Federated Learning in Massive MIMO 6G Networks: Convergence Analysis and Communication-Efficient Design

Federated Learning in Massive MIMO 6G Networks: Convergence Analysis and Communication-Efficient Design

Yuchen Mu, Navneet Garg, *Member, IEEE,* and Tharmalingam Ratnarajah, *Senior Member, IEEE*.

***Abstract*—In Federated learning (FL), model weights must be updated at local users and the base station (BS). These weights are subjected to uplink (UL) and downlink (DL) transmission errors due to the limited reliability of wireless channels. In this paper, we investigate the impact of imperfections in both UL and DL links. First, for a multi-user massive multi-input-multi-output (mMIMO) 6G network, employing zero-forcing (ZF) and minimum mean-squared-error (MMSE) schemes, we analyze the estimation errors of weights for each round. A tighter convergence bound for the communication-efficient FL algorithm is derived with $\mathcal{O}\left(T^{-1}\sigma_z^{-2}\right)$ behavior, where σ_z denotes the variance of overall communication error including the quantization noise. The analysis shows that the reliability of DL links is more critical than that of UL links; and the transmit power can be varied in training process to reduce energy consumption. We also vary the number of local training steps, average codeword length after quantization and scheduling policy to improve the communication efficiency. Simulations with image classification problems on MNIST, EMNIST and FMNIST datasets verify the derived bound and are useful to infer the minimum SNR required for successful convergence of the FL algorithm.**

***Index Terms*—Deep learning, federated learning, massive MIMO (mMIMO), 6G networks.**

I. Introduction

Rapid evolutions in the field of machine and deep learning during past few years [1] along with tremendous applications including computer vision [2], natural language processing [3] and medical images analysis [4] have introduced a great impact on our daily lives. For the proper working of a machine (deep) learning based application, deep learning requires a powerful machine with huge computational resource, and a significant amount of data in order to learn reliable models, which not only incurs the higher costs, but also damages the privacy while sharing users' data. Recently, federated learning (FL) [5] models have been studied, where models are learned in a distributed manner with limited cooperation without sharing data. Thus, compared with traditional "one node learning", where a central server/device collects all the data needed and performs training, federated learning has the benefits of utilizing local users' computing resources, such as graphics processing unit (GPU) on a local user's PC or central processing unit (CPU) on a smart phone. Since FL is a privacy-preserving algorithm [6]–[9], each user's dataset never leaves its local device. This is especially helpful in privacy critical scenarios, where local users' data contains sensitive or personal information that they may not be willing to share, e.g., medical images, different users' keyboard typed letters, browser and searching history, etc. Another advantage of FL algorithm in practical communication setting is the requirement of significantly fewer physical resources (transmit power and bandwidth), which are limited in a typical wireless system.

Regarding FL methods, averaging based FL algorithm (FEDAVG) was first proposed in [5] and was mathematically analyzed in [5] and [10], where local users run several local stochastic gradient descents in parallel on their local training sets, respectively and send the trained model weights to the central server. The server aggregates the received weights by averaging or combining linearly, followed by broadcasting of the updated global model weights back to users for further training. Such iterations continue until the weights converge.

A. Communication-efficient FL frameworks

Further, recently, trend in distributed learning is pushing the deployment of training and computing towards the network edge. In this way, the process of data acquisition and computation are decoupled, which is of interest to both academia and industry. Motivated by this, emergence of FL also pushes to integrate two originally decoupled areas: machine learning and wireless communications. The advanced signal processing and communication techniques can be utilized and applied to speed up the convergence of FL model and help to train a global model with better performance accounting the reliability of wireless links.

With rapid deployment of fifth generation (5G) communication infrastructures and networks on a worldwide scale, both industry and academia have begun to transfer their research focus to beyond 5G and sixth generation (6G) communications. Compared with 5G, in 6G communications, penetration of Artificial Intelligence (AI) into every aspect of heterogeneous and massive-scale networks is required; that is, achieving ubiquitous AI in 6G [11]. With its privacy-preserving nature, FL has been recognized as one of the most promising solutions to resolve and fulfill the challenges in 6G communication networks [12]. In this way, works like [13]–[18] investigate FL framework under a 5G or 6G scenario. Among these, the data privacy leakage issues are given in [13] and [14], while authors in [15]–[18] provides the key technical challenges that

Y. Mu, N. Garg, and T. Ratnarajah are with Institute for Digital Communications, The University of Edinburgh, Edinburgh, EH9 3FG, UK (e-mails: {yuchen.mu, ngarg, t.ratnarajah}@ed.ac.uk).

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/T021063/1.

FL may face in 6G scenarios. Meanwhile, these works have also demonstrated that the advantages provided by 6G, such as high performance networking, hyper-flexible architecture and AI-enhanced security [18], are also vital for accomplishing the requirements of FL frameworks in future communication networks; that is, 6G communication and FL can benefit each other mutually.

In the direction to integrate FL algorithms with communication-efficient framework under different wireless network settings, a general discussion can be seen in [19]. These design investigations can be categorized as follows.

1) Quantization schemes: This includes the effective quantization scheme for model weights in order to optimize the number of symbols for transmissions. The work in [20] considers a layered quantization scheme for downlink transmission, where various quantization gains are used to quantize the weights of different layers inside a deep learning model. Similarly, lattice quantization [21], sparsification [22], ternary gradients [23] and sparse ternary compression [24] are employed. These quantization schemes show their effectiveness in making the FL framework efficient. Among those, sparse ternary compression [24] outperforms others since it not only compresses both uplink and downlink streams with high compression rate, but also stays robust to the scenario where non-i.i.d. local users data is considered.

2) Resource allocation: Optimizing the communication resources, such as transmit energy and local computation energy consumption, is investigated in [25]–[30]. In [25], the loss function is minimized with the constraints on packet loss as well as bandwidth of wireless medium, and optimal transmit power for each user is derived. The work in [26] decomposes the non-convex energy optimization problem into several subproblems and derives corresponding closed-form solutions. The convergence analysis in [26] was later conducted by authors in [27]. In [28], a local computation latency constraint was considered in the optimization, and it was shown that training completion time in [28] was less than that in [26]. Two communication protocols namely, non-orthogonal multiple access (NOMA) and time division multiple access (TDMA) were considered in [29] to optimize the transmission rate and CPU frequency of local user devices. In [30], a resource-constrained FL framework for heterogeneous data is proposed and corresponding convergence analysis is also provided. Another resource allocation problem is to design the scheduling policies to allow a subset of the users to update the model such that the overall cost of resources can be reduced. In [31], an age-based scheduling policy has been proposed, which accounts for the recency of updates and the instantaneous channel quality of users. Proportional fair based scheduling policy is considered in [32], where users are selected based on signal-to-noise ratio (SNR). In [33], A greedy algorithm for scheduling the devices is proposed by analyzing the trade-off between local computation, communication latencies and whole FL system learning time. In [34], scheduling problem in FL is investigated under a unmanned aerial vehicles (UAVs) communication scenario. A joint optimization problem, which includes device selection, UAVs placement, and resource management, is formulated and solved to speed up the convergence. Similarly, [35] investigates user scheduling problem in a UAV-assisted FL framework. In this work, based on the multi-dimensional contract incentive design, a subset of the users will be selected by UAV in each global communication round. In [36], differences between transmitting model weights and the weights updates are compared analytically and in simulations.

3) Imperfect links: Many of the existing literature [37]–[40] assume only UL links to be imperfect and error-prone, while considering DL links to be error-free; that is, the fact that the aggregated model being broadcasted in DL links also experiences fading and interference, has not been taken into consideration in the existing works. In this work, we focus on the massive MIMO scenario, where both the UL and DL links are error-prone, which is of interest, since our results show that the reliability of DL-links have higher impact on the performance of FL algorithm.

B. Contributions

In this paper, we consider a massive MIMO 6G network consisting of a base station (BS) and multiple single antenna users. In this system, we integrate FL framework, where the parameters of deep learning models are updated at both BS (or server) side and local user side. In FL, at regular intervals, weights are transmitted via UL channels, updated by aggregation and transmitted back via DL channels. Fading and interference caused by the communication links will degrade the training performance compared with baseline, where error-free perfect links are considered. The main contributions of this paper can be summarized as follows.

1) *FL with massive MIMO 6G network*: We design and implement a FL framework with massive MIMO 6G communication network, considering imperfections in both UL and DL links. Employing zero-forcing (ZF) and minimum mean-squared-error (MMSE) precoding/combining methods, we analyze the communication performance of this system and compare it with two simulation baselines in existing literature.
2) *Simpler convergence proof and tighter bound*: For the system considered, we derive the upper bound on the convergence rate, which is tighter than that of FEDAVG. The bound also provides interesting insights about UL and DL communications. Our analysis shows that SNR in UL-DL communications can be kept lower at the lower iterations. This leads to improved power consumption and is also analyzed. It can also be concluded that reliability of DL links is more important than that of UL links. These analytical conclusions are verified via simulations.
3) *Communication-efficient system design*: Motivated by the convergence analysis, several variants of proposed FL framework with different degrees of communication-efficiency are designed and implemented where we adjust local training steps, transmit power, scheduling policy and average codeword length after quantization. These modifications could be leveraged to reduce the total communication cost incurred during training process.

Work	Convergence Analysis	Non-i.i.d. Dataset	Design Focus					Uplink			Downlink			Scheduling				Datasets			
			TP	TS	P/C	QS	SP	RSN	CF	IU	RSN	CF	IU	FP/RS	CSI	l_2	Others	MNIST	EMNIST	FMNIST	Others
[20]	✓	✓	–	–	–	✓	–	✓	–	–	✓	–	–	✓	–	–	–	✓	–	–	✓
[21]	✓	–	–	–	–	✓	–	✓	–	–	✓	–	–	✓	–	–	–	–	–	–	✓
[22]	–	–	–	–	–	✓	–	✓	–	–	✓	–	–	✓	–	–	–	✓	–	–	✓
[23], [41]	✓	–	–	–	–	✓	–	✓	–	–	–	–	–	✓	–	–	–	✓	–	–	✓
[24]	✓	✓	–	–	–	✓	✓	✓	–	–	✓	–	–	✓	–	–	✓	✓	–	✓	✓
[25]	✓	–	✓	–	–	–	✓	✓	✓	✓	✓	✓	✓	✓	–	–	✓	✓	–	–	–
[27], [28]	✓	✓	✓	–	–	–	–	✓	✓	–	–	–	–	✓	–	–	–	✓	–	–	✓
[29]	✓	–	✓	–	–	–	–	✓	✓	✓	–	–	–	✓	–	–	–	✓	–	–	–
[31]	–	–	–	–	–	–	✓	–	✓	–	–	–	–	✓	✓	–	✓	✓	–	–	–
[32]	✓	–	–	–	–	–	✓	✓	✓	✓	–	–	–	✓	–	–	✓	✓	–	–	–
[33]	–	✓	–	–	–	–	✓	✓	✓	–	–	–	–	✓	–	–	✓	✓	–	–	–
[38]	✓	✓	✓	–	–	–	✓	✓	✓	✓	–	–	–	✓	–	–	✓	✓	–	–	–
[40]	✓	✓	✓	–	–	–	✓	✓	✓	✓	–	–	–	✓	✓	–	✓	✓	–	–	–
[42]	✓	✓	–	–	✓	–	–	✓	✓	✓	–	–	–	✓	✓	–	–	✓	–	–	–
[43]	–	✓	–	–	–	–	✓	–	–	–	–	–	–	✓	–	–	✓	–	✓	–	✓
Proposed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	✓	–

TABLE I

COMPARISON WITH OTHER COMMUNICATION-EFFICIENT FL SYSTEMS. ABBREVIATIONS IN THE TABLE ARE DEFINED AS: "TP": TRANSMIT POWER, "TS": TRAINING STEP, "P/C": PRECODER/COMBINER, "QS": QUANTIZATION SCHEME, "SP": SCHEDULING POLICY, "RSN": RECEIVER SIDE NOISE, "CF": CHANNEL FADING, "IU": INTERFERENCE FROM OTHER USERS, "FP": FULL PARTICIPATION, "RS": RANDOM SCHEDULING POLICY, "CSI": CHANNEL STATE INFORMATION-BASED SCHEDULING POLICY, "l_2": l_2-NORM-BASED SCHEDULING POLICY.

Detailed analysis and corresponding simulation results are provided, communication cost comparison is also presented.

We consider image classification problem with MNIST, EMNIST and FMNIST datasets for deep federated learning, both i.i.d. and non-i.i.d. training dataset distributions are considered. Extensive simulations are carried out to include the effects of quantization, modulation, number of antennas, uplink and downlink SNR, ZF and MMSE precoding/combining, scheduling policy, fading and noise. Performance of the proposed system is also evaluated at different values of non-i.i.d. factor, which is a metric that quantitatively evaluates the non-i.i.d.-ness of users' training dataset distributions and will be explained in details later in Section IV. The connection between our theoretical and simulation results is highlighted. We compare our simulation results with two baselines: *1)* Both communication links (DL and UL) are assumed to be error-free. *2)* Only the UL is considered to be error-prone; that is, DL is assumed to be error-free. It is demonstrated that with UL-DL imperfections, FL converges similar to the convergence of baseline, with a very small constant gap. To the best of authors' knowledge, we are the first to evaluate FL algorithm under a comprehensive massive MIMO communication scenario with all imperfections considered and conduct corresponding convergence analysis rigorously. In Table I, we compare our communication-efficient FL framework with existing works in the literature, the comparison is made for convergence analysis, simulation settings, types of noise considered for communication, techniques used for improving communication-efficiency, datasets used for evaluations, etc. It can be witnessed that most of the work listed above try to improve communication-efficiency of a FL framework from only one perspective. On the other hand, some of the existing work investigate on multiple aspects, such as the work in [24] (quantization scheme and scheduling policy) and the work in [25], [38], [40] (transmit power and scheduling policy). However, the importance of DL transmission is not included or highlighted in the corresponding convergence analysis (the work in [25] considers the imperfections in DL channel, but non-i.i.d. dataset distribution is not discussed and the proposed framework is only evaluated on one dataset).

Organization

The rest of this paper is organized as follows. Section II and III detail the proposed system model and communication process included in the proposed framework, respectively. Simulation results are presented in Section IV. Section V concludes this paper.

Notations

$\mathcal{B}$, $\mathbf{B}, \mathbf{b}$, b represent a set, a matrix, a vector, and a scalar, respectively. $\mathbf{B}^H$ and $\mathbf{B}^{-1}$ are the Hermitian transpose and the inverse of $\mathbf{B}$, respectively. $\|\mathbf{B}\|_F$ and $\mathrm{tr}(\mathbf{B})$ are the Frobenius norm and the trace of $\mathbf{B}$, respectively. $\Re[b]$ means the real part of b. $\|\mathbf{b}\|_2$ is the l_2-norm of $\mathbf{b}$. $\mathrm{Cov}(\mathbf{b})$ is the covariance matrix of zero mean $\mathbf{b}$, i.e., $\mathbb{E}\{\mathbf{b}\mathbf{b}^H\}$, where $\mathbb{E}\{\cdot\}$ denotes the expectation operator. $\mathcal{CN}(b, \mathbf{B})$ represents a circularly symmetric complex Gaussian random vector with mean b and covariance matrix $\mathbf{B}$. The notation $\mathbf{I}_K$ is the identity matrix of size $K \times K$.

II. SYSTEM MODEL

We consider a wireless single-cell multi-user cellular 6G network, in which N users and a server or base station

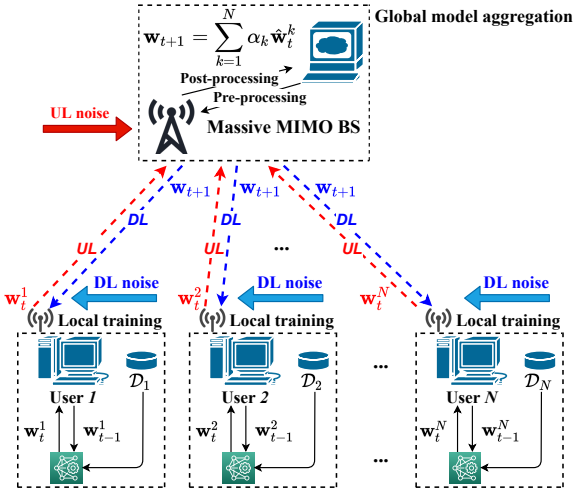


Fig. 1. Architecture of the federated learning framework with imperfect wireless links, where channel fading, intra-cell interference, quantization noise and receiver side noise are considered.

(BS) are included, as illustrated in Fig. 1. Details about the procedures presented in Fig. 1 are given in the subsequent subsections. The BS has N_{BS} number of antennas, while each user device is equipped with N_D antennas. The k^{th} user stores a local training dataset $\mathcal{D}_k = \{\mathbf{d}_i\}_{i=1}^{D_k}$ with size denoted by D_k. Each training data point $\mathbf{d}$ is stored as input-output pair $\{\mathbf{x}, y\}$, where $\mathbf{x}$ is the input to the model and y denotes the corresponding output. Total dataset size of all users within this cell can be defined as $D = \sum_{k=1}^{N} D_k$. We also assume all users' datasets are not independent and identically distributed (non-i.i.d.). Let $f(\mathbf{w}, \mathbf{d})$ denote the loss function to measure the loss between the model output based on $\mathbf{w}$ and labeled data point $\mathbf{d}$. For the k^{th} user, the loss function is calculated over all its local training data as

$$F_k(\mathbf{w}) \coloneqq \frac{1}{D_k} \sum_{\mathbf{d} \in \mathcal{D}_k} f(\mathbf{w}, \mathbf{d}). \tag{1}$$

The overall loss function for the data $\mathcal{D}_k, \forall k$ can be defined as

$$F(\mathbf{w}) \coloneqq \frac{1}{D} \sum_{k=1}^{N} \sum_{\mathbf{d} \in \mathcal{D}_k} f(\mathbf{w}, \mathbf{d}) = \sum_{k=1}^{N} \alpha_k F_k(\mathbf{w}), \tag{2}$$

where $\alpha_k = {}^{D_k}/_{D}$ is the weight of the k^{th} device such that $\alpha_k \geq 0, \forall k$ and $\sum_{k=1}^{N} \alpha_k = 1$. The optimization problem for finding optimal model parameters $\mathbf{w}^*$ such that minimizes the overall loss function can be given as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w}). \tag{3}$$

Therefore, the BS minimizes the overall loss $F(\mathbf{w})$, while each user optimizes its local loss function $F_k(\mathbf{w})$. Accounting the imperfections in UL-DL links, the major procedures of the FL algorithm for a typical (say t^{th}) communication round can be described as follows.

A. Local training

Let S denotes the number of steps in one communication round, and $\mathcal{T} = \left\{S_1, S_1 + S_2, \ldots, \sum_{r=1}^{T} S_r\right\}$ denote the time instances for each round. Then, the instances for local updates can be given as $\mathcal{S} = \left\{0, 1, 2, \ldots, S_1, \ldots, S_1 + S_2, \ldots, \sum_{r=1}^{T} S_r\right\} \setminus \mathcal{T}$, where T is the total number of communication rounds. In literature, $\mathcal{T}$ is selected uniformly, i.e., $S_r = S, r = 1, \ldots, T$, However, we set non-uniformly based on the training dataset and the communication errors.

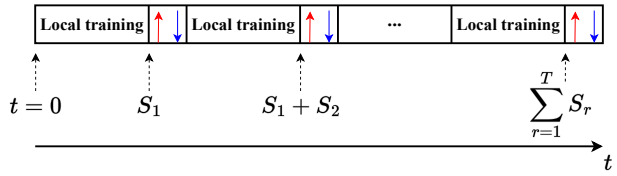


Fig. 2. Illustration of the timeline of communication rounds in federated leaning. Up and down arrows signify the uplink and downlink communications, respectively.

At the start of t^{th} round, each k^{th} user initializes the local weights by the received model from BS, sent via DL channels in the $(t-1)^{th}$ round. After initialization, each k^{th} user performs local stochastic gradient decent (SGD) steps using its local dataset as

$$\mathbf{w}_{t+1}^k = \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k, \xi_t^k), t \in \mathcal{S}, \tag{4}$$

where η_t is learning rate; ξ_t^k denotes a mini-batch, that is a subset chosen uniformly and independently from k^{th} user's local training dataset $\mathcal{D}_k$. This local training process is also presented in Fig. 1. At the end of S steps, the locally trained model, $\mathbf{w}_t^k$, is evaluated on the local testing set, and the averaged testing accuracy over all users is defined as the local testing accuracy for the t^{th} round.

B. Uploading model weights

Each user uploads the local trained model $\mathbf{w}_t^k$ to the central server or BS through UL channel (red dotted line in Fig. 1). For uplink transmissions, each user quantizes the local weights and employs suitable modulation (e.g., QPSK or QAM). Through the multi-path fading channel, the BS receives the noisy and distorted version of weights $\mathbf{w}_t^k, \forall k = 1, \ldots, N$. This uplink process can be represented using the mapping $\phi_{UL}\left(\mathbf{w}_t^k, \forall k\right)$.

C. Aggregation

After receiving uploaded parameters from all users, the BS first combines, demodulates and dequantizes the received weights ("post-processing" in Fig. 1), and the new updated global weights are obtained via aggregating these noisy weights. This aggregation step can be represented using the mapping $\phi_{av}(\cdot)$. At the end of aggregation, the BS has $\phi_{av}\left(\phi_{UL}\left(\mathbf{w}_t^k, \forall k\right)\right)$. At this point, the global testing accuracy is defined as the accuracy of the aggregated model tested on the testing set at server, which is different from the testing sets of users. Note that for the simplicity of aggregation, it is also assumed that the model weights of each user ($\mathbf{w}_t^k \in \mathbb{R}^{W \times 1}$) are real-valued and of same size $W \times 1$.

D. Downloading updated weights

In this step, BS broadcasts the aggregated model to all users via DL channel (blue dotted line in Fig. 1). The received model at the k^{th} user is denoted by the mapping $\phi_{k,DL}\left(\phi_{av}\left(\phi_{UL}\left(\mathbf{w}_t^k, \forall k\right)\right)\right)$, where $\phi_{k,DL}\left(\cdot\right)$ represents the downlink communication process for user k. This process includes quantization, modulation and precoding ("preprocessing" in Fig. 1) at BS, and the effects of transmission through the k^{th} multi-path downlink channel.

In this way, each round progresses until the global (or local) testing accuracy reaches convergence. An illustration of the FL timeline can be found in Fig. 2. In this figure, t denotes the time index for local step. $S_1, S_1+S_2, \ldots, \sum_{r=1}^{T} S_r \in \mathcal{T}$ are time instances when UL communication (red up arrow), weights aggregation and DL communication (blue down arrow) take place. Detailed analysis about these three mappings will be explained in the following section.

III. Uplink and downlink communication

Since BS is equipped with a large number of antennas, we consider ZF approach at the BS, and analyze the impacts on model weights using channel hardening effect. Performance of ZF techniques in massive MIMO system has been investigated in [44] and [45]. First, the update step of FL algorithm in a round can be expressed as

$$\mathbf{v}_{t+1}^k = \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k, \xi_t^k), \tag{5}$$

$$\mathbf{w}_{t+1}^k = \begin{cases} \mathbf{v}_{t+1}^k, & \text{if } t+1 \notin \mathcal{T}, \\ \phi_k\left(\mathbf{v}_{t+1}^k, \forall k\right), & \text{if } t+1 \in \mathcal{T}, \end{cases} \tag{6}$$

where the variable $\mathbf{v}_t$ is introduced to denote the immediate result of one step SGD. We define $\mathbf{g}_t = \sum_{k=1}^{N} \alpha_k \nabla F_k(\mathbf{w}_t^k, \xi_t^k)$ and $\bar{\mathbf{g}}_t = \sum_{k=1}^{N} \alpha_k \nabla F_k(\mathbf{w}_t^k)$. Therefore, $\mathbb{E}\left\{\mathbf{g}_t - \bar{\mathbf{g}}_t\right\} = \mathbf{0}$. The function $\phi_k(\cdot)$ represents the distortion arising due to uplink and downlink communications, which can be decomposed into uplink operation $\phi_{UL}(\cdot)$, averaging $\phi_{av}(\cdot)$, and downlink broadcast operation $\phi_{k,DL}(\cdot)$, i.e., a composition of three mappings as $\phi_k\left(\mathbf{w}_t^k, \forall k\right) = \phi_{k,DL} \circ \phi_{av} \circ \phi_{UL}\left(\mathbf{w}_t^k, \forall k\right)$.

A. Uplink mapping

In the uplink process, W real-valued weights are resized into complex-valued data streams for transmissions, say $\mathbf{W}_t^k \in \mathbb{C}^{N_D \times \tau_{UL}}$, where $\tau = \frac{W}{2N_D}$. The received signal at the massive MIMO BS can be written as

$$\mathbf{Y}_t^{(UL)} = \sum_{k=1}^{N} \mathbf{H}_t^k \mathbf{W}_t^k \sqrt{\frac{P_{UL}}{N_D \sigma_w^2}} + \mathbf{Z}_t^{(UL)}, \tag{7}$$

where entries of $\mathbf{W}_t^k$ are assumed to have zero mean and σ_w^2 variance; $\mathbf{Z}_t$ denotes zero-mean Gaussian noise with σ^2 variance, and P_{UL} is the transmit power constraint for each device, i.e., $tr\mathbb{E}\left\{\mathbf{W}_t^k \mathbf{W}_t^{kH}\right\} = \tau N_D \sigma_w^2$. With the zero forcing receiver with $N_{BS} \geq N \cdot N_D$ number of antennas, the estimated weights at the BS can be given as

$$\begin{aligned} \hat{\mathbf{W}}_t &= \sqrt{\frac{N_D \sigma_w^2}{P_{UL}}} \left(\mathbf{H}_t^H \mathbf{H}_t\right)^{-1} \mathbf{H}_t^H \mathbf{Y}_t^{(UL)} \\ &= \mathbf{W}_t + \sqrt{\frac{N_D \sigma_w^2}{P_{UL}}} \left(\mathbf{H}_t^H \mathbf{H}_t\right)^{-1} \mathbf{H}_t^H \mathbf{Z}_t^{(UL)}, \end{aligned}$$

where $\mathbf{H}_t = \left[\mathbf{H}_t^1, \ldots, \mathbf{H}_t^N\right]$ and $\mathbf{W}_t^T = \left[\mathbf{W}_t^{1T}, \ldots, \mathbf{W}_t^{NT}\right]$. Each column of the estimation error term has zero mean and covariance matrix as $\frac{N_D \sigma_w^2 \sigma^2}{P_{UL}} \left(\mathbf{H}_t^H \mathbf{H}_t\right)^{-1}$. Owing to channel hardening $\frac{1}{N_{BS}} \mathbf{H}_t^H \mathbf{H}_t \to \mathbf{I}$, we simplify the covariance matrix of the effective noise above as $\frac{N_D \sigma_w^2 \sigma^2}{P_{UL}} N_{BS}^{-1} \mathbf{I}_{N_D}$. The variable $\hat{\mathbf{W}}_t$ can be used to express the estimated model weight for each user at the BS as $\hat{\mathbf{w}}_t^k = \mathbf{w}_t^k + \mathbf{z}_{t,UL}^k$, where $\mathbf{z}_{t,UL}^k \sim \mathcal{N}\left(\mathbf{0}, \frac{N_D \sigma_w^2 \sigma^2}{2 P_{UL} N_{BS}} \mathbf{I}\right)$ and the factor of 2 comes from converting the complex random variables to the real ones.

B. Aggregation mapping

Having $\hat{\mathbf{W}}_t$ at the BS, the averaging operation can be given as

$$\phi_{av}\left(\hat{\mathbf{W}}_t\right) = \sum_{k=1}^{N} \alpha_k \hat{\mathbf{w}}_t^k \tag{8}$$

$$= \bar{\mathbf{w}}_t + \bar{\mathbf{z}}_{t,UL}, \tag{9}$$

where $\bar{\mathbf{w}}_t = \sum_{k=1}^{N} \alpha_k \mathbf{w}_t^k$, $\bar{\mathbf{z}}_{t,UL} \sim \mathcal{N}\left(\mathbf{0}, \frac{\|\boldsymbol{\alpha}\|_2^2 \sigma_w^2}{2} \frac{\sigma^2 N_D}{P_{UL} N_{BS}} \mathbf{I}\right)$ and $\|\boldsymbol{\alpha}\|_2^2 = \sum_{k=1}^{N} \alpha_k^2$.

C. Downlink mapping

The aggregated vector, say $\hat{\bar{\mathbf{w}}}_t = \phi_{av}\left(\hat{\mathbf{W}}_t\right)$, is then broadcasted to each user device via the base station through the DL channel. Since each device can receive only N_D data streams, the vector $\hat{\bar{\mathbf{w}}}_t$ is resized into a complex-valued matrix $\hat{\bar{\mathbf{W}}}_t \in \mathbb{C}^{N_D \times \tau}$ with $\tau = \frac{W}{2N_D}$. The received signal at the k^{th} device can be given as

$$\mathbf{Y}_{k,t}^{(DL)} = \mathbf{H}_t^{kH} \mathbf{F}_t \hat{\bar{\mathbf{W}}}_t p_{DL} + \mathbf{Z}_{k,t}^{(DL)},$$

where channel reciprocity is assumed; and $\mathbf{F}_t$ denotes the $N_{BS} \times N_D$ precoder such that $\mathbf{F}_t^H \mathbf{F}_t = \mathbf{I}$. The transmit power constraint (P_{DL}) is given as

$$p_{DL}^2 = \frac{P_{DL} \tau}{tr\mathbb{E}\left\{\mathbf{F}_t \hat{\bar{\mathbf{W}}}_t \hat{\bar{\mathbf{W}}}_t^H \mathbf{F}_t^H\right\}} = \frac{P_{DL}}{\hat{\sigma}_w^2 N_D},$$

where $\hat{\sigma}_w^2 = \|\boldsymbol{\alpha}\|_2^2 \sigma_w^2 \left(1 + \frac{\sigma^2 N_D}{P_{UL} N_{BS}}\right)$. Then, the estimates at the k^{th} user can be obtained as $\left(\mathbf{H}_t^{kH} \mathbf{F}_t\right)^{-1} \mathbf{Y}_{k,t}^{(DL)} p_{DL}^{-1}$. The corresponding covariance matrix for each column of the estimation error $\left(\mathbf{H}_t^{kH} \mathbf{F}_t\right)^{-1} \mathbf{Z}_{k,t}^{(DL)} p_{DL}^{-1}$ can be computed as

$$\begin{aligned} &\frac{\hat{\sigma}_w^2 N_D \sigma^2}{P_{DL}} \left(\mathbf{H}_t^{kH} \mathbf{F}_t\right)^{-1} \left(\mathbf{H}_t^{kH} \mathbf{F}_t\right)^{-H} \\ &= \frac{\hat{\sigma}_w^2 N_D \sigma^2}{P_{DL}} \left(\mathbf{F}_t^H \mathbf{H}_t^k \mathbf{H}_t^{kH} \mathbf{F}_t\right)^{-1} \\ &\approx \frac{\hat{\sigma}_w^2 N_D \sigma^2}{P_{DL}} N_D^{-1} \mathbf{I} = \frac{\hat{\sigma}_w^2 \sigma^2}{P_{DL}} \mathbf{I}, \end{aligned}$$

where $\mathbf{H}_t^k \mathbf{H}_t^{kH}$is approximated to $N_D \mathbf{I}_{BS}$, owing to channel hardening effect. Thus, we can write $\phi_{k,DL}\left(\hat{\bar{\mathbf{w}}}_t\right) = \hat{\bar{\mathbf{w}}}_t + \bar{\mathbf{z}}_{t,DL}^k$, where $\bar{\mathbf{z}}_{t,DL}^k \sim \mathcal{N}\left(\mathbf{0}, \frac{\hat{\sigma}_w^2 \sigma^2}{2P_{DL}}\mathbf{I}\right)$ is the Gaussian noise at the typical user.

Overall, one can express the composite uplink-downlink aggregation process as

$$\phi_k\left(\mathbf{v}_t^k, \forall k\right) = \phi_{k,DL} \circ \phi_{av} \circ \phi_{UL}\left(\mathbf{v}_t^k, \forall k\right) \tag{10a}$$

$$= \phi_{k,DL} \circ \phi_{av}\left(\hat{\mathbf{V}}_t\right) \tag{10b}$$

$$= \phi_{k,DL}\left(\bar{\mathbf{v}}_t + \bar{\mathbf{z}}_{t,UL}\right) \tag{10c}$$

$$= \bar{\mathbf{v}}_t + \underbrace{\bar{\mathbf{z}}_{t,UL} + \bar{\mathbf{z}}_{t,DL}^k}_{\bar{\mathbf{z}}_t^k}, \tag{10d}$$

where $\bar{\mathbf{v}}_t = \sum_{k=1}^N \alpha_k \mathbf{v}_t^k$, and the variance for the effective noise $\bar{\mathbf{z}}_t^k$, σ_z^2, can be seen in the proof of Lemma 1.

Remark: In the above mapping, the weights can be considered with quantization and modulation by introducing quantization error, e.g., with uniform quantization for real and imaginary parts in Eqn. (7), the resulting error variance can be given as sum of noise variance σ^2 and the variance of the quantization error $\frac{\Delta^2}{6}$, where Δ is the quantization step size. The total effective variance can be obtained similarly with additive quantization error variance, which affects the convergence in the same way as the noise variance σ_z^2.

Lemma 1. *For UL-DL imperfect communication, given any model $\mathbf{w}^*$, the l_2-norm of the model error is bounded as*

$$\mathbb{E}\left\|\bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\right\|_2^2 \leq \mathbb{E}\left\|\bar{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|_2^2 + W\sigma_z^2 \delta_{t+1\in\mathcal{T}}, \tag{11}$$

$$\sigma_z^2 = \frac{\|\boldsymbol{\alpha}\|_2^2 \sigma_w^2}{2} \frac{\sigma^2}{P_{DL}} \left[\left(\frac{P_{DL}}{\sigma^2} + \|\boldsymbol{\alpha}\|_2^2\right) \frac{\sigma^2 N_D}{P_{UL} N_{BS}} + \|\boldsymbol{\alpha}\|_2^2\right],$$

where $\delta_{t\in\mathcal{T}}$ is an indicator function, taking value 1, when $t \in \mathcal{T}$, and 0 otherwise.

Proof: Proof is given in Appendix-A. ∎

The above result shows that if the UL-DL links are perfect, i.e., $\sigma_z^2 = 0$, the models can be aggregated perfectly, and leads to similar errors. In other words, when $\sigma_z^2 = 0$, $\bar{\mathbf{v}}_{t+1} = \bar{\mathbf{w}}_{t+1}$. Regarding the effective noise variance σ_z^2, it can be seen that the downlink reliability is more important than the uplink's, since in the above expression of σ_z^2, the downlink SNR $\frac{P_{DL}}{\sigma^2}$ is the main factor in division.

D. Assumptions on the loss function

Following assumptions for the loss function $F_k(\cdot), k = 1, \ldots, N$ are made for the convergence analysis [10], [20], [46], [47].

Assumption 1. *(Smoothness)*: $F_1, \ldots, F_N$ are all L-smooth, for $\forall \mathbf{v}, \mathbf{w}$, $F_k(\mathbf{v}) \leq F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{L}{2}\|\mathbf{v}-\mathbf{w}\|_2^2$.

Assumption 2. *(Convexity)*: $F_1, \ldots, F_N$ are all μ-strongly convex, for $\forall \mathbf{v}, \mathbf{w}$, $F_k(\mathbf{v}) \geq F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v}-\mathbf{w}\|_2^2$.

Assumption 3. *(Bounded variance for SGD)*: The variance of SGD satisfies

$$\mathbb{E}\left\|\nabla F_k\left(\mathbf{w}_t^k, \xi_t^k\right) - \nabla F_k\left(\mathbf{w}_t^k\right)\right\|_2^2 \leq \sigma_k^2, \tag{12}$$

which leads to $\mathbb{E}\left\|\sum_{k=1}^N \alpha_k \nabla F_k(\mathbf{w}_t^k, \xi_t^k) - \nabla F(\bar{\mathbf{w}}_t)\right\|_2^2 \leq \sum_k \alpha_k^2 \sigma_k^2$.

Assumption 4. *(Uniformly bounded gradient)*: l_2-norm of gradients satisfies

$$\mathbb{E}\left\|\nabla F_k\left(\mathbf{w}_t^k, \xi_t^k\right)\right\|_2^2 \leq G^2, \forall k, \tag{13}$$

which provides $\mathbb{E}\left\|\nabla F(\bar{\mathbf{w}}_t)\right\|_2^2 \leq \|\boldsymbol{\alpha}\|_2^2 G^2$.

Assumptions 1 and 2 are the standard ones for l_2-norm regularized linear regression, logistic regression, and softmax classifier. Assumptions 3 and 4 are based on bounded gradients and are helpful in bounding the model-error variance while analyzing the convergence. The following result utilizes μ-strong convexity assumption (Assumption 2) to bound the next step error.

Lemma 2. *Let Assumption 1-4 hold. Then, with respect to the optimized model weights $\mathbf{w}^*$, the l_2-norm of the next step model update error can be bounded as*

$$\mathbb{E}\left\|\bar{\mathbf{v}}_{t+1} - \mathbf{w}^*\right\|_2^2 \leq (1 - \eta_t \mu)\, \mathbb{E}\left\|\bar{\mathbf{w}}_t - \mathbf{w}^*\right\|_2^2 \tag{14a}$$

$$+ \eta_t^2 \left(\sum_k \alpha_k^2 \sigma_k^2 + G^2\right). \tag{14b}$$

Proof: Proof is provided in Appendix-B. ∎

The above result shows that if η_t and μ are chosen properly, one can show the model updates leads to non-increasing errors. This important result is proved in the following theorem, which employs the smoothness of F_k (Assumption 1).

Theorem 3. *Let Assumption 1-4 hold and σ_z^2 be defined therein. Then, the proposed FL algorithm satisfies*

$$\mathbb{E}\left\{F\left(\bar{\mathbf{w}}_T\right)\right\} - F^* \leq \frac{L}{2} \cdot \frac{\nu}{T + \beta_1}, \tag{15}$$

where $\eta_t = \frac{\beta_0}{t+\beta_1}$, $\nu \geq \frac{\beta_0^2 C}{\beta_0 \mu - 1}$, $\beta_1 > 0$, and $C = \sum_k \alpha_k^2 \sigma_k^2 + G^2 + W\frac{\sigma_z^2}{\eta_t^2}\delta_{t+1\in\mathcal{T}}$, $\beta_0 > \frac{1}{\mu}$.

Proof: Proof is provided in Appendix-C. ∎

Theorem 3 tells that the modeling error for the UL-DL FL algorithm reduces in $\mathcal{O}\left(T^{-1}\sigma_z^{-2}\right)$ manner, that is the algorithm converges in time, and the accuracy is inversely proportional to communication error variance σ_z^2, whose expression accounts for and explains the effects of the communication constraints on the convergence. We can reduce the communication cost via two factors, viz., varying the transmit power (equivalently varying SNRs) and reducing the number of communicating rounds (equivalently varying the number of local steps). Note that the quantization error also affects the variance σ_z^2, proving the number of quantization bits to be the third factor. The convergence bound linearly depends on the L-smoothness, whereas it is inversely proportional to the μ-convexity; that is, the stronger the convexity is, the lower the convergence error is. For a FL training process, we can define the overall learning cost as:

Overall communication cost

= # of communication rounds × # of model parameters

× # quantization bits × normalized transmit power. (16)

In our simulation, we make attempt to vary these factors as follows.

Communication rounds: From the convergence analysis, we see that weights converge as iterations progress, which means there is less need to communicate weights at later iterations. Keeping this fact in mind, we set two modifications for improving the communication cost. The first one includes the decreasing number of local steps, i.e., $S_1 \geq S_2 \geq \cdots \geq S_T$, where the equality $S_r = S_{r+1}$ implies that for much larger iterations near convergence, reducing local iterations cannot further improve the performance, e.g., including the case $S_r = 1$ for r large. The second modification is to select users by comparing the l_2-norm of difference in weights, that is, at $t = \sum_{\hat{r}=1}^{r} S_{\hat{r}}$, sort the values of $\left\|\mathbf{w}_t^k - \mathbf{w}_{t-S_r}^k\right\|_2, \forall k$ and choose $N_0 < N$ users with higher values. It provides more opportunities to users who have less convergence with respect to others. These modifications not only improve the communication cost of transmission, but also provide other users to have more diversity in physical layer communications, i.e., more reliable channels.

Remark: Since in practice, distances between users and BS are different, thus, it could be more reasonable to design a update policy $(S_{k1}, \ldots, S_{kT})$ based on a user's local situation rather than the global updating policy $(S_1, \ldots, S_T)$, e.g., within a global communication round, users that are closer to the BS may be able to perform more local training steps compared with those who are far from the BS.

Transmit power: Since σ_z^2 depends on uplink and downlink SNRs, this gap reduces to a constant if $\sigma_z^2 \propto T$. It also shows that the SNR in UL-DL communications can be increased as the iterations progress, that is, the effective noise variance σ_z^2 should be proportional to the step size η_t^2, that is, $\sigma_z^2 \propto \eta_t^2 \propto \frac{1}{(t+\beta_1)^2}$. At the same time, In the learning process, at lower iterations, when weights are not converged, we can decrease the transmit power, since they are being averaged for more updates. Similarly, for higher iterations, when weights are close to convergence, one can increase the transmit power for accurate communications and averaging. Towards that, let $P_{DL,r} = P_{DL,0}(1 - \exp(-\gamma r))$ and $P_{UL,r} = P_{UL,0}(1 - \exp(-\gamma r))$.

Therefore, for T rounds, the incurred energy consumption is given as

$$\begin{aligned}\mathcal{E}_{1T} &= N \sum_{r=1}^{T} P_{UL,0}(1 - \exp(-\gamma r)) \\ &\quad + \sum_{r=1}^{T} P_{DL,0}(1 - \exp(-\gamma r)) \qquad (17) \\ &= (NP_{UL,0} + P_{DL,0}) \left(T - e^{-\gamma} \frac{1 - e^{-\gamma T}}{1 - e^{-\gamma}}\right). \qquad (18)\end{aligned}$$

On the other hand, conventionally, when constant power P_0 is used for all communication rounds, the total incurred power consumption is $\mathcal{E}_{2T} = (NP_{UL,0} + P_{DL,0})\, T > \mathcal{E}_{1T}$. Therefore, the above analysis shows that by varying SNR, the energy consumption of the system can be improved without losing the performance at the convergence. The normalized value of transmit power $\mathcal{P}$ can be defined as:

$$\mathcal{P} = \frac{\mathcal{E}_{1T}}{\mathcal{E}_{2T}} = \frac{\left(T - e^{-\gamma} \frac{1 - e^{-\gamma T}}{1 - e^{-\gamma}}\right)}{T}. \qquad (19)$$

In our communication cost comparison (see in next section), this variable serves as a weighting factor that reflects the improvement of transmit power consumption introduced by varying P_r during training.

Quantization: We employ the uniform quantization scheme for the model weights prior to transmission. We vary the number of quantization bits to reduce the overhead in each round. In subsubsection IV-D5, we also consider another quantization scheme based on stochastic rounding [20]. Specifically, the procedures of this quantization scheme include scale up, stochastic rounding, limit and scale down.

Overall, the aforementioned features and factors in our system can be summarized in the Algorithm 1. Similar to Fig. 1, Algorithm 1 explains the procedures of proposed FL framework, where lines $9-10$ and $26-27$ denote transmit power variation, followed by the transmission process (line 11 and line 28, for DL and UL, respectively). At local user side, lines $17-20$ represent the local training procedures, where

Algorithm 1 Communication-efficient FL algorithm

1: **Initialization:**
2: Initialize $T, N, N_0, P_{UL,0}, P_{DL,0},$
3: **for** $r = 1, \ldots, T$ **do**
4: **if** $r = 1$ **then**
5: All users initialize its local weights $\mathbf{w}_0^k, \forall k$
6: **else**
7: Downlink transmission:
8: 1. BS quantizes, modulates and precodes $\hat{\hat{\mathbf{w}}}_t$
9: 2. Value of $P_{DL,r}$ is updated as:
10: $P_{DL,r} = P_{DL,0}(1 - \exp(-\gamma r))$
11: 3. BS broadcasts $\hat{\hat{\mathbf{w}}}_t$ with power $P_{DL,r}$.
12: **end if**
13: **for** $k = 1, \ldots, N$ **do**
14: **if** $r \neq 1$ **then**
15: $\mathbf{w}_t^k = \phi_{k,DL}(\hat{\hat{\mathbf{w}}}_t)$
16: **end if**
17: **for** $t = \sum_{\hat{r}=1}^{r} S_{\hat{r}} - S_r, \ldots, \sum_{\hat{r}=1}^{r} S_{\hat{r}} - 1$ **do**
18: Update local weights as:
19: $\mathbf{w}_{t+1}^k = \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k, \xi_t^k)$
20: **end for**
21: **end for**
22: Sort the values of $\left\|\mathbf{w}_{t+1}^k - \mathbf{w}_{t+1-S_r}^k\right\|_2, \forall k$.
23: Select $N_0 < N$ users with higher values.
24: Uplink transmission:
25: 1. Selected users quantize and modulate weights.
26: 2. Value of $P_{UL,r}$ is updated as:
27: $P_{UL,r} = P_{UL,0}(1 - \exp(-\gamma r))$
28: 3. Users upload weights with power $P_{UL,r}$.
29: 4. BS combines, demodulates and dequantizes the received weights.
30: 5. BS performs aggregation to produce $\hat{\hat{\mathbf{w}}}_{t+1}$.
31: **end for**

local steps variation is applied. Lines $22-23$ stand for the l_2-norm scheduling policy before UL transmission.

Further, in Theorem 3, as $t \to \infty$, $\eta_t \to 0$, that is, for convergence of FL, the UL-DL communication links should be reliable, where the importance of DL links' reliability can be inferred from Lemma 1. Moreover, it can be noted that the above bound in Eqn. (15) is much tighter than the one, originally expressed in literature [10][1].

The following result expresses the convergence bound for the case when the learning rate is constant throughout the training process.

Corollary 4. *When learning rate is set to be a constant through whole training process, i.e., $\eta_t = \eta$, the average model accuracy is bounded by a constant gap with respect to the optimum model as*

$$\mathbb{E}\left\{F\left(\bar{\mathbf{w}}_T\right)\right\} - F^* \leq \frac{L}{2} \cdot \frac{\eta C}{\mu}, \tag{20}$$

where $C = \sum_k \alpha_k^2 \sigma_k^2 + G^2 + W \frac{\sigma_z^2}{\eta^2} \delta_{t+1 \in \mathcal{T}}$.

Proof: Proof is derived in Appendix-D. ∎

The above bound shows that for the constant learning rate η, the FL learning algorithm converges to a constant gap.

IV. Simulation results

A. Simulations setup

For an image classification task, we choose MNIST [48], EMNIST [49] and FMNIST [50] datasets for both i.i.d. and non-i.i.d. training data distributions. MNIST contains gray-scale images of size 28×28 for handwriting of numbers, ranging from 0 to 9. EMNIST is a more challenging dataset, from which we choose EMNIST-letters dataset containing gray-scale images of size 28×28 for handwriting of English letters, i.e., 26 classes. Similar to but more challenging than MNIST, FMNIST also has 10 classes, which correspond to 10 categories of clothes.

1) Training data preparation: For data augmentation purpose, we apply random rotation (ranging from -20 degree to 20 degree), random horizontal translation (ranging from -3 pixels to 3 pixels) and random vertical translation (ranging from -3 pixels to 3 pixels) for each training image of MNIST, EMNIST and FMNIST datasets.

2) Communication setup: As practiced and investigated in [25], [40] and [51], we consider $N = 25$ users, each with single $N_D = 1$ antennas. The learning weights of local and global models are stored in 32-bit float point format. A uniform quantization is employed to quantize the model weights, followed by Quadrature Phase Shift Keying (QPSK) for modulation. For i.i.d. Rayleigh fading model, channel state information (CSI) is assumed constant within each communication round, and is known at the BS side. Note that for the purpose of simplicity, in this paper we only consider perfect CSI. For the scenario with imperfect CSI, such as [52], similar convergence analysis and simulations can be conducted by including the channel estimation error in the noise variance.

3) Local model: For MNIST dataset, the local deep learning model comprises of a convolutional neural network (CNN) model with three convolutional layers, a fully connected layer, and a final classification layer with softmax function. Each convolutional layer is followed by a batch normalization layer, a ReLU activation layer, and a 2×2 max-pooling layer. Three convolutional layers contain 8, 16 and 32 filters, respectively. All filters are of size 3×3. When EMNIST and FMNIST datasets are considered, the local deep learning model is same to that used in MNIST scenario except that no batch normalization layers are used.

4) MNIST: Each user has 50 training samples locally, and has a testing set of 200 samples. For i.i.d. case, overall training and testing sets are randomly shuffled and evenly splitted up among users in the cell, whereas in the non-i.i.d. case, for each local user, two labels are randomly selected for dominating labels. We define the fraction of the training set occupied by the samples with dominating labels as "non-i.i.d. factor". Note that the testing images in testing sets still have uniform distribution, that is, i.i.d. testing data samples. At each user, training parameters are set as follows. In the "fix local steps" scheme, learning rate is set to be 0.01, the number of local epochs is set to be 3 and for 50 training images, batch size is 10, leading to a total number of local updates to be $S = 3 \times 5 = 15$. In the scenario where number of local steps is varied, learning rate and batch size are set to be 0.01 and 10, respectively. S_r decreases according to $S_r = \frac{D_k}{\lfloor \frac{\xi_t^k}{} \rfloor} \left(6 - \lfloor \frac{r-1}{15} \rfloor\right)$, where $\lfloor\rfloor$ denotes floor operation.

5) EMNIST and FMNIST: For these two datasets, each user has 100 training samples locally, and has a testing set of 200 samples. At each user, training parameters are set as follows. Local learning rate 0.01, the numbers of local training epochs and batch size are 7 and 10, respectively.

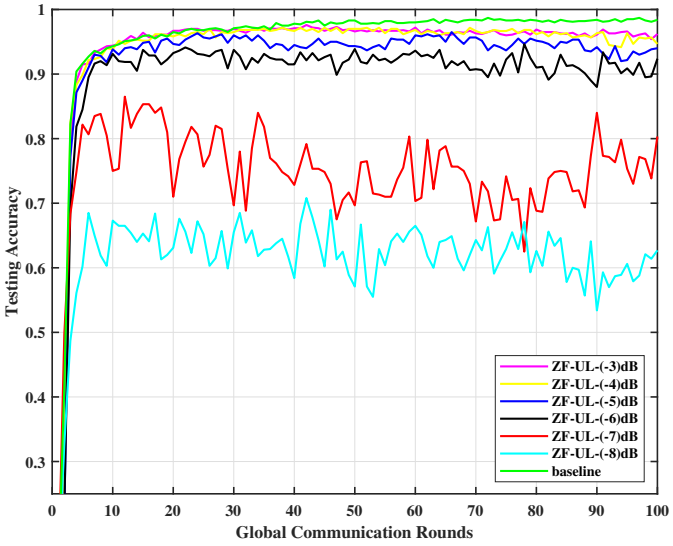
6) Server/BS setup: The server, or the massive MIMO base station, is considered to be equipped with 64 of antennas (except in subsections IV-C and IV-D, where 128 antennas are considered at BS). Note that 64 and 128 are common choices for 5G and 6G massive MIMO communication scenario [53]–[55]. Unless otherwise noted, the number of global communication rounds is set to $T = 100$. The BS contains a global testing set, which contains 200 testing samples, and is used for getting global server side testing accuracy in each round.

In the following subsections, we will present our simulation results. We start with simulations on i.i.d. MNIST dataset (subsections IV-B and IV-C), where we mainly focus on the impacts introduced by UL and DL transmission errors with different values of SNRs. After that, simulation results on non-i.i.d. MNIST under different system schemes proposed in this paper are compared and analyzed (subsection IV-D), where transmit power, number of local steps, quantization bits and non-i.i.d. factor are varied. Simulations on EMNIST and FMNIST datasets are provided in subsection IV-E.

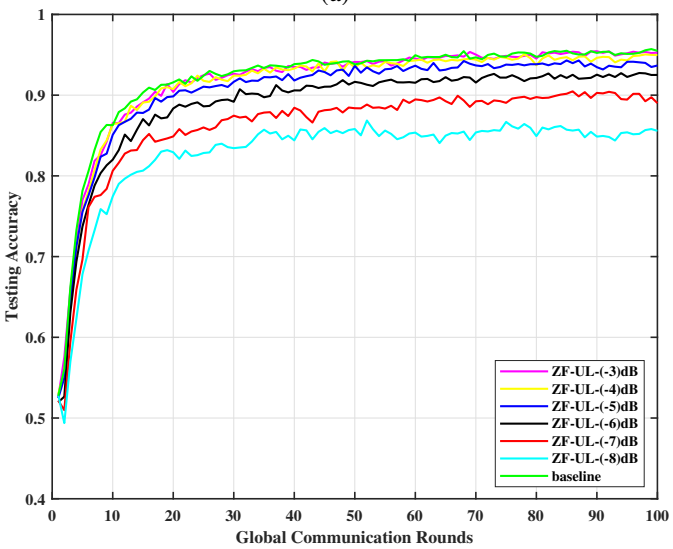
B. Testing accuracy for fixed learning rate

1) For different UL-SNRs and a fixed DL-SNR:

[1]For perfect UL-DL links ($\sigma_z^2 = 0$), the bound from [10] can be written as $\mathbb{E}\left\{F\left(\bar{\mathbf{w}}_T\right)\right\} - F^* \leq \frac{L}{2} \frac{\nu}{T + \beta_1}$, where $\eta_t = \frac{\beta_0}{t+\beta_1}$, $\nu = \frac{2}{\mu}\left(\frac{2}{\mu} C + \frac{\mu}{2} \gamma \mathbb{E}\left\|\mathbf{w}_1 - \mathbf{w}^*\right\|_2^2\right)$, $\beta_1 = \gamma$, $\beta_0 = \frac{2}{\mu}$, $\Gamma = F^* - \sum_{k=1}^N \alpha_k F_k^*$ and $C = \sum_k \alpha_k^2 \sigma_k^2 + 6L\Gamma + 8(S-1)^2 G^2$.

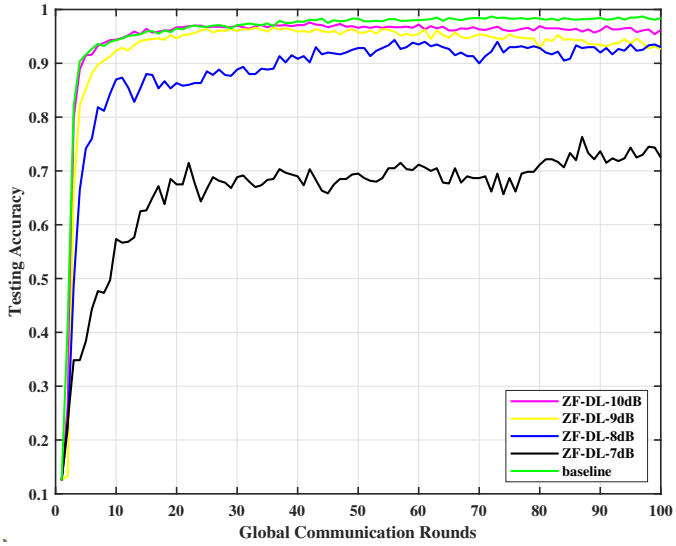


(a)

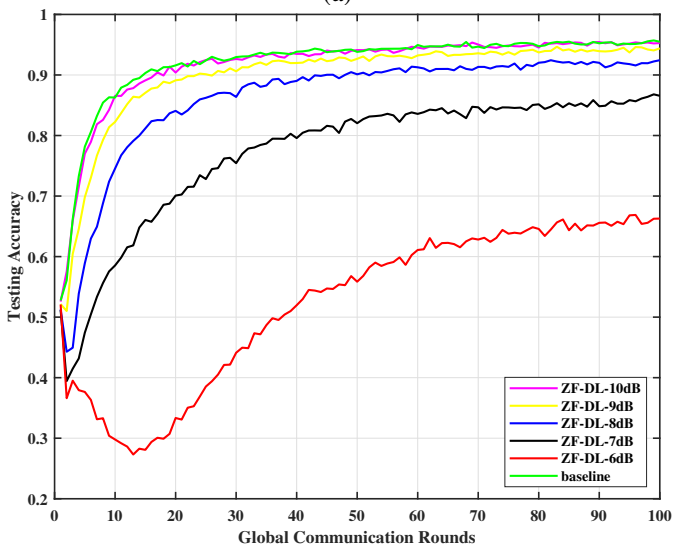


(b)

Fig. 3. Testing accuracy of the training progress at server side in (a), and local user side in (b), for different SNRs of UL links while DL-SNR is fixed to 10 dB; that is, $P_{DL,0}/\sigma^2 = 10$ dB.



(a)



(b)

Fig. 4. Testing accuracy of the training progress at server side in (a), and local user side in (b), for different SNRs of DL links while UL-SNR is fixed to -3 dB; that is, $P_{UL,0}/\sigma^2 = -3$ dB.

Fig. 3 plots the progress of testing accuracy in different communication rounds at both server and user sides (in (a) and (b) respectively), when UL-SNR is varied and DL-SNR is kept fixed to 10 dB. Simulations are averaged over 10 times.

From both Figs. 3 (a) and (b), it can be seen that increasing the SNR improves the performance of the trained models. At higher SNRs, testing accuracy is similar to that of baseline approach (perfect links). In the Fig 3, the gap at the convergence, is due to an additional term of σ_z^2 arising in Theorem 3 and Corollary 4. At the server side, there are more fluctuations in accuracy as compared to that at user side. It is due to the fact that the server receives the noisy version of locally updated weights owing to imperfect UL links. Also, the model at the server side is updated once every S steps. The almost steady behavior at local users is due to the reason that the accuracy is computed just before uplink transmission, that is, after the S steps local updates of received model weights. Note that green curves in Figs. 3 (a) and (b) denote the baseline considered in [10], where both UL and DL channels are considered to be error-free.

2) For different DL-SNRs and a fixed UL-SNR:

Fig. 4 depicts the plots of testing accuracy progress when UL-SNR is fixed to -3 dB, and downlink SNR is varied between 7 dB to 10 dB. Similar to Fig. 3, the accuracy at the sever side has more fluctuations that that at the users' side, and the accuracy is increased close to that of baseline approach for positive increments in DL-SNRs. The notable fact here is that the DL-SNR has a direct impact on the accuracies at both sides, as inferred in Lemma 1 and Theorem 3. This can be observed by comparing Fig. 4 with Fig. 3. At the server side, the fluctuations become less in Fig. 4 (a) as SNR is increased, as compared to Fig. 3 (a). Similar observations can be seen

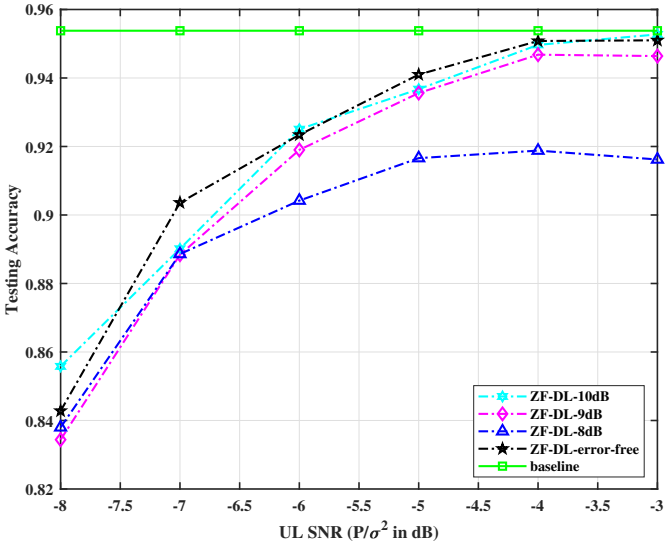


Fig. 5. Converged local testing accuracy versus SNRs after $T = 100$ global communication rounds.

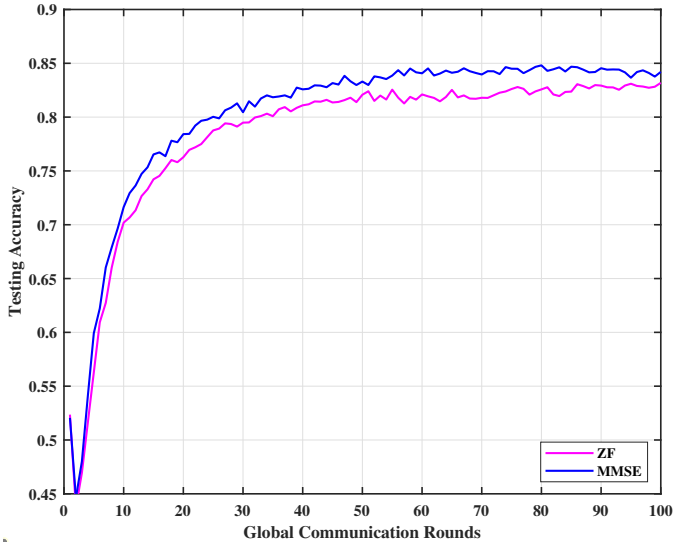


Fig. 6. Testing accuracy at local user side on MNIST i.i.d. dataset when different precoding methods are implemented. $N_{BS} = 64$. Values of DL-SNR and UL-SNR are 8 dB and -8 dB, respectively; that is, $P_{DL,0}/\sigma^2 = 8$ dB, $P_{UL,0}/\sigma^2 = -8$ dB.

for Fig. 4 (b) and Fig. 3 (b). In this figure, the decline and rise of the accuracy at the first several rounds is due to random initialization of weights. When DL channel quality is relatively bad, e.g., DL-SNR equals 6 dB, such trend is more obvious and thus again, verifying our claim that DL channel quality is more important than that of UL. Similar to Fig. 3, green curves in Figs. 4 (a) and (b) denote the baseline with perfect communication links introduced in [10]

Figure 5 shows the converged local testing accuracy versus UL-SNRs with different plots depicting different DL-SNRs. It can be seen that the testing accuracy increases with increase in any of UL or DL SNR, and for higher SNR, it reaches close to the perfect case, i.e., the baseline approach. In this figure, apart from comparing our system's performance with baseline (green curve) introduced by [10], we also plot the simulation results based on the assumption that only DL links are considered to be error-free (black curve) and this assumption has been made in many existing work [37]–[40].

3) For different precoding methods:

In Fig. 6, we plot the average testing accuracy on i.i.d. MNIST dataset with ZF and MMSE precoding/combining methods. It can be witnessed that as compared with ZF, MMSE is able to achieve a slightly better performance and converges faster. In high SNR regime, this difference becomes negligible, therefore, for the rest of simulations, we adopt the ZF strategy. It can also be noted that the convergence bound depends on the statistical error variance which is approximately similar in high SNR regime for both the precoding strategies.

4) Scheduling policy:

In Fig. 7, the average testing accuracies at local user side are compared under full participation (magenta curve), UL-channel-state-based scheduling policy (red curve) and l_2-norm-based scheduling policy (blue curve). In both partial participation cases, within each global communication round, 15 users out of 25 are selected. In UL channel state scheduling policy, users with better UL communication links quality are selected, while in the l_2 scheduling policy, such selection

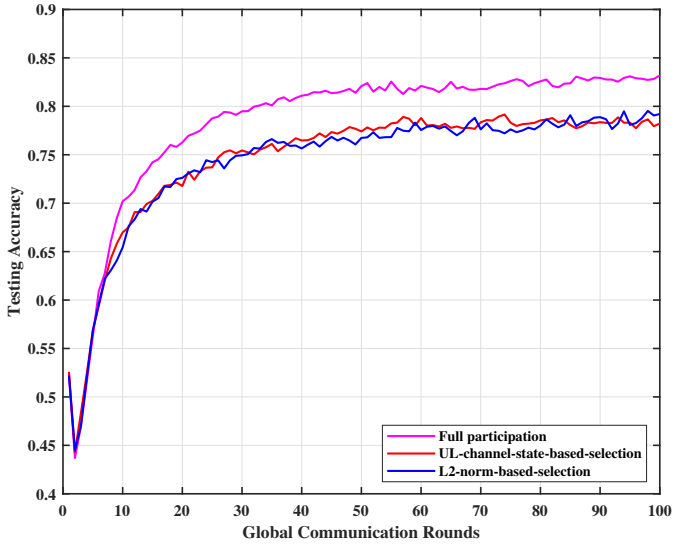


Fig. 7. Testing accuracy at local user side on MNIST i.i.d. dataset when full participation, UL-channel-state-based scheduling policy and l_2-norm-based scheduling policy are considered. Values of DL-SNR and UL-SNR are 8 dB and -8 dB, respectively; that is, $P_{DL,0}/\sigma^2 = 8$ dB, $P_{UL,0}/\sigma^2 = -8$ dB.

is based on the criteria explained in Algorithm 1. Thus, UL channel state and l_2 scheduling policy can reduce the communication cost at the expense of testing performance reduction, that is, local models converge slower with less participations.

C. Testing accuracy for varying SNR and learning rate

In this subsection, we vary learning rate η_t, UL-SNR and DL-SNR through the training process. The value of η_t is decreased according to Theorem 3, i.e., $\eta_t = \frac{\beta_0}{t+\beta_1}$ with $\beta_0 = 2, \beta_1 = 100$, and both DL and UL SNRs are increased as $P_{UL,r} = P_{UL,0}(1 - \exp(-\gamma r))$ and $P_{DL,r} =$

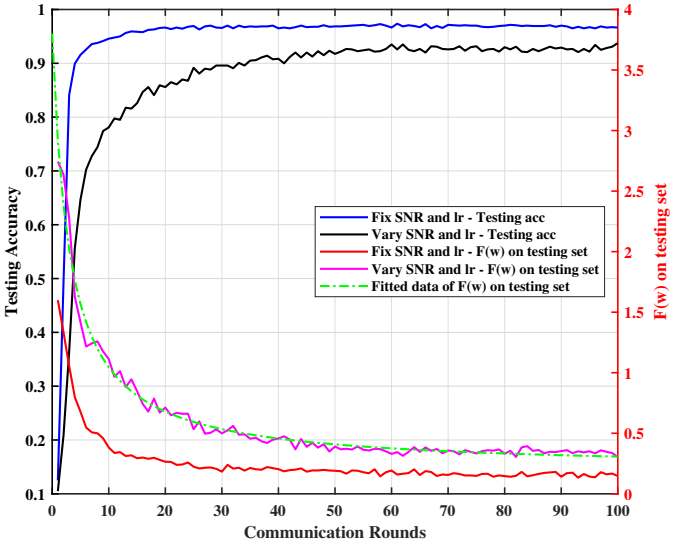


Fig. 8. Progresses of testing accuracy (left-axis) and testing loss $F(\mathbf{w})$ (right axis) at the sever side to compare the effects of varying and fixed SNRs through training process. The fitted line for varying SNR shows $\mathcal{O}\left(T^{-1}\sigma_z^{-2}\right)$ behavior as proved in Theorem 3.

TABLE II
AVERAGED TESTING ACCURACY OF LOCAL USER MODELS AT END OF TRAINING WHEN I.I.D. TRAINING DATASET IS USED.

Schemes \ Quantization bits		16	14	12	10
Vary S_r	Vary P_r	95.52%	94.28%	95.44%	92.82%
	Fix P_r	95.32%	96.30%	95.90%	92.60%
Fix S_r	Vary P_r	94.14%	93.68%	93.62%	91.70%
	Fix P_r	94.86%	94.26%	94.66%	91.28%

TABLE III
AVERAGED TESTING ACCURACY OF LOCAL USER MODELS AT END OF TRAINING WHEN NON-I.I.D. TRAINING DATASET IS USED.

Schemes \ Quantization bits		16	14	12	10
Vary S_r	Vary P_r	79.01%	77.36%	76.82%	70.16%
	Fix P_r	82.42%	82.80%	81.32%	75.50%
Fix S_r	Vary P_r	73.29%	73.06%	71.56%	60.04%
	Fix P_r	80.27%	80.12%	79.00%	70.36%

$P_{DL,0}(1-\exp(-\gamma r))$ with $\gamma = 0.2$, $P_{UL,0}/\sigma^2 = -7$ dB, $P_{DL,0}/\sigma^2 = 7$ dB. Number of antennas at BS $N_{BS} = 128$.

Fig. 8 plots the testing accuracy on i.i.d. MNIST (left-axis) and the values of the loss-function $F(\mathbf{w})$ for the testing set (right-axis) versus the communication rounds. By comparing the performance for varying SNRs with the scenario where SNRs are fixed, it can be observed that in the former, models converge slower than that in the latter. However, varying SNR provides the advantage of energy efficiency improvement. The value of γ can be adjusted to trade-off energy efficiency with the convergence rate. To show the loss function decreases by $\mathcal{O}\left(T^{-1}\sigma_z^{-2}\right)$, we fit the curve with the function $a + \frac{b}{T+c}$, which verifies the theorem 3.

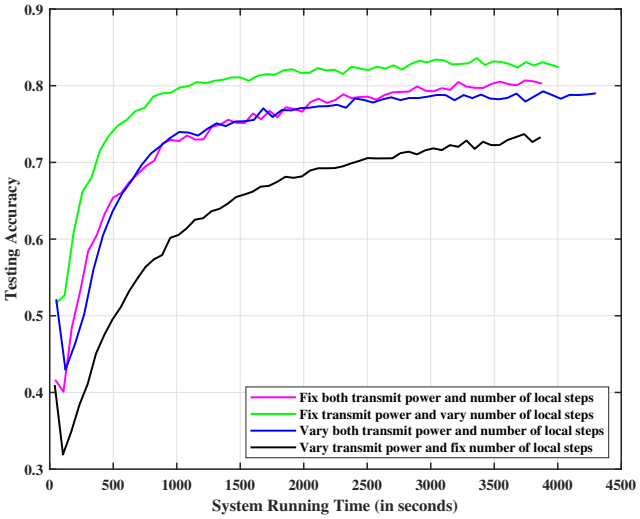


Fig. 9. Testing accuracy on non-i.i.d. MNIST dataset versus system running time under different system schemes. Number of quantization bit is 16. $N_{BS} = 128$. Non-i.i.d. factor = 70%. Values of DL-SNR and UL-SNR are 7 dB and -7 dB, respectively; that is, $P_{DL,0}/\sigma^2 = 7$ dB, $P_{UL,0}/\sigma^2 = -7$ dB.

D. Simulations on non-i.i.d. MNIST

1) Effects of varying local steps and transmit power:

In Fig. 9, we plot the averaged local testing accuracy versus the system running time as training progresses under different schemes with non-i.i.d. training data. Number of communication rounds is set to be $T = 60$ and non-i.i.d. factor is 70%. We average the accuracy of each scenario for 10 instances. It can be observed that under same system running time, varying local step scheme with fixed transmit power (green curve) gives best model performance at the end of training. While magenta and blue curves show similar testing accuracy through training process, the difference is that the scheme corresponding to blue curve incurs less communication cost since both transmit power and number of local steps are varied in this scheme. Black curve represents the scheme where transmit power is varied and number of local steps is fixed, it can be witnessed that it returns worst result among all schemes under consideration, implying that the learning algorithm requires more iterations to converge, based on the transmit power variation model.

2) Effects of varying quantization bits:

In Table II and Table III, we summarize the averaged testing accuracy of local user models at the end of training under different system schemes and quantization bits conditions on both i.i.d. and non-i.i.d. training datasets. For the same simulation setup in the previous subsubsection with $N_{BS} = 128$, $P_{DL,0}/\sigma^2 = 7$ dB, $P_{UL,0}/\sigma^2 = -7$ dB and 70% non-i.i.d. factor, we can observe that in both i.i.d. and non-i.i.d. cases, as number of quantization bits is increased, better testing performance can be seen at the expense of more communication cost. Since the difference introduced by varying number of local steps and transmit power can hardly be witnessed in the i.i.d. case, in the non-i.i.d. scenario, this difference stands out and becomes more obvious as shown in Fig. 9 and Table III. Fixing transmit power through whole training process gives better performance than "varying power" policy, this is due to

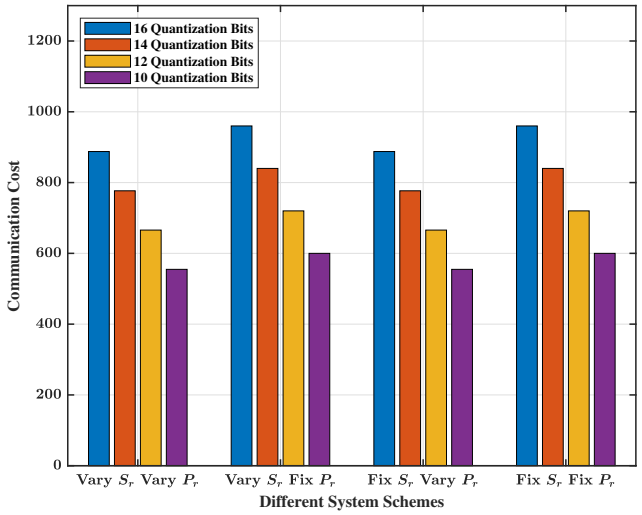


Fig. 10. Visualization of communication cost under different systems schemes. Non-i.i.d. MNIST dataset is considered.

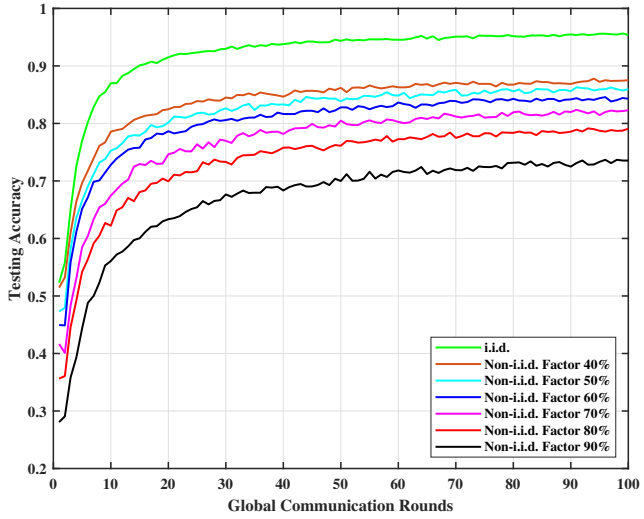


Fig. 11. Testing accuracy on MNIST dataset with different non-i.i.d. factors. Number of quantization bit is 16. $N_{BS} = 128$. Values of DL-SNR and UL-SNR are 7 dB and -7 dB, respectively; that is, $P_{DL,0}/\sigma^2 = 7$ dB, $P_{UL,0}/\sigma^2 = -7$ dB.

the fact that on average, higher SNR values introduce fewer transmission errors.

Another interesting finding is that, in non-i.i.d. case, varying number of local steps not only accelerates the convergence but also leads to apparently better testing results at the end of training compared with its "fix local steps" counterpart. For instance, if we compare first and third rows of Table III (where transmit power is varied), we can observe 5% and 10% better testing performance in the former case when the constrains for quantization bits are 12 and 10, respectively. Similar trend can also be witnessed when second and fourth rows are compared in Table III, where the transmit power is fixed.

3) Comparison of communication cost:

In Fig. 10, communication cost incurred by different system schemes in Table III is visualized and compared by utilizing a bar chart. Since all local users' models have same architecture,

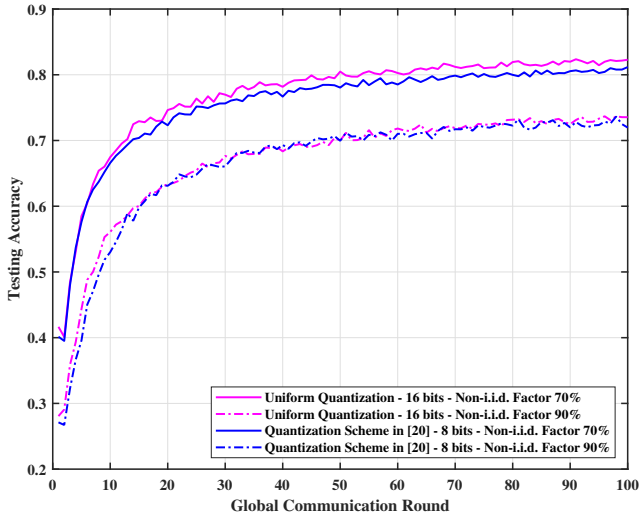


Fig. 12. Testing accuracy on non-i.i.d. MNIST dataset when two quantization schemes are used. $N_{BS} = 128$. Values of DL-SNR and UL-SNR are 7 dB and -7 dB, respectively; that is, $P_{DL,0}/\sigma^2 = 7$ dB, $P_{UL,0}/\sigma^2 = -7$ dB.

when plotting the bar chart, we do not consider "# of model parameters" as shown in Eqn. (16), leaving the remaining three factors under considerations:

1) Communication rounds: As stated in subsubsection IV-D1, number of communication rounds T is set to be 60 for all system schemes.
2) Normalized transmit power: For $T = 60$ and $\gamma = \frac{1}{5}$, $\mathcal{P} = 0.9247$. When "fix P_r" strategy is used, normalized power is 1.
3) Quantization bits: We use the same values $\mathcal{B}$ as the ones we present in Table. III, that is, $16, 14, 12$ and 10 bits, respectively.

As shown in Fig. 10, for a given system scheme, decreasing number of quantization bits leads to less communication cost. While "vary S_r vary P_r" and "fix S_r vary P_r" yield same communication cost, the former scheme is able to provide better performance in terms of testing accuracy (as shown in Fig. 9). At the same time, even though system schemes with "fix P_r" characteristic incur more communication cost, higher testing accuracy can be witnessed at the end of training, compared with their "vary P_r" counterparts.

4) Different values of non-i.i.d. factors:

In Fig. 11, we vary the non-i.i.d. factor and results are averaged over 10 instances. With 16 quantization bits, $N_{BS} = 128$, $P_{DL,0}/\sigma^2 = 7$ dB, and $P_{UL,0}/\sigma^2 = -7$ dB, the performance degradation can be witnessed as the value of non-i.i.d. factor is increased.

5) Different quantization schemes:

In this subsubsection, we consider another quantization scheme based on stochastic rounding [20]. The corresponding result for 8 bits quantization is plotted in Fig. 12 and compared with the uniform quantization (16 bits). We utilize two different values (70% and 90%) of non-i.i.d. factor and it can be seen from Fig. 12 that for both quantization schemes, the system performances in terms of testing accuracy are approximately same; that is, saving half of communication cost.

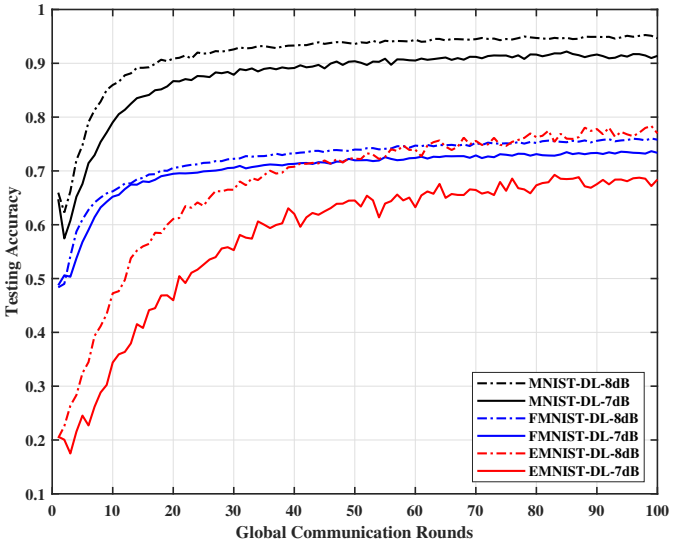


Fig. 13. Testing accuracy at local user side on MNIST, EMNIST and FMNIST datasets. $N_{BS} = 64$ and number of local epochs is 7. Value of UL-SNR is fixed to -3 dB; that is, $P_{UL,0}/\sigma^2 = -3$ dB.

E. Simulations on EMNIST and FMNIST dataset

In Fig. 13, testing accuracies are plotted for i.i.d. MNIST, FMNIST and EMNIST datasets with different DL-SNRs and a fixed UL SNR -3 dB. Number of local epochs is set to be 7. Compared with MNIST and FMNIST, training history for EMNIST dataset has more fluctuations, since EMNIST dataset has more classes (26).

Similarly, FMNIST is also more challenging to classify as compared with MNIST, despite the same number of classes (10) in these two datasets. This is due to the fact that when dimension reduction techniques, e.g., uniform manifold approximation and projection (UMAP) [56] and minimum-distortion embedding (MDE) [57] are used, FMNIST samples with different labels are more densely distributed than that of MNIST, leading to overlapped classes and reduced testing accuracy.

V. Conclusion

In this paper, we have integrated the federated learning framework with a multi-user massive MIMO 6G network, considering the imperfections of both UL and DL links. At the massive MIMO BS, ZF and MMSE schemes have been used to null the co-channel interference. The tighter convergence rate bound has been derived in a simpler manner incorporating these imperfections, than the one in literature. This $\mathcal{O}\left(T^{-1}\sigma_z^{-2}\right)$ bound also has revealed that DL-SNR and UL-SNR can be increased as training goes on, motivated by which we improve the communication-efficiency of the system by adjusting local training steps, transmit power and quantization bits. Moreover, it has been shown that compared with the reliability of UL links, quality of DL links plays a more significant role in this system. Simulation results for images recognition task with MNIST, EMNIST and FMNIST datasets have verified the theoretical insights for fixed and varying SNRs. The testing accuracy has been compared with the scenario with the perfect links. Value of loss function has also been plotted to verify the $\mathcal{O}\left(T^{-1}\sigma_z^{-2}\right)$ performance.

Future work includes expanding the proposed framework to multi-cell scenario, conducting corresponding convergence analysis and investigating new quantization schemes for the proposed FL framework.

Appendices

A. Proof of Lemma 1

Proof: The left hand side of lemma can be simplified as

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{w}}_{t+1}-\mathbf{w}^*\right\|_2^2 \\
&=\begin{cases}\mathbb{E}\left\|\bar{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|_2^2, & \text{if } t+1\notin\mathcal{T},\\ \mathbb{E}\left\|\bar{\mathbf{v}}_{t+1}+\sum_{k=1}^N\alpha_k\bar{\mathbf{z}}_t^k-\mathbf{w}^*\right\|_2^2, & \text{if } t+1\in\mathcal{T},\end{cases}\\
&=\mathbb{E}\left\|\bar{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|_2^2+\mathbb{E}\left\|\sum_{k=1}^N\alpha_k\bar{\mathbf{z}}_t^k\right\|_2^2\delta_{t+1\in\mathcal{T}}\\
&=\mathbb{E}\left\|\bar{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|_2^2+W\sigma_z^2\delta_{t+1\in\mathcal{T}},
\end{aligned}$$

where the variance of the effective noise $\sum_{k=1}^N\alpha_k\bar{\mathbf{z}}_t^k$ is obtained as

$$\begin{aligned}
W\sigma_z^2&=\mathbb{E}\left\|\sum_{k=1}^N\alpha_k\bar{\mathbf{z}}_t^k\right\|_2^2\\
&=\mathbb{E}\left\|\sum_{k=1}^N\alpha_k\left(\bar{\mathbf{z}}_{t,UL}+\bar{\mathbf{z}}_{t,DL}^k\right)\right\|_2^2\\
&=\sum_{k,j=1}^N\alpha_k\alpha_j tr\mathbb{E}\left(\bar{\mathbf{z}}_{t,UL}+\bar{\mathbf{z}}_{t,DL}^k\right)\left(\bar{\mathbf{z}}_{t,UL}^H+\bar{\mathbf{z}}_{t,DL}^{jH}\right)\\
&=\sum_{k,j=1}^N\alpha_k\alpha_j tr\mathbb{E}\left\{\bar{\mathbf{z}}_{t,UL}\bar{\mathbf{z}}_{t,UL}^H+\bar{\mathbf{z}}_{t,DL}^k\bar{\mathbf{z}}_{t,DL}^{jH}\right\}\\
&=\frac{\|\boldsymbol{\alpha}\|_2^2\sigma_w^2}{2}\frac{\sigma^2N_D}{P_{UL}N_{BS}}tr\mathbf{I}_W+\sum_k^N\alpha_k^2tr\mathbb{E}\left\{\bar{\mathbf{z}}_{t,DL}^k\bar{\mathbf{z}}_{t,DL}^{kH}\right\}\\
&=\frac{\|\boldsymbol{\alpha}\|_2^2\sigma_w^2}{2}\frac{\sigma^2N_D}{P_{UL}N_{BS}}W+\frac{\|\boldsymbol{\alpha}\|_2^2\hat{\sigma}_w^2}{2}\frac{\sigma^2}{P_{DL}}W\\
&=\frac{\|\boldsymbol{\alpha}\|_2^2\sigma_w^2}{2}\frac{\sigma^2}{P_{DL}}\left[\frac{P_{DL}N_D}{P_{UL}N_{BS}}+\frac{\hat{\sigma}_w^2}{\sigma_w^2}\right]W\\
&=\frac{\|\boldsymbol{\alpha}\|_2^2\sigma_w^2}{2}\frac{\sigma^2}{P_{DL}}\left[\frac{P_{DL}N_D}{P_{UL}N_{BS}}+\|\boldsymbol{\alpha}\|_2^2\left(1+\frac{\sigma^2N_D}{P_{UL}N_{BS}}\right)\right]W\\
&=\frac{\|\boldsymbol{\alpha}\|_2^2\sigma_w^2}{2}\frac{\sigma^2}{P_{DL}}\left[\left(\frac{P_{DL}}{\sigma^2}+\|\boldsymbol{\alpha}\|_2^2\right)\frac{\sigma^2N_D}{P_{UL}N_{BS}}+\|\boldsymbol{\alpha}\|_2^2\right]W.
\end{aligned}$$

■

B. Proof of Lemma 2

Proof: The left hand side of lemma can be simplified as

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{v}}_{t+1}-\mathbf{w}^*\right\|_2^2 \\
&=\mathbb{E}\left\|\bar{\mathbf{w}}_t-\eta_t\mathbf{g}_t-\mathbf{w}^*\pm\eta_t\nabla F(\bar{\mathbf{w}}_t)\right\|_2^2 \\
&\overset{(a)}{=}\mathbb{E}\left\|\bar{\mathbf{w}}_t-\eta_t\nabla F(\bar{\mathbf{w}}_t)-\mathbf{w}^*\right\|_2^2+\eta_t^2\mathbb{E}\left\|\mathbf{g}_t-\nabla F(\bar{\mathbf{w}}_t)\right\|_2^2 \\
&=\mathbb{E}\left\|\bar{\mathbf{w}}_t-\mathbf{w}^*\right\|_2^2+\eta_t^2\mathbb{E}\left\|\mathbf{g}_t-\nabla F(\bar{\mathbf{w}}_t)\right\|_2^2+\eta_t^2\left\|\nabla F(\bar{\mathbf{w}}_t)\right\|_2^2 \\
&\quad-2\eta_t\mathbb{E}\left(\bar{\mathbf{w}}_t-\mathbf{w}^*\right)^T\nabla F(\bar{\mathbf{w}}_t) \\
&\overset{(b)}{\leq}\mathbb{E}\left\|\bar{\mathbf{w}}_t-\mathbf{w}^*\right\|_2^2+\eta_t^2\mathbb{E}\left\|\mathbf{g}_t-\nabla F(\bar{\mathbf{w}}_t)\right\|_2^2+\eta_t^2\left\|\nabla F(\bar{\mathbf{w}}_t)\right\|_2^2 \\
&+2\eta_t\mathbb{E}\left[F(\mathbf{w}^*)-F(\bar{\mathbf{w}}_t)-\frac{\mu}{2}\|\bar{\mathbf{w}}_t-\mathbf{w}^*\|_2^2\right] \\
&\overset{(c)}{\leq}\left(1-\eta_t\mu\right)\mathbb{E}\left\|\bar{\mathbf{w}}_t-\mathbf{w}^*\right\|_2^2+\eta_t^2\sum_k\alpha_k^2\left(\sigma_k^2+G^2\right) \\
&+2\eta_t\mathbb{E}\left[F(\mathbf{w}^*)-F(\bar{\mathbf{w}}_t)\right] \\
&\overset{(d)}{\leq}\left(1-\eta_t\mu\right)\mathbb{E}\left\|\bar{\mathbf{w}}_t-\mathbf{w}^*\right\|_2^2+\eta_t^2\sum_k\alpha_k^2\left(\sigma_k^2+G^2\right),
\end{aligned}$$

where in (a), $\mathbf{g}_t=\sum_{k=1}^N\alpha_k\nabla F_k(\mathbf{w}_t^k,\xi_t^k)$ with $\mathbb{E}\left\{\mathbf{g}_t-\bar{\mathbf{g}}_t\right\}=\mathbf{0}$ is used; (b) is obtained from μ-strong convexity, i.e., $-2(\mathbf{w}-\mathbf{v})^T\nabla F(\mathbf{w})\leq 2F(\mathbf{v})-2F(\mathbf{w})-\mu\|\mathbf{v}-\mathbf{w}\|_2^2$; in (c), the result from Assumption 3 and 4 is used; in (d), the fact $F(\mathbf{w}^*)\leq F(\bar{\mathbf{w}}_t)$ is used. ∎

C. Proof of Theorem 3

Proof: For the L-smooth function, we write

$$\begin{aligned}
&\mathbb{E}\left\{F\left(\bar{\mathbf{w}}_{t+1}\right)\right\}-F^* && (21)\\
&\leq\mathbb{E}\left\{\left(\bar{\mathbf{w}}_{t+1}-\mathbf{w}^*\right)^T\nabla F_k(\mathbf{w}^*)\right\}+\frac{L}{2}\Delta_{t+1} && (22)\\
&\overset{(a)}{=}\frac{L}{2}\Delta_{t+1} && (23)\\
&\overset{(b)}{\leq}\frac{L}{2}\cdot\frac{\nu}{t+1+\beta_1}, && (24)
\end{aligned}$$

where in (a), zero mean result $\mathbb{E}\left\{\bar{\mathbf{w}}_t-\mathbf{w}^*\right\}=\mathbf{0}$ is used, and $\Delta_t=\mathbb{E}\|\bar{\mathbf{w}}_t-\mathbf{w}^*\|_2^2$; the inequality, $\Delta_t\leq\frac{\nu}{t+\beta_1}$ in (b) is proved via induction as follows. Let $C=\sum_k\alpha_k^2\sigma_k^2+G^2+W\frac{\sigma_{\tilde{s}}^2}{\eta_t^2}\delta_{t+1\in\mathcal{S}}$, $\nu\geq\frac{\beta_0^2C}{\beta_0\mu-1}$, $\eta_t=\frac{\beta_0}{t+\beta_1}$, where $\beta_0>\frac{1}{\mu}$, $\beta_1>0$. Then, definition of ν guarantees that the inequality in (b) holds for $t=0$, assume the conclusion holds for some $t>0$, by using the results of Lemma 1 and Lemma 2, we write

$$\begin{aligned}
\Delta_{t+1}&\leq\left(1-\eta_t\mu\right)\Delta_t+\eta_t^2C && (25)\\
&\leq\frac{\nu}{t+\beta_1}-\frac{\beta_0\mu\nu}{\left(t+\beta_1\right)^2}+\frac{\beta_0^2C}{\left(t+\beta_1\right)^2} && (26)\\
&=\frac{\nu(t+\beta_1)-\nu}{\left(t+\beta_1\right)^2}+\frac{\nu-\beta_0\mu\nu+\beta_0^2C}{\left(t+\beta_1\right)^2} && (27)\\
&\leq\frac{\nu}{t+1+\beta_1}+0. && (28)
\end{aligned}$$

∎

D. Proof of Corollary 4

Proof: When learning rate η is a constant, the model also converges, however, with fixed gap from the optimal values. From Eqn. (25) in the proof of Theorem 3 above, we replace $\eta_t=\eta$ and prove $\Delta_t\leq\frac{\eta C}{\mu}$ via induction as follows

$$\begin{aligned}
\Delta_{t+1}&\leq\left(1-\eta\mu\right)\Delta_t+\eta^2C && (29)\\
&\leq\left(1-\eta\mu\right)\frac{\eta C}{\mu}+\eta^2C=\frac{\eta C}{\mu}. && (30)
\end{aligned}$$

Then, from the inequality in Eqn. (23), it follows that using the L-smooth function, we have

$$\mathbb{E}\left\{F\left(\bar{\mathbf{w}}_{t+1}\right)\right\}-F^*\leq\frac{L}{2}\Delta_{t+1}\leq\frac{L}{2}\frac{\eta C}{\mu}.$$

∎

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 700–714, 2015.

[2] D. A. Forsyth, and J. Ponce, *Computer vision: a modern approach*. Pearson, 2012.

[3] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. SAAnchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," In *Proc. of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, Apr. 2017.

[6] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, *et al.*, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, Nov. 2021.

[7] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, and H. Ludwig, "Hybridalpha: An efficient approach for privacy-preserving federated learning," In *Proc. of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 13–23, Nov. 2019.

[8] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H.Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," In *Proc. of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 1–11, Nov. 2019.

[9] M. Hao, H. Li, G. Xu, S. Liu, and H. Yang, "Towards efficient and privacy-preserving federated deep learning," In *Proc. of the IEEE International Conference on Communications*, pp. 1–6, May. 2019.

[10] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," In *Proc. of the International Conference on Learning Representations,*, Apr. 2020.

[11] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6g: ai empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[12] T. Li, *et al.*, "Federated learning: challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May. 2020.

[13] Y. Liu, J. Peng, J. Kang, A. M. Iliyasu, D. Niyato, and A. A. A. El-Latif, "A secure federated learning framework for 5g networks," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 24–31, Aug. 2020.

[14] M. Isaksson, and K. Norrman, "Secure federated learning in 5g mobile networks," In *Proc. of the IEEE Global Communications Conference*, Dec. 2020.

[15] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6g communications: challenges, methods, and future directions," *China Communications*, vol. 17, no. 9, pp. 105–118, Sep. 2020.

[16] Z. Yang, M. Chen, K. K. Wong, *et al.*, "Federated learning for 6g: applications, challenges, and opportunities", *Engineering*, vol. 8, pp. 33–41, Jan. 2022.

[17] H. Yang, A. Alphones, Z. Xiong, *et al.*, "Artificial-intelligence-enabled intelligent 6g networks," *IEEE Network*, vol. 34, no. 6, pp. 272–280, Nov. 2020.

[18] Y. Xiao, G. Shi, and M. Krunz, "Towards ubiquitous ai in 6g with federated learning," *IEEE Communications Magazine*, 2020.

[19] J. Konecný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: strategies for improving communication efficiency," In *Proc. of the NIPS Workshop on Private Multi-Party Machine Learning*, pp. 1–10, Dec. 2016.

[20] S. Zheng, C. Shen, and X. Chen, "Design and analysis of uplink and downlink communications for federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2150–2167, Jul. 2021.

[21] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Federated learning with quantization constraints," In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8851–8855, May. 2020.

[22] A. F. Aji, and K. Heafield, "Sparse communication for distributed gradient descent," In *Proc. of the Empirical Methods in Natural Language Processing*, Sep. 2017.

[23] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: ternary gradients to reduce communication in distributed deep learning," In *Proc. of the Advances in Neural Information Processing Systems*, Dec. 2017.

[24] F. Sattler, S. Wiedemann, K. R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[25] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A Joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[26] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: optimization model design and analysis," In *Proc. of the IEEE Conference on Computer Communications*, pp. 1387–1395, Apr. 2019.

[27] C. Dinh, N. H. Tran, M. N. Nguyen, C. S. Hong, W. Bao, A. Y. Zomaya, and V. Gramoli, "Federated learning over wireless networks: convergence analysis and resource allocation," *IEEE/ACM Transactions on Networking*, vol. 29, no. 1, pp. 398–409, Feb. 2021.

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

[29] X. Mo, and J. Xu, "Energy-efficient federated edge learning with joint communication and computation design," *Journal of Communications and Information Networks*, vol. 6, no. 2, pp. 110–124, Jun. 2021.

[30] Y. Liu, Y. Zhu, and J. Q. James, "Resource-constrained federated learning with heterogeneous data: formulation and analysis," *IEEE Transactions on Network Science and Engineering*, Nov. 2021.

[31] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8743–8747, May. 2020.

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

[33] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," In *Proc. of the IEEE International Conference on Communications*, pp. 1–6, Jun. 2020.

[34] H. Yang, J. Zhao, Z. Xiong, *et al.*, "Privacy-preserving federated learning for uav-enabled networks: learning-based joint scheduling and resource management," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3144–3159, Oct. 2021.

[35] W. Y. B. Lim, S. Garg, Z. Xiong, Y. Zhang, D. Niyato, C. Leung, and C. Miao, "Uav-assisted communication efficient federated learning in the era of the artificial intelligence of things," *IEEE Network*, vol. 35, no. 5, pp. 188–195, Sep. 2021.

[36] X. Wei, and C. Shen, "Federated learning over noisy channels: convergence analysis and design examples," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1253–1268, Jun. 2022.

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

[38] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[39] M. M. Amiri, and D. Gündüz, "Machine learning at the wireless edge: distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, Mar. 2020.

[40] M. M. Amiri, and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, May. 2020.

[41] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: a communication-efficient federated learning method with periodic averaging and quantization," In *Proc. of International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031, Aug. 2020.

[42] Y. S. Jeon, M. M. Amiri, J. Li, *et al.*, "A compressive sensing approach for federated learning over massive mimo communication systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1990–2004, Nov. 2020.

[43] M. Duan, D. Liu, X. Chen, *et al.*, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 59–71, Jan. 2021.

[44] A. Papazafeiropoulos, H. Ngo and T. Ratnarajah, "Performance of massive mimo uplink with zero-forcing receivers under delayed channels", *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3158–3169, Apr. 2017.

[45] A. Papazafeiropoulos and T. Ratnarajah, "Deterministic equivalent performance analysis of time-varying massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 10, pp. 5795–5809, Oct. 2015.

[46] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Soteria: provable defense against privacy leakage in federated learning from representation perspective," In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9311–9319, Jun. 2021.

[47] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd with faster convergence and less communication: demystifying why model averaging works for deep learning," In *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 5693–5700, Jul. 2019.

[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," In *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[49] G. Cohen, S. Afshar, J. Tapson, *et al.*, "Emnist: extending mnist to handwritten letters," In *Proc. of the International Joint Conference on Neural Networks*, May. 2017.

[50] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*. [Online]. Available: https://arxiv.org/abs/1708.07747.

[51] Z. Qin, GY. Li and H. Ye, "Federated learning and wireless communications," *IEEE Wireless Communications*, vol. 28, no. 5, pp. 134–140, Oct. 2021.

[52] H. Yang, Z. Xiong, J. Zhao, *et al.*, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

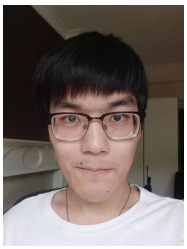
[53] M. A. Albreem, M. Juntti, S. Shahabuddin, "Massive mimo detection techniques: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3109–3132, Aug. 2019.

[54] R. Chataut and R. Akl, "Massive mimo systems for 5g and beyond networks-overview, recent trends, challenges, and future research direction," *Sensors*, vol. 20, no. 10, May. 2020.

[55] E. G. Larsson, O. Edfors, F. Tufvesson, *et al.*, "Massive mimo for next generation wireless systems," *IEEE communications magazine*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[56] L. McInnes, J. Healy, and J. Melville, "Umap: uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*. [Online]. Available: https://arxiv.org/abs/1802.03426.

[57] A. Agrawal, A. Ali and S. Boyd, "Minimum-distortion embedding," *Foundations and Trends in Machine Learning*, vol. 14, no. 3, pp. 211–378, Sep. 2021.



Yuchen Mu received the B.Eng. degree in electronic engineering (telecommunications) from De Montfort University, Leicester, U.K., in 2019, and the M.Sc. degree in signal processing and communications from The University of Edinburgh, Edinburgh, U.K., in 2020. He was awarded Best Final Year Student (electrical and electronic engineering) by De Montfort University and Top Student (M.Sc. in signal processing and communications) by The University of Edinburgh, in 2019 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Institute for Digital Communications. His research interests include federated learning, deep learning, cellular communications, and massive MIMO.

Navneet Garg (Member, IEEE) received the B.Tech. degree in electronics and communication engineering, and the M.Tech. degree in digital communications. He has completed the Ph.D. degree in June 2018 from the department of electrical engineering at the IIT Kanpur, India. From July 2018-Jan. 2019, he visited The University of Edinburgh, UK. From Feb. 2019-2020, he is employed as a research associate in Heriot-Watt University, UK. Since Feb. 2020, he is working as a research associate in The University of Edinburgh, UK. His main research interests include wireless communications, signal processing, optimization, and machine learning.

Tharmalingam Ratnarajah (Senior Member, IEEE) is currently with the Institute for Digital Communications, The University of Edinburgh, UK, as a Professor in Digital Communications and Signal Processing. His research interests include signal processing and information theoretic aspects of beyond 5G wireless networks, full-duplex radio, mmWave communications, random matrices theory, interference alignment, statistical and array signal processing and quantum information theory. He has published over 400 publications in these areas and holds four U.S. patents. He has supervised 16 PhD students and 21 postdoctoral research fellows and raised $11+ million USD of research funding.