\mathbf{h}_j] = 0$, $\mathbb{V}\mathrm{ar}[\mathbf{h}_i^H\mathbf{h}_j] = \frac{1}{M}$, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent, we have

$$\begin{aligned}
\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\right\|^2 &= \mathbb{E}\left\|\frac{1}{K}\sum_{i\in[K]}\mathbf{h}_s^H\mathbf{h}_i(\mathbf{v}_{t+1}^i - \mathbf{w}_{t+1-E}) + \frac{1}{K}\mathbf{N}_{t+1}\mathbf{h}_s + \mathbf{w}_{t+1-E} - \frac{1}{K}\sum_{i\in[K]}\mathbf{v}_{t+1}^i\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{K}\sum_{k\in[K]}\hat{\mathbf{x}}_k - \frac{1}{K}\sum_{k\in[K]}\mathbf{x}_k\right\|^2 \\
&= \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k + \sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j + \mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k - \sum_{k\in[K]}\mathbf{x}_k\right\|^2 \\
&= \frac{1}{K^2}\Bigg[\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\right\|^2 + \mathbb{E}\left\|\mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\mathbf{x}_k\right\|^2
\end{aligned}$$

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

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 $\mathbb{E}\left[\mathrm{Re}\left(\frac{z_i^k}{g_k}\right)\right]=\mathrm{Re}\left(\mathbb{E}\left[z_i^k\right]\frac{1}{\mathbb{E}[g_k]}\right)=0$, and $\mathbb{V}\mathrm{ar}\left[\mathrm{Re}\left(\frac{z_i^k}{g_k}\right)\right]\leq\frac{\mathbb{E}\left[\mathrm{Re}({z_i^k}^H z_i^k)\right]}{\mathbb{E}[\mathrm{Re}(g_k^H g_k)]}=\frac{1/(2\mathsf{SNR}_{DL})}{1/2(1+K/M)}=\left(\frac{M}{K+M}\right)\frac{1}{\mathsf{SNR}_{\mathsf{DL}}}$, from which we can easily obtain $\mathbb{E}\left[\tilde{\mathbf{z}}_{t+1}^k\right]=\mathbf{0}$ and $\mathbb{V}\mathrm{ar}\left[\tilde{\mathbf{z}}_{t+1}^k\right]=\left(\frac{M}{K+M}\right)\frac{d}{\mathsf{SNR}_{\mathsf{DL}}}$. Therefore, we have $\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right]=\frac{1}{N}\sum_{k=1}^N\mathbb{E}\left[\mathbf{w}_{t+1}^k+\tilde{\mathbf{z}}_{t+1}^k\right]=\frac{1}{N}\sum_{k=1}^N\mathbf{w}_{t+1}^k+$ $\frac{1}{N}\sum_{k\in[K]}\mathbb{E}\left[\tilde{\mathbf{z}}_{t+1}^k\right]=\overline{\mathbf{w}}_{t+1}$, 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:

$$
\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2=\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|^2\leq(1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2+\eta_t^2\left[\sum_{k=1}^N\frac{H_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2\right]. \tag{23}
$$

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:

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

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

$$\begin{aligned}
&\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 = \mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \\
&\leq (1-\eta_t\mu)\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + (1-\eta_t\mu)\mathbb{E}\|\overline{\mathbf{w}}_t - \overline{\mathbf{p}}_t\|^2 + \eta_t^2\left[\sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right] \\
&\leq (1-\eta_t\mu)\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \eta_t^2\left[\sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{dMK}{N^2(K+M)}\right]. \qquad (24)
\end{aligned}$$

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 $\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2$. We have

$$\begin{aligned}
\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &= \|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1} + \overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\|^2 \\
&= \underbrace{\|\overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}\|^2}_{B_1} + \underbrace{\|\overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\|^2}_{B_2} + \underbrace{2\left\langle \overline{\mathbf{w}}_{t+1} - \overline{\mathbf{u}}_{t+1}, \overline{\mathbf{u}}_{t+1} - \mathbf{w}^* \right\rangle}_{B_3}.
\end{aligned} \qquad (25)$$

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

$$\begin{aligned}
\|\overline{\mathbf{u}}_{t+1} - \mathbf{w}^*\|^2 &= \|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1} + \overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 \\
&= \underbrace{\|\overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}\|^2}_{C_1} + \underbrace{\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2}_{C_2} + \underbrace{2\left\langle \overline{\mathbf{u}}_{t+1} - \overline{\mathbf{v}}_{t+1}, \overline{\mathbf{v}}_{t+1} - \mathbf{w}^* \right\rangle}_{C_3}.
\end{aligned} \qquad (26)$$

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

$$\begin{aligned}
&\mathbb{E}\|\overline{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^*\|^2 + \frac{4}{K}\left[\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right]\eta_t^2E^2H^2 + \frac{N-K}{N-1}\frac{4}{K}\eta_t^2E^2H^2 \\
&\leq (1-\eta_t\mu)\mathbb{E}\|\overline{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \eta_t^2\left[\sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2\right. \qquad (27) \\
&\left. + \frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2\right].
\end{aligned}$$

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

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|^2 \leq (1-\eta_t\mu)\mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2$$
$$+\left[\sum_{k=1}^{N}\frac{H_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2+\frac{4}{K}\left(\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2+\frac{N-K}{N-1}\frac{4}{K}E^2H^2+\frac{dMK}{N^2(K+M)}\right]. \tag{28}$$

Let $\Delta_t = \mathbb{E}\left\|\overline{\mathbf{w}}_t-\mathbf{w}^*\right\|^2$. From (23), (24), (27) and (28), 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 $B=\sum_{k=1}^{N}\frac{H_k^2}{N^2}+6L\Gamma+8(E-1)^2H^2+\frac{4}{K}\left(\frac{K}{M}+\frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2+\frac{N-K}{N-1}\frac{4}{K}E^2H^2+\frac{dMK}{N^2(K+M)}$. 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:

$$\begin{aligned}\Delta_{t+1}&\leq\left(1-\frac{2}{t+\gamma}\right)\Delta_t+\frac{4B}{\mu^2(t+\gamma)^2}=\frac{t+\gamma-2}{(t+\gamma)^2}v+\frac{4B}{\mu^2(t+\gamma)^2}\\&=\frac{t+\gamma-1}{(t+\gamma)^2}v+\left(\frac{4B}{\mu^2(t+\gamma)^2}-\frac{v}{(t+\gamma)^2}\right)\leq\frac{v}{t+\gamma+1}.\end{aligned}$$

By the L-smoothness of f and $v\leq\frac{4B}{\mu^2}+(1+\gamma)\Delta_1$, we can prove the result in (18).

Appendix D
Additional Convergence Analysis

In this subsection, we provide convergence analysis when random orthogonalization is adopted in both uplink and downlink communications. Due to the change on downlink communication mechanism, we need to change of the definition of $\mathbf{p}_{t+1}^k$ in Appendix A, while keeping the same definition of other local training variables:

$$\mathbf{p}_{t+1}^k\triangleq\begin{cases}\mathbf{v}_{t+1}^k & \text{if } t+1\notin\mathcal{I}_E,\\ \mathbf{h}_k^H\mathbf{h}_s\mathbf{w}_{t+1}^k+\mathbf{z}_{t+1}^k & \text{if } t+1\in\mathcal{I}_E \text{ and } k\in[K],\\ \mathbf{w}_{t+1}^k & \text{if } t+1\in\mathcal{I}_E \text{ and } k\notin[K].\end{cases}$$

Based on the new definition of $\mathbf{p}_{t+1}^k$, we next need to establish Lemma 5, similar as Lemma 4.

Lemma 5. *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* $\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1}-\overline{\mathbf{w}}_{t+1}\right\|^2\leq\left(d+\frac{dK}{M}\right)\frac{\eta_t^2}{1-\mu\eta_t}$.

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

$$\mathbb{E}\left[\overline{\mathbf{p}}_{t+1}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\mathbf{p}_{t+1}^{k}\right] = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_{t+1}^{k} + \mathbb{E}\left[\frac{1}{N}\sum_{k\in[K]}\mathbf{z}_{t+1}^{k}\right] = \overline{\mathbf{w}}_{t+1}.$$

We next evaluate the value of $\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right\|^2$,

$$\mathbb{E}\left\|\overline{\mathbf{p}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right\|^2 = \mathbb{E}\left\|\frac{1}{N}\left[\sum_{k\in[k]}\mathbf{h}_k^H\mathbf{h}_s\mathbf{w}_{t+1}^k + \sum_{k\notin[k]}\mathbf{w}_{t+1}^k\right] + \frac{1}{N}\sum_{k\in[K]}\mathbf{z}_{t+1}^k - \overline{\mathbf{w}}_{t+1}\right\|^2$$

$$= \frac{1}{N^2}\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_s\mathbf{w}_{t+1}^k + \sum_{k\in[K]}\mathbf{z}_{t+1}^k - \sum_{k\in[K]}\mathbf{w}_{t+1}^k\right\|^2$$

$$= \frac{1}{N^2}\left[\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_s\mathbf{w}_{t+1}^k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\mathbf{z}_{t+1}^k\right\|^2 + \mathbb{E}\left\|\sum_{k\in[K]}\mathbf{w}_{t+1}^k\right\|^2 - 2\mathbb{E}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_s\mathbf{w}_{t+1}^k\sum_{k\in[K]}\mathbf{w}_{t+1}^k\right]\right]$$

$$= \frac{1}{N^2}\mathbb{V}\mathrm{ar}\left[\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_s\mathbf{w}_{t+1}^k\right] + \frac{1}{N^2}\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{z}_{t+1}^k\right\|^2$$

$$= \frac{K}{N^2M}\sum_{k\in[K]}\left\|\mathbf{w}_{t+1}^k\right\|^2 + \frac{dK}{N^2}\frac{1}{\mathsf{SNR}_{\mathsf{UL}}} \le \frac{K}{N^2M}\sum_{k\in[K]}\left\|\mathbf{w}_{t+1}^k\right\|^2 + \frac{dK}{N^2}\frac{\eta_t^2}{1-\mu\eta_t}$$

□

Building on Lemma 5, we next establish Theorem 3 for the convergence of random orthogonalization when applies to both uplink and downlink communications.

Theorem 3 (***Convergence for random orthogonalization in both uplink and downlink***)**.** *Consider a wireless FL task that applies random orthogonalization for both uplink and downlink communication design. With Assumptions 1-4, for some $\gamma \ge 0$, if we select the learning rate as $\eta_t = \frac{2}{\mu(t+\gamma)}$ and downlink SNR scales $\mathsf{SNR}_{\mathsf{DL}} \ge \frac{1-\mu\eta_t}{\eta_t^2}$ at learning round t, we have*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \le \frac{L}{2(t+\gamma)}\left[\frac{4B}{\mu^2} + (1+\gamma)\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2\right] + C, \tag{29}$$

for any $t \ge 1$, where

$$B \triangleq \sum_{k=1}^{N}\frac{H_k^2}{N^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{N-K}{N-1}\frac{4}{K}E^2H^2 + \frac{4}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}_{\mathsf{UL}}}\right)E^2H^2 + \frac{dK}{N^2}. \tag{30}$$

and

$$C \triangleq \frac{K}{N^2M}\sum_{k\in[K]}\left\|\mathbf{w}_{t+1}^k\right\|^2.$$

Proof. The proof of Theorem 3 basically follows the proof of Theorem 1. The only difference is we use Lemma 5 instead of Leamma 4 in (24), so that we can obtain the upper bound in 3. □

We note that due to the existence of interference, the convergence suffers from a constant gap C from optimum. However, the constant decreases linearly as the increase of M.

References

[1] X. Wei, C. Shen, J. Yang, and H. V. Poor, "Random orthogonalization for federated learning in massive MIMO systems," in *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] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.

[5] 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.

[6] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimal power control for over-the-air computation," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

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

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

[9] 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.

[10] 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.

[11] 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)*. IEEE, 2020, pp. 1221–1230.

[12] 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.

[13] ——, "Gradient estimation for federated learning over massive MIMO communication systems," *arXiv preprint arXiv:2003.08059*, 2020.

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

[15] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," in *ICML Workshop on Coding Theory for Machine Learning*, 2019.

[16] 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 (ICLR)*, 2020.

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

[18] S. Hu, J. Goetz, K. Malik, H. Zhan, Z. Liu, and Y. Liu, "Fedsynth: Gradient compression via synthetic data in federated learning," *arXiv preprint arXiv:2204.01273*, 2022.

[19] H. Zhu and Q. Ling, "Byzantine-robust aggregation with gradient difference compression and stochastic variance reduction for federated learning," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4278–4282.

[20] 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," *arXiv preprint arXiv:2001.05713*, 2020.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.*, 2022.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Over-the-air federated learning with energy harvesting devices," *arXiv preprint arXiv:2205.12869*, 2022.

[38] 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.

[39] 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.

[40] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Time-correlated sparsification for efficient over-the-air model aggregation in wireless federated learning," *arXiv preprint arXiv:2202.08420*, 2022.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

[49] S. Xia, J. Zhu, Y. Yang, Y. Zhou, Y. Shi, and W. Chen, "Fast convergence algorithm for analog federated learning," *arXiv preprint arXiv:2011.06658*, 2020.

[50] 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.

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

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

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

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

[55] 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.

[56] 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.

[57] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Trans. Cogn. Commun. Netw.*, vol. Early access, pp. 1–1, 2022.

[58] 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.

[59] L. Deng, “The MNIST database of handwritten digit images for machine learning research,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[60] 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.