ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9838988

Random Orthogonalization for Federated Learning in Massive MIMO Systems

Xizixiang Wei*, Cong Shen*, Jing Yang†, H. Vincent Poor‡
* Department of Electrical and Computer Engineering, University of Virginia, USA
† Department of Electrical Engineering, The Pennsylvania State University, USA
‡ Department of Electrical and Computer Engineering, Princeton University, USA

***Abstract*—We propose a novel uplink communication method, 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 model aggregation 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, while significantly reducing the channel estimation overhead at the receiver. 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

Communication overhead is widely considered one of the primary bottlenecks for federated learning (FL) [1], [2], as a FL task consists of multiple learning rounds, each of which requires uplink and downlink model exchange between clients and the server. Compared with downlink broadcasting, uplink communication is more challenging in FL. Due to the strigent power constraint at edge devices, channel noise and fading have more conspicuous impacts on uplink communications. More importantly, the limited uplink communication resources may severely limit the *scalability* of FL, negatively affecting one of its primary features [3].

To tackle the scalability problem in FL uplink communications, several over-the-air computation (also known as *AirComp*) mechanisms have been exploited in wireless FL (see [4] and the references therein). Instead 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 model (global model) directly at the FL server. Zhu et al. [5] propose an analog aggregation framework which "inverts" the fading channel at each transmitter, so that the sum model can be directly obtained at the server. However, the fundamental limitation of analog aggregation is that it requires channel state information at transmitter (CSIT). The process of enabling CSIT is complicated and the precision of CSIT is often worse than the channel state information at receiver (CSIR). Besides, analog aggregation essentially requires a channel inversion power control, which is well known to "blow up" when channel is in deep fade. Moreover, analog aggregation does not naturally extend to multiple-input and multiple-output (MIMO) systems where the uplink channels become vectors, which makes channel inversions at the transmitters nontrivial.

This paper aims at designing a simple-yet-effective uplink FL communication and model aggregation method. To address the scalability challenge in FL, we explore another design degree of freedom (d.o.f.) in modern wireless systems: *massive MIMO*. The proposed framework only requires the BS to estimate a summation channel, which significantly alleviates the burden on uplink channel estimation in FL. Moreover, this approach is agnostic to the number of clients, making it attractive for the scalability of FL. By tightly integrating the channel hardening and favorable propagation properties of massive MIMO, the proposed scheme, 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. To analyze the performances of random orthogonalization, we derive the Cramer-Rao lower bounds (CRLBs) of the average model estimation as a theoretical benchmark. Moreover, taking both interference and noise into consideration, a novel convergence bound of FL is derived for the proposed method over massive MIMO channels. Notably, we establish an explicit relationship among the convergence rate, the number of clients K, and the number of antennas M, which provides practical design guidance for wireless FL. Numerical results validate the effectiveness and efficiency of the proposed method.

The potential of MIMO for wireless FL has attracted interest recently. MIMO beamforming design to optimize FL has been studied in [6], [7]. Coding, quantization, and compressive sensing over a (massive) MIMO channel for FL has been studied in [8]–[10]. Nevertheless, none of these works tightly incorporates the unique properties of massive MIMO in the FL uplink communication design. On the other hand, massive MIMO can also be utilized in a straightforward manner, e.g., one can use traditional 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,

XW and CS were partially supported by the US National Science Foundation (NSF) under ECCS-2033671. JY was supported in part by the NSF under CNS-1956276 and CNS-2114542. HVP was supported in part by the NSF under CCF-1908308.

this heuristic approach requires large channel estimation overhead, especially in massive MIMO. Decoding individual local models also makes it easier for the server to sketch the data distribution of a client. Moreover, matrix inversion operations in ZF or MMSE detectors are computationally demanding, which increases the complexity and latency.

II. System Model

A. FL Model

Consider a FL task with a central server and K clients. Each client $k \in [K]$ stores a (disjoint) local dataset $\mathcal{D}_k$, with its size denoted by D_k. The size of the total data is $D \triangleq \sum_{k\in[K]} D_k$. We use $f_k(\mathbf{w})$ to denote the local loss function at client k, which measures how well a machine learning (ML) model with parameter $\mathbf{w} \in \mathbb{R}^d$ fits its local dataset. The global objective function over all K clients is $f(\mathbf{w}) = \sum_{k\in[K]} p_k f_k(\mathbf{w})$, 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: $\mathbf{w}^* \triangleq \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$. A typical wireless FL pipeline is illustrated in Fig. 1. Specifically, this pipeline 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 all devices over the downlink wireless channel.
2) **Local computation.** Each client uses its local data to train a local model improved upon the received global model $\mathbf{w}_t$. We assume that mini-batch stochastic gradient descent (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 $\mathbf{w}$.
3) **Uplink communication.** Each 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: $\mathbf{w}_{t+1} = \Sigma_{k\in[K]} p_k \tilde{\mathbf{w}}_{t+1}^k$. For simplicity, we assume that each local dataset has equal size, hence $p_k = \frac{1}{K}$.

This work focuses on steps 3 and 4 in the FL pipeline. In particular, we take advantage of the unique properties of massive MIMO to design efficient FL uplink communication and server aggregation.

B. Communication Model

Consider a massive MIMO system equipped with M antennas at the BS (server) where K single-antenna devices (clients) are involved in the aforementioned FL task. At the uplink step of the t-th round, each client transmits the differential between the received global model and the computed new local model $\mathrm{x}_t^k = \mathbf{w}_t - \mathbf{w}_{t+1}^k \in \mathbb{R}^d, \forall k \in [K]$ to the BS[1], 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,

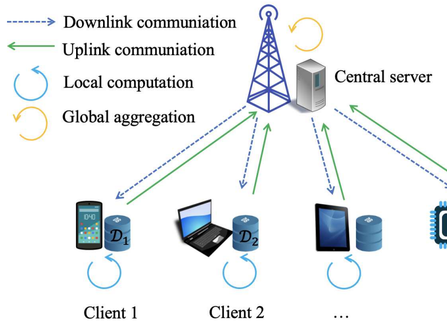

Fig. 1. The wireless FL pipeline.

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

$$\mathbf{y}_i = \sqrt{P} \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \mathbf{n}_i, \ \ \forall i = 1, \cdots, d, \tag{1}$$

where P is the maximum transmit power of each client, $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ is the wireless channel between k-th client and BS, and $\mathbf{n}_i \in \mathbb{C}^{M\times 1}$ is the uplink noise. We assume normalized symbol power $\mathbb{E}\left\|x_{k,i}\right\|^2 = 1$, normalized Rayleigh block fading channel[3] $\mathbf{h}_k \sim \mathcal{CN}(0, \frac{1}{M}\mathbf{I})$ in d slots, and independent and identically distributed (i.i.d.) Gaussian noise $\mathbf{n}_i \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$. We define the signal-to-noise ratio (SNR) as $\mathsf{SNR} \triangleq P/\sigma^2$, and w.l.o.g. we set $P = 1$. 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[4] can be written as

$$\mathbf{y}_i = \mathbf{H}\mathbf{x}_i + \mathbf{n}_i. \tag{2}$$

Eqn. (2) is a standard MIMO 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 $\{x_{k,i}\}_{i=1}^d$ individually and obtaining the aggregated parameter $\tilde{x}_i \triangleq \sum_{k\in[K]} \hat{x}_{k,i}$ by a summation is inefficient. We propose a novel method that allows the BS to compute $\tilde{x}_i$ directly. 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

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

III. Random Orthogonalization

We study a wireless FL framework where the global model can be directly obtained at the BS via a simple operation. By exploring favorable propagation and channel hardening in massive MIMO, our proposed FL framework obtains the global model by the following three main steps.

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

[2] In general, differential model parameters can be transmitted over any d shared time-frequency resources. For simplicity, we use d time slots here.

[3] Large-scale pathloss and shadowing effect is assumed to be taken care of by, e.g., open loop power control [11].

[4] 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.

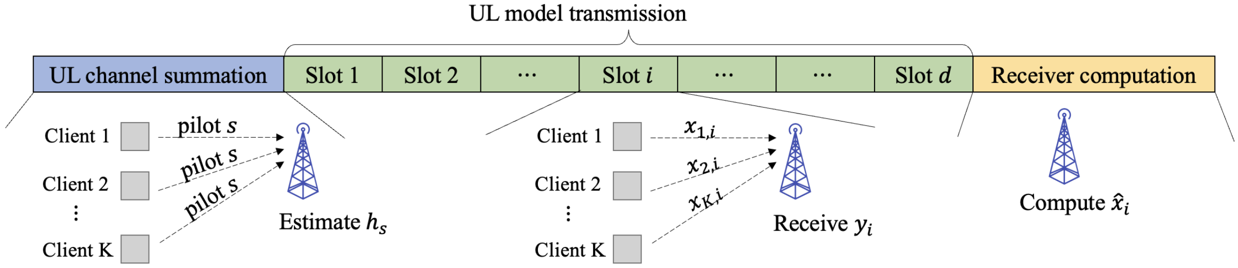

Fig. 2. An illustration of the proposed uplink FL design with massive MIMO.

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

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

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). For simplicity, we assume perfect sum channel estimation at the BS.

(2) Uplink model transmission. All clients transmit model differential parameters $\{x_{k,i}\}_{i=1}^{d}$ to the BS in d time slots. The received signal for each differential model element is

$$\mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \mathbf{n}_i, \quad \forall i = 1, \cdots, d.$$

(3) Receiver computation. The BS estimates each aggregated parameter via a simple *linear projection* operation:

$$\begin{aligned}\tilde{x}_i &= \mathbf{h}_s^H \mathbf{y}_i = \sum_{k\in[K]} \mathbf{h}_k^H \sum_{k\in[K]} \mathbf{h}_k x_{k,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i \\ &\overset{(a)}{=} \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i}}_{\text{Signal}} + \underbrace{\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i}}_{\text{Interference}} + \underbrace{\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i}_{\text{noise}} \\ &\overset{(b)}{\approx} \sum_{k\in[K]} x_{k,i}, \quad \forall i = 1, \cdots, d. \end{aligned} \tag{4}$$

The above three-step procedure is illustrated in Fig. 2. Based on Eqn. (4), BS then computes the global model via Eqn. (3) and begins the next communication round. As shown in part (a) of Eqn. (4), inner product $\mathbf{h}_s^H \mathbf{y}_i$ can be regarded as the combination of three parts: signal, interference, and noise. We next show that, taking advantage of two fundamental properties of massive MIMO, the approximation (b) in Eqn. (4) is asymptotically error-free, as the number of antennas at the BS 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 [12]:

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

In practical systems, when M is large but finite, we have

$$\mathbb{E}_{\mathbf{h}} \left[\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i} \right] = \sum_{k\in[K]} x_{k,i}, \tag{5}$$

and

$$\mathbb{V}\text{ar}_{\mathbf{h}} \left[\sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i} \right] = \frac{\sum_{k\in[K]} x_{k,i}^2}{M} \tag{6}$$

for the signal part of (4).

Favorable propagation. Since channels between different users are independent random vectors, massive MIMO also offers favorable propagation [12]:

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

Similarly, when M is finite, we have

$$\mathbb{E}_{\mathbf{h}} \left[\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} \right] = 0, \tag{7}$$

and

$$\mathbb{V}\text{ar}_{\mathbf{h}} \left[\sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} \right] = \frac{(K-1)\sum_{k\in[K]} x_{k,i}^2}{M}. \tag{8}$$

Furthermore, the expectation of the noise part in (4) is zero. Therefore, $\tilde{x}_i$ in Eqn. (4) 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 FL.

To gain some insight of random orthogonality, we approximate the average signal-to-interference-plus-noise-ratio (SINR) after the operation in Eqn. (4) as

$$\begin{aligned}&\mathbb{E}[\text{SINR}_i] \approx \\ &\frac{\mathbb{E}_{\mathbf{h},x} \left\| \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{h}_k x_{k,i} \right\|^2}{\mathbb{E}_{\mathbf{h},\mathbf{n},x} \left\| \sum_{k\in[K]} \sum_{j\in[K], j\neq k} \mathbf{h}_k^H \mathbf{h}_j x_{j,i} + \sum_{k\in[K]} \mathbf{h}_k^H \mathbf{n}_i \right\|^2} \\ &= \frac{M}{K-1+1/\mathsf{SNR}},\end{aligned} \tag{9}$$

which grows linearly with M for a fixed K. We note that Eqn. (9) 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 IV-B.

Remark 1. *Unlike the analog aggregation method in [5], random orthogonalization does not require any CSIT, and 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 [6] and traditional MIMO decoders. Moreover, the global model is obtained after a single linear projection, which improves the privacy and reduces the system latency.*

Remark 2. *The proposed framework assumes a perfect estimation of $\mathbf{h}_s$ and requires channel hardening and favorable propagation. In practical systems, to improve the accuracy of the estimate $\hat{\mathbf{h}}_s$, BS can adopt multiple pilots for channel estimation. We will provide more details on the robustness of the proposed scheme over imperfect $\hat{\mathbf{h}}_s$ and evaluate the circumstances where channel hardening and favorable propagation are not fully offered, e.g. correlated channels, in the journal version.*

IV. Performance Analysis

In this section, we analyze the performances of random orthogonalization in FL. We first derive CRLBs of the estimates of global model parameters as the theoretical benchmark of the proposed method. Then, by an ML model convergence analysis, we investigate the relationship between the number of involved clients K and the number of BS antennas M. We show that random orthogonalization has the potential to achieve nearly the same convergence rate as the interference-free case in massive MIMO systems.

A. Cramer-Rao Lower Bounds

Recall that the received signal is $\mathbf{y}_i = \mathbf{H}\mathbf{x}_i + \mathbf{n}_i$. Denoting $\boldsymbol{\mu} = \mathbf{H}\mathbf{x}_i$, we have that $\mathbf{y}_i \sim \mathcal{CN}(\boldsymbol{\mu}, \frac{1}{\mathsf{SNR}}\mathbf{I})$. The Fisher information matrix (FIM) of the estimation of $\mathbf{x}_i$ is

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

After inserting $\frac{\partial \boldsymbol{\mu}(\mathbf{x}_i)}{\partial \mathbf{x}_i} = \mathbf{H}$ into FIM, we have $\mathbf{F} = 2 \cdot \mathsf{SNR} \cdot \mathrm{Re}(\mathbf{H}^H\mathbf{H})$. The CRLBs are given by the inverse of the FIM $\mathbf{C}_{\hat{\mathbf{x}}_i} = \mathbf{F}^{-1}$. CRLB expresses a lower bound on the variance of unbiased estimators, stating that the variance of any such estimator is at least as high as the inverse of the FIM. Eqn. (4) has shown that the proposed method leads to an unbiased estimation of the global model; 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) $\mathbb{E}\left\|\mathbf{x}_i - \hat{\mathbf{x}}_i\right\|^2$ to evaluate the performance of global model estimation.

B. Convergence analysis

To simplify the analysis, we assume[5] $E = 1$, which is also referred to as *parallel SGD* [13], and make the following standard assumptions that are commonly adopted in the convergence analysis of FedAvg and its variants; see [13]–[16].

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

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

Assumption 3. ***Unbiased SGD:*** *$\forall k \in [K]$, $\mathbb{E}[\nabla\tilde{f}_k(\mathbf{w})] = \nabla f_k(\mathbf{w})$;*

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

Lemma 1 (***One-step convergence***)**.** *Based on Assumptions 1-4 and selecting learning rate $\eta_t \leq 1/(2\mu)$, $\forall t \in [T]$, the following inequality holds for parallel SGD:*

$$\begin{aligned}\mathbb{E}\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1 - 2\mu\eta_t)\mathbb{E}\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2 \\ + \eta_t^2\left[1 + \frac{K}{M} + \frac{1}{\mathsf{SNR}}\right]\frac{H^2}{K}.\end{aligned} \tag{10}$$

Proof. We introduce an auxiliary error-free global model $\bar{\mathbf{w}}_{t+1} = \frac{1}{K}\mathbf{w}_{t+1}^k$. We first have

$$\begin{aligned}&\mathbb{E}\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2 = \mathbb{E}\left\|\mathbf{w}_{t+1} - \bar{\mathbf{w}}_{t+1} + \bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2 \\ &= \mathbb{E}\underbrace{\left\|\mathbf{w}_{t+1} - \bar{\mathbf{w}}_{t+1}\right\|^2}_{A_1} + \mathbb{E}\underbrace{\left\|\bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|^2}_{A_2} \\ &+ 2\mathbb{E}\underbrace{\left\langle \mathbf{w}_{t+1} - \bar{\mathbf{w}}_{t+1}, \bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\rangle}_{A_3}.\end{aligned} \tag{11}$$

Note that $\mathbb{E}[A_3] = 0$. Then, $\mathbb{E}[A_2]$ can be obtained from a well-known result [14]:

$$\mathbb{E}\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2 \leq (1 - 2\mu\eta_t)\mathbb{E}\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2 + \eta_t^2\frac{H^2}{K}. \tag{12}$$

We finally focus on $\mathbb{E}[A_1]$. Based on (6) and (8), we have

$$\begin{aligned}&\mathbb{E}\left\|\mathbf{w}_{t+1} - \bar{\mathbf{w}}_{t+1}\right\|^2 = \mathbb{E}\left\|\frac{1}{K}\sum_{k\in[K]}\mathbf{x}_k - \frac{1}{K}\sum_{k\in[K]}\hat{\mathbf{x}}_k\right\|^2 \\ &= \frac{1}{K^2}\mathbb{E}\left\|\sum_{k\in[K]}\mathbf{h}_k^H\mathbf{h}_k\mathbf{x}_k + \sum_{k\in[K]}\sum_{j\in[K],j\neq k}\mathbf{h}_k^H\mathbf{h}_j\mathbf{x}_j\right. \\ &\left. + \mathbf{N}_{t+1}\sum_{k\in[K]}\mathbf{h}_k - \sum_{k\in[K]}\mathbf{x}_k\right\|^2 \\ &= \eta_t^2\frac{\sum_{k\in[K]}\mathbb{E}\left\|\nabla\tilde{f}_k(\mathbf{w})\right\|^2}{K^2}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}}\right) \\ &\leq \eta_t^2\frac{H^2}{K}\left(\frac{K}{M} + \frac{1}{\mathsf{SNR}}\right),\end{aligned} \tag{13}$$

where $\mathbf{N}_{t+1} \in \mathbb{C}^{d\times M}$ is the stack of noise $\mathbf{n}_i^H$ $\forall i = 1, \cdots, d$. Plugging (12) and (13) back to (11) completes the proof. □

[5] We will address the general case of $E > 1$ in the journal version.

Building on Lemma 1, we next present a complete convergence upper bound for random orthogonalization. Due to space limitation, the proof of Theorem 1 is omitted and will be reported in the journal version.

Theorem 1 (***Convergence for random orthogonalization***)**.** *With Assumptions 1-4, for some $\gamma \geq 0$, if we select the learning rate as $\eta_t = \frac{2}{\mu(t+\gamma)}$, we have*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)} \left[\frac{4B}{\mu^2} + (1+\gamma) \left\| \mathbf{w}_0 - \mathbf{w}^* \right\|^2 \right], \tag{14}$$

for any $t \geq 1$, where

$$B \triangleq \left[1 + \frac{K}{M} + \frac{1}{\mathsf{SNR}} \right] \frac{H^2}{K}. \tag{15}$$

Lemma 1 and Theorem 1 illustrate that there are two main factors that influence the convergence rate of FL *in the high SNR regime*: **variance reduction** and **channel interference**. In particular, the definition of B in (15), which appears in both Lemma 1 and Theorem 1, 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 [17], which is captured by the $\frac{H^2}{K}$ term in (10) and (14). However, due to the existence of interference, the convergence rate is determined by both variance reduction and channel interference, shown as $\frac{H^2}{K}$ and $\frac{(K/M+1/\mathsf{SNR})H^2}{K}$ terms in (15). This suggests that the desired variance reduction may be adversely impacted if channel interference dominates the convergence bound. In particular, when $M >> K$, we have $\frac{1}{K} >> \frac{K/M+1/\mathsf{SNR}}{K}$ when SNR is high, 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 bring the advantage of variance reduction.

Remark 3. *In massive MIMO, a BS is usually equipped with hundreds of antennas. Although there may exist large number of users participating in FL, only a small number of them are simultaneously active [6], especially in millimeter wave cells whose coverage are usually small. 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-free convergence rate with low communication and computation overhead in massive MIMO systems.*

V. Experiments

We evaluate the performances of random orthogonalization for uplink FL communications through numerical experiments. From a communication performance perspective, we compare the proposed method with the traditional MIMO detector to compute the global model. Then, we use a real-world FL task to evaluate the learning performance of the proposed method.

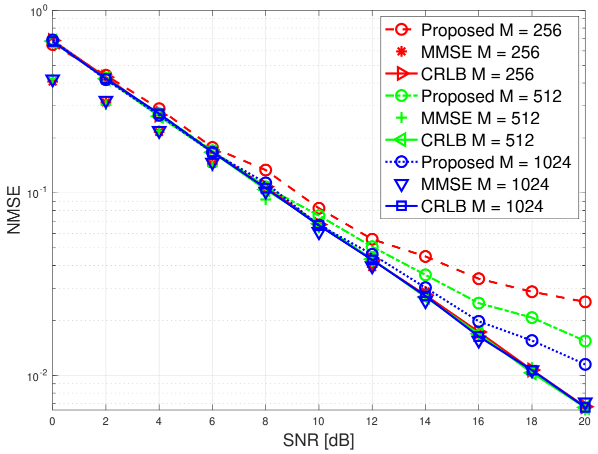

Fig. 3. NMSE of the global ML model parameters versus SNR.

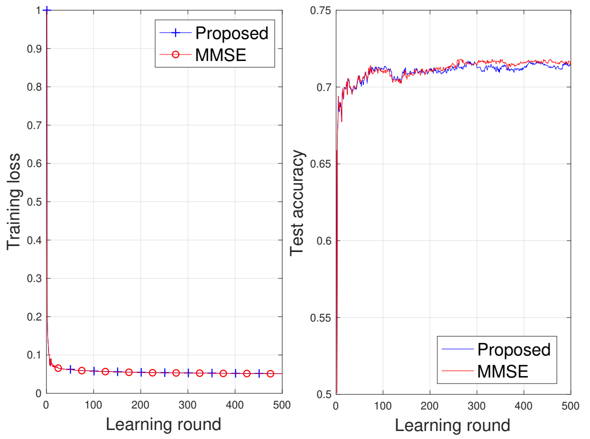

Fig. 4. Training loss and test accuracy of a SVM FL task of random orthogonalization and MMSE.

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 normalized mean square error (NMSE) of the computed global model to evaluate the system performance. All NMSE results in the simulation are obtained from $2,000$ Monte Carlo experiments. Fig. 3 compares the NMSE performance of the proposed scheme with a MMSE decoder as well as CRLB under different system SNRs. As illustrated in Fig. 3, the proposed method performs nearly identically to the MMSE decoder in low and moderate SNRs under different antenna configurations (see SNR $\leq$ 12 dB). As the SNR increases, the dominant factor affecting system performances becomes the interference among different users. Unlike the MMSE decoder that can cancel all interferences when $K \leq M$ at high SNR, Eqn. (9) shows that, for a given K and M, the proposed framework has a fixed (approximate) SIR $= \frac{K-1}{M}$ as SNR $\rightarrow \infty$, which explains why the performance of the proposed scheme

TABLE I
COMPUTATION TIME COMPARISON BETWEEN RANDOM ORTHOGONALIZATION AND MMSE DECODER

# antennas	Total CPU time (second)		
(M)	Proposed	MMSE	Proposed/MMSE
256	0.0186	2.7141	0.68%
512	0.0303	12.4155	0.24%
1024	0.0448	82.3530	0.05%

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. 3 that the performance gap between the proposed method and MMSE reduces, from about 5 dB when $M = 256$ to about 2 dB when $M = 1024$ at SNR = 20 dB. Note that our method only requires $1/K$ of channel estimation overheard compared with MMSE, and this advantage is more significant when the BS is equipped with larger number of antennas.

We next focus on the low-latency benefit of random orthogonalization. Table I compares the computation time of the proposed scheme and MMSE decoder with SNR = 10 dB. 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 is less than 1% of the MMSE baseline. Especially, when $M = 1024$, despite the 0.3 dB NMSE performance loss compared with the MMSE decoder (as shown in Fig. 3), the computation time of the proposed method is only 0.05% of the MMSE baseline. The results suggest that the random orthogonalization framework is attractive in massive MIMO systems, because it has nearly identical NMSE performance to CRLB but requires much less channel estimation overhead and achieves extremely lower system latency than the MMSE decoder.

B. Learning performance

We use a classification task to evaluate the ML model accuracy and convergence rate of the proposed random orthogonalization approach. In particular, we implement a support vector machine (SVM) to classify even and odd numbers in the MNIST handwritten-digit dataset [18], with $d = 784$. We consider a BS with $M = 256$ antennas and $K = 8$ active clients involved in this task. The size of the local training set at each client is 500, the size of the test set is $2,000$, and we set $E = 1$. Fig. 4 reports the training loss and test accuracy of the proposed method and MMSE decoder with SNR = 10 dB. Although the MSE of the global model at the BS during the learning process is about 2 dB worse for random orthogonalization as shown in Fig. 3, the actual learning performances of the two methods are nearly identical, further validating the effectiveness of random orthogonalization.

VI. CONCLUSION

Leveraging the unique characteristics of channel hardening and favorable propagation in massive MIMO, we have proposed a novel uplink communication and processing 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. 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 asymptotically achieves the same convergence rate as vanilla FL with perfect communications, and were further validated with numerical experiments. More importantly, random orthogonalization improves the scalability of FL, which is a critical feature that is often bottlenecked by the limited wireless resources.

REFERENCES

[1] 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.

[2] 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.

[3] P. Kairouz *et al.*, “Advances and open problems in federated learning,” *arXiv preprint arXiv:1912.04977*, 2019.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.