Model Compression via Pattern Shared Sparsification in Analog Federated Learning under Communication Constraints

Jin-Hyun Ahn, Mehdi Bennis, *Fellow, IEEE*, and Joonhyuk Kang, *Member, IEEE*

***Abstract*—Recently, it has been shown that analog transmission based federated learning enables more efficient usage of communication resources compared to the conventional digital transmission. In this paper, we propose an effective model compression strategy enabling analog FL under constrained communication bandwidth. To this end, the proposed approach is based on pattern shared sparsification by setting the same sparsification pattern of parameter vectors uploaded by edge devices, as opposed to each edge device independently applying sparsification. In particular, we propose specific schemes for determining the sparsification pattern and characterize the convergence of analog FL leveraging these proposed sparsification strategies, by deriving a closed-form upper bound of convergence rate and residual error. The closed-form expression allows to capture the effect of communication bandwidth and power budget to the performance of analog FL. In terms of convergence analysis, the model parameter obtained with the proposed schemes is proven to converge to the optimum of model parameter. Numerical results show that leveraging the proposed pattern shared sparsification consistently improves the performance of analog FL in various settings of system parameters. The improvement in performance is more significant under scarce communication bandwidth and limited transmit power budget.**

***Index Terms*—Distributed learning, federated learning, over-the-air computation, compression, local gradient accumulation.**

I. Introduction

As the computational capabilities of edge devices continue to improve, distributed machine learning (ML) has become one of the promising technologies to alleviate the heavy cost of collecting training datasets in centralized ML. Centralized ML is impractical when the huge amount of datasets are collected through wireless communication due to the limited communication resources such as power, time, and bandwidth. *Federated Learning* (FL) was recently proposed as a privacy-preserving distributed ML scheme based on first-order distributed optimization updates and periodic exchanges of model parameter information between devices (or clients) and a parameter server (PS) [1]–[5]. In numerous previous works including [6]–[8], the superiority of FL has been validated through numerical results and convergence analysis in IID and non-IID datasets [9].

Nonetheless, to enable FL, tackling challenges related to the large number of model is still an open problem. According to [10], the 50-layer ResNet [11] has 26 millions model parameters, and approximately 138 millions for VGGNet [12], while the available channel bandwidth is relatively small due to constrained communication resources, e.g., 1 LTE frame of 5MHz bandwidth and 10ms duration can carry 6000 complex symbols. Thus, the main performance bottleneck in FL is in the uplink communication, that is from devices to PS [13]. To alleviate this problem, there is a rich literature to reduce the communication requirements in FL. Besides the periodic communication [3], [14], [15], one well-known approach is *lossy compression* through quantization or sparsification of parameter vectors. The quantization technique reduces the amount of information to be uploaded by quantizing each entry of the parameter vector with low-bit precision [16]–[19] and the sparsification techniques selectively sends the entries of the parameter vector [20]–[23]. The most recent works propose to accelerate the convergence of FL by applying the gradient tracking, momentum, and sketching [24]–[27].

Recently, there has been a great effort to investigate methods to improve the performance of FL taking into account wireless communication. Considering the physical and network layer aspects, [28]–[33] have established a new approach for improving the performance of FL through scheduling policy and allocation of communication resources such as power, time, and frequency. On the other hand, wireless communication links bring new opportunities due to the superposition property of wireless transmission. As demonstrated in [34]–[36], the superposition property can be leveraged to enable over-the-air computation (AirComp) of aggregation of signals sent from multiple devices without decoding each signal separately.

The idea of AirComp has been leveraged to improve the efficiency of communication resources for model uploading in uplink. This is done by directly estimating the average parameters from the superposition of signals transmitted by multiple devices over a multi-access channel, when each device adopts analog communication. In fact, it is not necessary for devices to transmit their parameters through orthogonal

J.-H. Ahn is with MGH/BWH Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston MA 02114 USA (jahn13@mgh.harvard.edu). J. Kang is with the School of Electrical Engineering (EE), Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Korea (jhkang@kaist.edu). M. Bennis is with the Centre for Wireless Communications, University of Oulu, Oulu 90014, Finland (e-mail: mehdi.bennis@oulu.fi). This work was performed when the first author was at KAIST.

Manuscript received September 13, 2021; revised March 30, 2022; accepted June 20, 2022. Date of publication xx xx, 2022. “‘This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-0-01787) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation)”

0000–0000/00$00.00 © 2021 IEEE

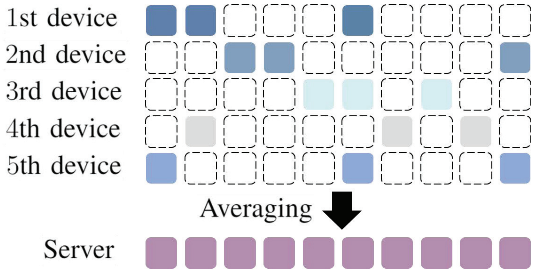


Fig. 1: Example of *curse of primal averaging*.

links for separate decoding since FL requires the sum of parameters. As a result, analog FL with AirComp outperforms FL with conventional digital communication [10], [37], [38]. The literature including [38]–[41] has addressed the issues for effective implementation of analog FL such as user selection, power allocation, and transmission protocol. In particular, model aggregation in broadband analog transmission has been studied and the latency gain obtained by AirComp has been investigated, compared to the conventional orthogonal access in [38].

Despite the efficient usage of communication bandwidth under analog FL, it is still needed to compress the parameters due to the available bandwidth. In the digital domain, lossy compression at any required level is possible [37]. However, quantization cannot be applied and the characteristic of analog transmission makes the compression different from the compression in the digital domain. While the issue is essential, there are few related works comparing digital domain as discussed next. This constitutes the focus of this work.

A. Related Work

To the authors' best knowledge, only references [10], [37] have addressed model compression enabling analog FL under constrained communication bandwidth. Starting from these works, we propose a new sparsification strategy for analog FL, which shows considerable improvement in performance in terms of test accuracy.

In [10], [37], local top-k sparsification [42] is considered to compress the gradient estimates and local gradient accumulation [20] is applied to accelerate the performance. Each device independently sparsifies the gradient estimate by choosing the k largest entries with absolute value. Then it accumulates the unchosen entries to keep track of gradient estimates losses due to compression and adds the accumulation to the gradient estimate obtained at the next iteration. The local top-k sparsification and local gradient accumulation are commonly used in the literature since they have empirically shown great performance [8], [17], [20], [43]. However, when the local top-k sparsification is considered in analog transmission, only the top-k values can be sent without the information of indices due to the characteristic of analog transmission. To solve this issue, a novel projection scheme inspired by compressed sensing (CS) is proposed in [10], [37].

In the scheme, the devices apply the local top-k sparsification and simultaneously transmit the compressed vectors which are projected to the available communication bandwidth. Then, the PS reconstructs the summation of gradient estimates by adopting the recovery algorithms in CS. For the reliable reconstruction of the gradient estimate, the sparsity level of the original vector must meet the specific constraint [44]. However, it suffers *curse of primal averaging*, also discussed in [45], where the desired summation of gradient estimate may be rendered dense. Although each gradient estimate is sparse, the averaging step in FL will definitely yield the dense solution as illustrated in Fig. 1.

Motivated by the drawback of local top-k sparsification in analog FL, we propose a novel pattern shared sparsification (PSS) by setting the same sparsification pattern of gradient estimates among edge devices. Unlike the local top-k sparsification, it is guaranteed that summation of gradient estimates can be reconstructed without any recovery error since the indices to be sparsified and reconstructed are previously shared. Furthermore, there is no loss of communication bandwidth when PSS is considered since the gradient estimates are compressed to the same dimension with communication bandwidth. In this work, we propose three wireless implementations of analog FL based on the proposed PSS.

B. Our Contributions

- For effective compression in analog FL under the constrained communication bandwidth, we propose a cooperative compression strategy based on exploiting unbiased gradient estimates while local top-k sparsification yields biased gradient estimates. In particular, three schemes for determining the shared pattern in PSS are respectively established and validated numerically. Through numerical results with various settings of system parameters, the PSS schemes consistently outperforms the local top-k sparsification and the performance improvement is more significant under scarce communication bandwidth and limited power budget.
- We prove the convergence of analog FL with PSS by deriving the closed-form upper bound of the convergence rate and residual error. Leveraging the closed-form expression, we describe the effect of communication bandwidth and power budget to the performance of analog FL with an explicit formula. The effect of communication bandwidth and power budget obtained in the analysis coincides with the results observed in the numerical results. Moreover, the gain coming from gradient accumulation is derived as an explicit formula in the comparison between the convergence rate with gradient accumulation and convergence rate without gradient accumulation.

II. Problem Definition

A. System Set-Up

We consider a federated learning system comprising a parameter server and M edge devices. Let $\boldsymbol{\theta} \in \mathbb{R}^D$ denote the shared network parameter to be optimized. The local loss $F_m(\boldsymbol{\theta})$ at the m-th device is $F_m(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_m|}\sum_{\boldsymbol{u}\in\mathcal{D}_m} f(\boldsymbol{\theta}, \boldsymbol{u})$, where $\mathcal{D}_m$ is the dataset allocated at the m-th device and $f(\cdot)$

is the loss function determined by the network model. The global loss $F(\boldsymbol{\theta})$ is defined as

$$F(\boldsymbol{\theta}) = \frac{1}{\left|\bigcup_{m=1}^{M} \mathcal{D}_m\right|} \sum_{\boldsymbol{u} \in \bigcup_{m=1}^{M} \mathcal{D}_m} f(\boldsymbol{\theta}, \boldsymbol{u}) \overset{(a)}{=} \frac{1}{M} \sum_{m=1}^{M} F_m(\boldsymbol{\theta}), \tag{1}$$

where (a) holds if $|\mathcal{D}_m|$ are equal for all m. The goal in FL is to learn the optimized parameter vector $\boldsymbol{\theta}^*$ minimizing (1), namely

$$\boldsymbol{\theta}^* = \operatorname{argmin} F(\boldsymbol{\theta}). \tag{2}$$

To obtain $\boldsymbol{\theta}^*$, the model parameter is updated through iterative stochastic gradient descent (SGD) allowing the parallel computation of gradients at the edge devices. The parameter vector $\boldsymbol{\theta}_t$ at the t-th iteration is updated according to

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{g}_m(\boldsymbol{\theta}_t), \tag{3}$$

where η is the learning rate and $\boldsymbol{g}_m(\boldsymbol{\theta}_t) \in \mathbb{R}^D$ is the stochastic gradient of the shared model parameter $\boldsymbol{\theta}_t$ computed at the m-th device as $\boldsymbol{g}_m(\boldsymbol{\theta}_t) = \frac{1}{|\mathcal{B}_m|} \sum_{\boldsymbol{u} \in \mathcal{B}_m} \nabla f(\boldsymbol{\theta}_t, \boldsymbol{u})$, using the available subdataset $\mathcal{B}_m$. Following the iterative SGD with parallel computation at the device, referred to as distributed SGD (DSGD) or FedSGD, we establish the baseline FL with error-free links in Algorithm 1.

Algorithm 1: Federated Learning (FL)

```
for each iteration $t = 0, \dots, T$
    for each device $m = 1, \dots, M$
        download from PS the global parameter vector $\boldsymbol{\theta}_t$
        set initial learning model with $\boldsymbol{\theta}_t$
        do SGD update using the available subdataset $\mathcal{B}_m$
        upload the locally obtained gradient $\boldsymbol{g}_m(\boldsymbol{\theta}_t)$
    compute (3) at PS
```

B. Communication Model

During the uplink communication of each global iteration, all edge devices share a fading uplink multiple-access channel

$$\mathbf{y} = \sum_{m=1}^{M} h_m \mathbf{x}_m + \mathbf{z}, \tag{4}$$

where h_m is the quasi-static flat fading channel from the m-th device to the AP; $\mathbf{x}_m$ is the $I \times 1$ signal transmitted by the m-th device; and $\mathbf{z}$ is $I \times 1$ noise vector with IID $\mathcal{CN}(0,1)$ entries. The number of channel uses, I, is the allowed number of channel uses during the given time resource with the allocated frequency bandwidth at each global iteration, *i.e.*, the maximum number of transmitted symbols at each global iteration, considering both time resource and frequency bandwidth. Edge devices have power constraints per channel use at each global iteration, $\|\mathbf{x}_m\|_2^2/I \le P$ for $m = 1, \dots, M$. Downlink broadcast communication from the AP to the devices is assumed to be error free in order to focus on the effect of the more challenging multi-access uplink channel.

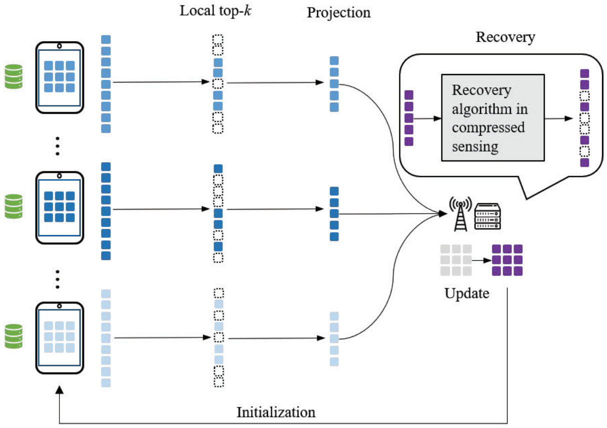


Fig. 2: Analog federated learning enabled by local top-k sparsification in [10], [37].

III. Analog Federated Learning

In this section, we establish the wireless implementations of the analog FL schemes. First of all, we review the analog-DSGD (A-DSGD) [10], [37], where each device applies a local top-k sparsification, as a benchmark scheme. We refer to A-DSGD as "local top-k" to highlight the difference of local top-k sparsification with the sparsification schemes proposed in this paper. We also propose three wireless implementations of analog FL based on PSS for the sparsification of gradient estimates being uploaded. Each implementation is developed based on a respective scheme for determining the shared sparsity pattern. The proposed schemes are referred to as *PSS with random selection*, *PS-guided PSS*, and *device-guided PSS*.

Under AirComp in analog FL, all devices simultaneously transmit their information multiplied by a scalar factor to the AP in an uncoded manner. The PS decodes the desired sum directly after scaling the received signal (4). Therefore, the scalar factor should be determined for establishing each scheme of analog FL. Different types of methods for determining the scalar factor have been studied in the literature including [46], [47], namely full-power transmission, channel inversion, and optimal power control for AirComp. In this paper, channel inversion is considered to enable the alignment of scalar factors among devices as [38], but extensions are conceptually straightforward. This requires the m-th device to have knowledge of the channel to the AP, h_m and the AP to have all channels, h_m, $m = 1, \dots, M$.

A. Compression with Local Top-k Sparsification

In order to enable dimensionality reduction as in CS, a pseudo-random matrix $\boldsymbol{A} \in \mathbb{R}^{2I \times D}$ with IID entries $\mathcal{N}(0, 1/2I)$ is generated and shared between the PS and the devices before the start of the global iterations. At the start of the t-th global iteration, the m-th device downloads the global parameter vector $\boldsymbol{\theta}_t$ and sets it as the initial model of local training, as described in Algorithm 1. After the local computation update, the gradient vector $\boldsymbol{g}_m(\boldsymbol{\theta}_t) \in \mathbb{R}^D$ is obtained.

To compress the gradient vector with reduced dimension, $\boldsymbol{g}_m(\boldsymbol{\theta}_t)$ is sparsified to $\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$ defined as

$$\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t) = \mathrm{Top}_k\left(\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t\right), \tag{5}$$

where $k(< 2I)$ is a parameter determining the sparsification ratio, Top_k denotes top-k sparsification, and $\boldsymbol{\Delta}_m^t$ is an accumulated gradient which is updated as

$$\boldsymbol{\Delta}_m^{t+1} = \boldsymbol{\Delta}_m^t + \boldsymbol{g}_m(\boldsymbol{\theta}_t) - \boldsymbol{g}_m^{sp}. \tag{6}$$

After computing $\widehat{\boldsymbol{g}}_m(\boldsymbol{\theta}_t) = \boldsymbol{A}\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t) \in \mathbb{R}^{2I}$, the m-th device transmits $\frac{\gamma}{h_m}\boldsymbol{v}_m^t \in \mathbb{C}^I$ whose i-th entry, $v_{m,i}^t$, is defined as

$$\mathrm{Re}(v_{m,i}^t) = \widehat{g}_{m,2i-1}(\boldsymbol{\theta}_t), \quad \mathrm{Im}(v_{m,i}^t) = \widehat{g}_{m,2i}(\boldsymbol{\theta}_t) \tag{7}$$

by encoding two different values of $\widehat{\boldsymbol{g}}_m(\boldsymbol{\theta}_t)$, i.e., $\widehat{g}_{m,2i-1}(\boldsymbol{\theta}_t)$ and $\widehat{g}_{m,2i}(\boldsymbol{\theta}_t)$, in the in-phase and quadrature components. The scaling factor γ_t is determined by

$$\gamma_t = \min_m \frac{|h_m^t|\sqrt{PI}}{\|\boldsymbol{v}_m^t\|_2}, \tag{8}$$

ensuring the alignment of scalar factors under power constraints.

The PS receives $\boldsymbol{y}_t \in \mathbb{C}^I$ in (4), which is the the aggregation of $\boldsymbol{x}_m^t = \frac{\gamma_t}{h_m^t}\boldsymbol{v}_m^t$, $m = 1, \ldots, M$ through the fading uplink multiple-access channels. From the noisy observation, the PS estimates $\sum_{m=1}^M \widehat{\boldsymbol{g}}_m(\boldsymbol{\theta}_t)$ by $\boldsymbol{u}_t \in \mathbb{R}^{2I}$ whose i-th entry is defined as follows:

$$u_{t,i} = \begin{cases} \dfrac{\mathrm{Re}(y_{t,j})}{\gamma_t}, & \text{if } i = 2j-1, \\ \dfrac{\mathrm{Im}(y_{t,j})}{\gamma_t}, & \text{if } i = 2j, \end{cases} \tag{9}$$

for $j = 1, \ldots, I$ by scaling $\boldsymbol{y}_t$ and decoding to the real values for further recovery, where $y_{t,j}$ is the j-th element of $\boldsymbol{y}_t$. For recovery from compressed vector, the approximate message passing (AMP) [44] is applied to $\boldsymbol{u}_t$ and the reconstructed vector, $\mathrm{AMP}_{\boldsymbol{A}}(\boldsymbol{u}_t)$, is used to update the model parameter $\boldsymbol{\theta}_t$ to the new model parameter, $\boldsymbol{\theta}_{t+1}$ as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{M}\mathrm{AMP}_{\boldsymbol{A}}(\boldsymbol{u}). \tag{10}$$

See Fig. 2 for an illustrative example.

B. Pattern Shared Sparsification (PSS)

In the local top-k sparsification, the m-th device sparsifies the gradient estimate as in (5). For the recovery without any indices of non-zero entries, the sparsified gradient estimates are projected onto the communication bandwidth, multiplied by a previously shared pseudo-random measurement matrix $\boldsymbol{A}$, as in CS. Then, all the devices simultaneously transmit the compressed vectors to the PS based on analog communication. After receiving the aggregation of compressed vectors, the PS acquires

$$\mathrm{Rec}\left(\sum_{m=1}^M \boldsymbol{A}\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)\right) = \mathrm{Rec}\left(\boldsymbol{A}\sum_{m=1}^M \boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)\right), \tag{11}$$

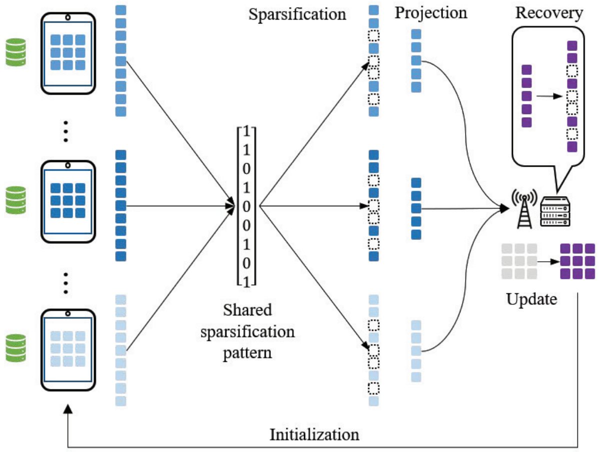


Fig. 3: Analog FL enabled by pattern shared sparsification.

where Rec denotes recovery algorithms in CS such as Lasso, orthogonal matching pursuit (OMP) and AMP as used in [10], [37].

We note that the sum of sparsified gradient estimates, $\sum_{m=1}^M \boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$, is obtained as observed in (11). It can lead to the *non-sparsity* of the original vector for recovery since $\sum_{m=1}^M \boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$ is not guaranteed to be sparse even though $\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$ is sparse for each m. This *curse of primal averaging* degrades the recovery performance and thereby the performance of analog FL. This is due to the irregular sparsity pattern among gradient estimates, which has been considered as a main drawback of local top-k sparsification [26], [43]. Furthermore, as a coupling problem with non-sparsity, the recovery algorithm is too bandwidth-consuming since the non-zero entries should be much less than the communication bandwidth for reliable recovery.

To address this issue, each device applies the previously shared pattern of sparsification and direct projection of the non-zero elements to the available communication bandwidth, when the proposed PSS is considered. Therefore, it is guaranteed that $\sum_{m=1}^M \boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$ can be reconstructed without any recovery error since the indices to be sparsified and reconstructed are previously shared. Furthermore, there is no loss of communication bandwidth since the gradient estimates are compressed to the same dimension with communication bandwidth. The key point for developing PSS is to seek a shared sparsification pattern at each global iteration. For this goal, we propose three different schemes for sparsification which are *PSS with random selection*, *PS-guided PSS*, and *device-guided PSS*.

1) PSS with random selection: As a first reference scheme, we propose PSS with random selection where PS randomly selects the entries of gradient vector to be non-zero and share the determined sparsity pattern with devices at the start of each global iteration. Using this scheme, the uploaded entries of gradient vectors might not contain significant elements having large absolute values. However, this scheme enables the same sparsity pattern among the gradient estimates of devices and the most efficient usage of communication bandwidth. In fact, when PSS with random selection is applied, each device can

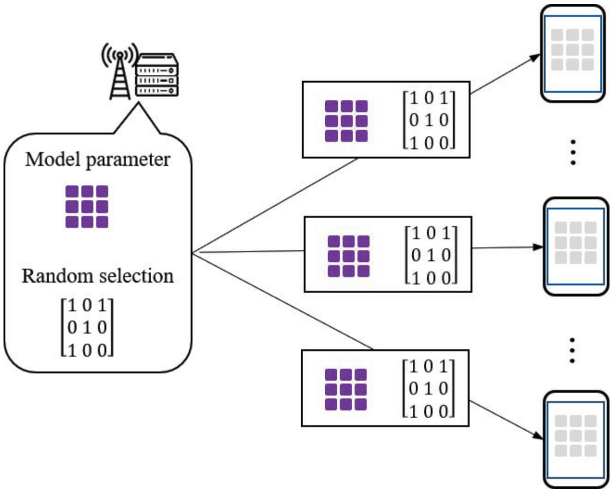


Fig. 4: Download in PSS with random selection.

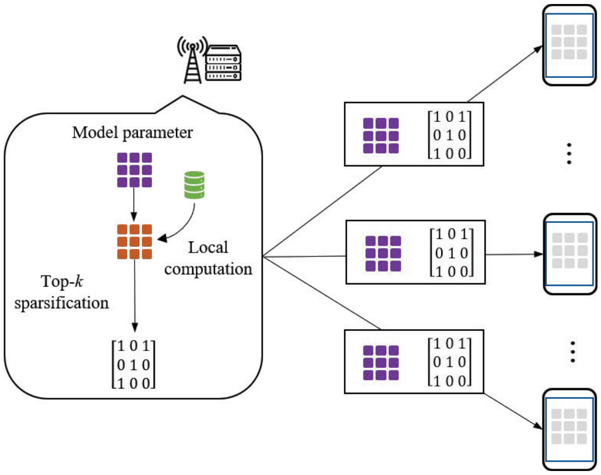


Fig. 5: Download in PS-guided PSS.

upload $2I$ elements of gradient vector which is the maximum number using the available number of channel uses, I, without the need to compress to a deeper level for the recovery in CS.

At the start of the t-th global iteration, the m-th device downloads the model parameter $\boldsymbol{\theta}_t \in \mathbb{R}^D$ and the sparsity pattern which is defined by a diagonal matrix $\boldsymbol{S}_t \in \{0,1\}^{D\times D}$, as shown in Fig. 4. The diagonal entries $s_{t,d}$, $1 \le d \le D$ of $\boldsymbol{S}_t$ are 0 or 1 when $\sum_{d=1}^{D} s_{t,d} = 2I$, defining the random selection of sparsified entries. The m-th device sets the initial model of local training with $\boldsymbol{\theta}_t$ and computes the gradient estimate $\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right) \in \mathbb{R}^D$ as described in Algorithm 1. For compression, $\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right)$ is sparsified to $\boldsymbol{g}_m^{sp}\left(\boldsymbol{\theta}_t\right)$,

$$\boldsymbol{g}_m^{sp}\left(\boldsymbol{\theta}_t\right) = \boldsymbol{S}_t\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right) + \boldsymbol{\Delta}_m^t\right), \tag{12}$$

where $\boldsymbol{\Delta}_m^t$ is the accumulated gradient updated as (6). The projection $\widehat{\boldsymbol{g}}_m\left(\boldsymbol{\theta}_t\right)$ of $\boldsymbol{g}_m^{sp}\left(\boldsymbol{\theta}_t\right)$ is obtained by eliminating all zeros in $\boldsymbol{g}_m^{sp}\left(\boldsymbol{\theta}_t\right)$ to have only non-zero entries in the reduced dimension, and $\widehat{\boldsymbol{g}}_m\left(\boldsymbol{\theta}_t\right)$ can be expressed as follows:

$$\widehat{\boldsymbol{g}}_m\left(\boldsymbol{\theta}_t\right) = \widehat{\boldsymbol{S}}_t\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right) + \boldsymbol{\Delta}_m^t\right), \tag{13}$$

where $\widehat{\boldsymbol{S}}_t$ is the $2I \times D$ matrix which is obtained by removing d-th row of $\boldsymbol{S}_t$ when $s_{t,d} = 0$.

Then, the m-th device transmits $\frac{\gamma_t}{h_m^t}\boldsymbol{v}_m^t \in \mathbb{C}^I$ whose i-th entry, $v_{m,i}^t$, is defined as (7) and the scaling factor γ_t is determined by (8), ensuring the alignment of scalar factors under power constraints. After the PS receives $\boldsymbol{y}_t \in \mathbb{C}^I$ in (4), the PS estimates $\sum_{m=1}^{M} \widehat{\boldsymbol{g}}_m\left(\boldsymbol{\theta}_t\right)$ with $\boldsymbol{u}_t \in \mathbb{R}^{2I}$ whose i-th entry is defined as (9). For recovery from the compression vector, each element of $\boldsymbol{u}_t$ is matched to the entry of gradient vector based on the known sparsity pattern $\boldsymbol{S}_t$ which can be expressed as $\widehat{\boldsymbol{S}}_t^T\boldsymbol{u}_t$, where $\widehat{\boldsymbol{S}}_t^T$ denotes transpose of $\widehat{\boldsymbol{S}}_t$. Finally, the reconstructed vector, $\widehat{\boldsymbol{S}}_t^T\boldsymbol{u}_t$, is used to update the model parameter $\boldsymbol{\theta}_t$ to the new model parameter, $\boldsymbol{\theta}_{t+1}$ as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{M}\widehat{\boldsymbol{S}}_t^T\boldsymbol{u}_t. \tag{14}$$

2) PS-guided PSS: Here, the PS determines the sparsity pattern by selecting the significant entries of the gradient vector computed at the PS. We assume that the PS has its local dataset which can be a public dataset as in many literature including [48]. This scheme enables the same sparsity pattern among the gradient estimates of devices and the most efficient usage of communication bandwidth similar to PSS with random selection. Furthermore, the obtained sparsity pattern is more relevant to the original gradient estimates than the sparsity pattern obtained by random selection since the pattern is guided by the gradient estimate computed at the PS with its local dataset.

At the start of t-th global iteration, the PS computes the gradient estimate, $\boldsymbol{g}_{PS}\left(\boldsymbol{\theta}_t\right) \in \mathbb{R}^D$, with its local dataset, denoted by $\mathcal{D}_{PS}$. Then the sparsity pattern is obtained by selecting the $2I$ entries having large absolute values in $\boldsymbol{g}_{PS}\left(\boldsymbol{\theta}_t\right)$ which can be expressed as a diagonal matrix $\boldsymbol{S}_t \in \{0,1\}^{D\times D}$ whose diagonal entries indicate the selected $2I$ entries. Accordingly, the vector $\boldsymbol{s}_t$ of the diagonal entries in $\boldsymbol{S}_t$ is $\mathrm{Top}_{2I}\left(\boldsymbol{g}_{PS}\left(\boldsymbol{\theta}_t\right)\right)$ whose non-zero elements are converted to ones. After each device downloads $\boldsymbol{\theta}_t$ and $\boldsymbol{S}_t$, the same procedure used in the PSS with random selection is followed. The downloading process in PS-guided PSS is illustrated in Fig. 5.

3) Device-guided PSS: Here, one of the devices determines the sparsity pattern by selecting the significant entries of gradient vector and shares such pattern with the PS and other devices with the help of PS. The sparsity pattern guides the compression at the device level, maintaining the same sparsity pattern among gradient estimates. In the implementation of this scheme, the first $\rho I\,(0 < \rho < 1)$ channels are used for uploading the determined sparsity pattern from a device to the PS in the digital domain and the remaining $(1-\rho)\,I$ channels are used for uploading the compressed gradient vectors from the other devices to the PS in the analog fashion. To this end, we assume that the devices and PS are allowed to communicate in a hybrid fashion enabling the selective use of analog and digital communication. We additionally assume that the communication in the downlink is error free.

At the start of t-th global iteration, devices download the model parameter $\boldsymbol{\theta}_t \in \mathbb{R}^D$ and set the initial model of local training with it, as described in Algorithm 1. After the local computation for update, the $\widehat{m}$-th device which has the

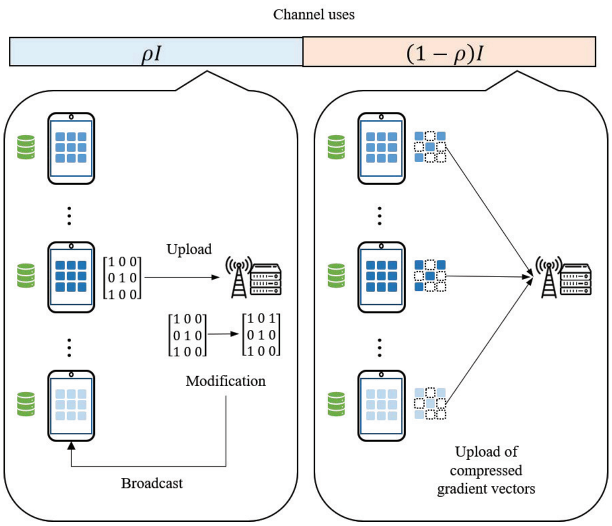

Channel uses
ρI
$(1-\rho)I$
Upload
$\begin{bmatrix}1&0&0\\0&1&0\\1&0&0\end{bmatrix}$
$\begin{bmatrix}1&0&0\\0&1&0\\1&0&0\end{bmatrix} \rightarrow \begin{bmatrix}1&0&1\\0&1&0\\1&0&0\end{bmatrix}$
Modification
Broadcast
Upload of compressed gradient vectors

Fig. 6: Device-guided PSS.

best channel gain to the PS, namely, $\widehat{m} = \underset{m}{\operatorname{argmax}} |h_m|$, determines the sparsity pattern by selecting the q entries of $\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t$ having large absolute values. The sparsity pattern is expressed as a diagonal matrix $\boldsymbol{S}_t \in \{0,1\}^{D\times D}$ whose diagonal entries indicate the selected q entries. Accordingly, the vector of diagonal entries, $\boldsymbol{s}_t$ is $\operatorname{Top}_q(\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t)$ whose non-zero elements are converted to ones, where q is chosen as follows. When ρI is the number of channel uses allowed for uploading $\boldsymbol{S}_t$ in a digital domain, the number of bits that can be transmitted from $\widehat{m}$-th device is upper bounded by

$$B_t = \rho I \log_2 \left(1 + \frac{|h_{\widehat{m}}|^2 P}{\rho}\right), \tag{15}$$

while the number of bits for the transmission of $\boldsymbol{S}_t$ is

$$b_t = \left\lceil \log_2 \binom{D}{k} \right\rceil. \tag{16}$$

Therefore, q is chosen as the largest integer satisfying $b_t \leq B_t$. After receiving $\boldsymbol{S}_t$ for the ρI channels, PS modifies the sparsity pattern $\boldsymbol{S}_t$ by adding the randomly selected $2(1-\rho)I - q$ entries in the remaining $D-q$ entries of sparsity pattern. In fact, the q entries for sparsification are previously determined and guided by the $\widehat{m}$-th device and the remaining number of entries that can be uploaded by each device is $2(1-\rho)I - q$.

For the sparsification of gradient estimates, each $\boldsymbol{g}_m(\boldsymbol{\theta}_t)$ is sparsified to $\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$ as in (12) and projected to $\widehat{\boldsymbol{g}}_m(\boldsymbol{\theta}_t)$ as in (13). Then, the m-th device transmits $\frac{\gamma_t}{h_m^t}\boldsymbol{v}_m^t \in \mathbb{C}^{(1-\rho)I}$ whose i-th entry, $v_{m,i}^t$, is defined as (7). The scaling factor γ_t is determined by

$$\gamma_t = \min_{m \neq \widehat{m}} \frac{|h_m^t| \sqrt{PI}}{\|\boldsymbol{v}_m^t\|_2}, \tag{17}$$

modified from (8) since the $\widehat{m}$-th device transmits $\boldsymbol{S}_t$ for the initial $(1-\rho)I$ channel uses. After the PS receives $\boldsymbol{y}_t \in \mathbb{C}^{(1-\rho)I}$ in (4), the same procedure used in the PSS with random selection is followed. The overall process of device-guided PSS is illustrated in Fig. 6.

IV. Convergence Analysis

In this section, we provide the convergence analysis of analog FL leveraging PSS for gradient sparsification. We consider the system set-up and communication model introduced in Section II and establish the convergence analysis of PSS with random selection, as a reference scheme. Since the other PSS schemes outperform the PSS with random selection as illustrated in the Section V, this analysis can be applied to the other PSS schemes. The expectation of the squared l_2 norm between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}^*$, $\mathrm{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2^2\right]$ is the main measure for the convergence of $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}^*$. We first introduce preliminaries and then provide the convergence result and related analysis.

A. Preliminaries

We assume the available subdataset $\mathcal{B}_m$ of the m-th device is equal to the allocated dataset $\mathcal{D}_m$ for simplicity of the convergence analysis. Thus, the gradient vector $\boldsymbol{g}_m(\boldsymbol{\theta}_t)$ is $\nabla F_m(\boldsymbol{\theta}_t)$ and the update of $\boldsymbol{\theta}_t$ in (3) is expressed as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{1}{M} \sum_{m=1}^{M} \nabla F_m(\boldsymbol{\theta}_t). \tag{18}$$

We introduce the following assumptions related to the characteristic of loss, which facilitate the convergence analysis:

Assumption 1. The local losses $F_1, \ldots, F_M$ are μ-strongly convex; $\forall\ \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$

$$F_m(\boldsymbol{a}) - F_m(\boldsymbol{b}) \geq \langle \boldsymbol{a} - \boldsymbol{b}, \nabla F_m(\boldsymbol{b}) \rangle + \frac{\mu}{2} \|\boldsymbol{a} - \boldsymbol{b}\|_2^2, \tag{19}$$

Assumption 2. The expectation of squared l_2 norm of local losses $F_1, \ldots, F_M$ are bounded by G; $\forall\ \boldsymbol{a} \in \mathbb{R}^d$

$$\mathrm{E}\left[\|\nabla F_m(\boldsymbol{a})\|_2^2\right] \leq G^2, \tag{20}$$

Assumption 3. The local losses $F_1, \ldots, F_M$ are L-smooth; $\forall$ $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$

$$F_m(\boldsymbol{a}) - F_m(\boldsymbol{b}) \leq \langle \boldsymbol{a} - \boldsymbol{b}, \nabla F_m(\boldsymbol{b}) \rangle + \frac{L}{2} \|\boldsymbol{a} - \boldsymbol{b}\|_2^2, \tag{21}$$

Next, we introduced a lemma which was also used in [49].

Lemma 1. *When random selection is considered for determining the sparsification pattern, the sparsity pattern of uploaded gradients follows a uniform distribution, thus*

$$\mathrm{E}\left[\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)\right] = \frac{2I}{D} \mathrm{E}\left[\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t\right], \tag{22}$$

$$\mathrm{E}\left[\|\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)\|_2^2\right] = \frac{2I}{D} \mathrm{E}\left[\|\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t\|_2^2\right], \tag{23}$$

for each $m = 1, \ldots, M$ and $t = 0, \ldots, T$.

Proof: In (22), we have

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)\right] &\overset{(a)}{=} \mathrm{E}\left[\boldsymbol{S}_t\left(\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t\right)\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[\boldsymbol{S}_t\left(\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t\right) \mid \boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t\right]\right] \\ &\overset{(b)}{=} \frac{2I}{D} \mathrm{E}\left[\boldsymbol{g}_m(\boldsymbol{\theta}_t) + \boldsymbol{\Delta}_m^t\right], \end{aligned} \tag{24}$$

where (a) holds following the definition of $\boldsymbol{g}_m^{sp}$ in (12) and (b) is due to $\mathrm{E}[\boldsymbol{S}_t] = \frac{2I}{D}\boldsymbol{I}_D$ for a $D \times D$ identity matrix $\boldsymbol{I}_D$.

In (23), we have

$$\mathrm{E}\left[\left\|\boldsymbol{g}_m^{sp}\left(\boldsymbol{\theta}_t\right)\right\|_2^2\right] \tag{25}$$

$$\begin{aligned}
&= \mathrm{E}\left[\left\|\boldsymbol{S}_t\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right)+\boldsymbol{\Delta}_m^t\right)\right\|_2^2\right]\\
&= \mathrm{E}\left[\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right)+\boldsymbol{\Delta}_m^t\right)^T\boldsymbol{S}_t^T\boldsymbol{S}_t\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right)+\boldsymbol{\Delta}_m^t\right)\right]\\
&= \mathrm{E}\left[\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right)+\boldsymbol{\Delta}_m^t\right)^T\boldsymbol{S}_t\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right)+\boldsymbol{\Delta}_m^t\right)\right]\\
&= \frac{2I}{D}\mathrm{E}\left[\left\|\boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right)+\boldsymbol{\Delta}_m^t\right\|_2^2\right].
\end{aligned}\tag{26}$$

■

B. Convergence Result

We prove the convergence of analog FL with local gradient accumulation when PSS with random selection is considered for determining the sparsity pattern of uploaded gradient vectors. We consider two cases for the loss, which are (i) Assumption 1+Assumption 2 and (ii) Assumption 1+Assumption 2+Assumption 3. For each case, we provide corresponding convergence result and the relevant analysis.

Theorem 1. *When the case (i) is considered, we have*

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right] \leq A\left(t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{D\eta G^2}{2I\mu}\left(1+\frac{\Psi}{P}\right), \tag{27}$$

for $\eta<\frac{1}{\mu}$, where $\Psi=\mathrm{E}\left[\frac{1}{|h_m|^2}\right]$, $A(t)=\prod_{t'=1}^{t}a(t')$, and

$$a(t')=1-\eta\mu\left(1-\left(1-\frac{2I}{D}\right)^{t'}\right). \tag{28}$$

Proof: See Appendix A. ■

Theorem 2. *When the case (ii) is considered, we have*

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right] \leq B(t)\mathrm{E}\left[\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{D\eta G^2\Psi}{2I\left(\mu-L^2\eta\right)P}, \tag{29}$$

for $\eta<\frac{\mu}{L^2}$, where $\Psi=\mathrm{E}\left[\frac{1}{|h_m|^2}\right]$, $B(t)=\prod_{t'=1}^{t}b(t')$, and

$$b(t')=1-\eta\left(\mu-L^2\eta\right)\left(1-\left(1-\frac{2I}{D}\right)^{t}\right). \tag{30}$$

Proof: See Appendix B. ■

Corollary 1. *When the case (i) is considered, $\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\xrightarrow{t}\epsilon$ for*

$$\eta<\min\left(\frac{1}{\mu},\frac{2PI\mu}{\left(P+\Psi\right)D\eta G^2}\epsilon\right). \tag{31}$$

Proof: If $\eta<\frac{1}{\mu}$, we have $0<a(t')<1$ and

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\xrightarrow{t}\frac{D\eta G^2}{2I\mu}\left(1+\frac{\Psi}{P}\right), \tag{32}$$

from Theorem 1, where $\frac{D\eta G^2}{2I\mu}\left(1+\frac{\Psi}{P}\right)<\epsilon$. ■

Corollary 2. *When the case (ii) is considered, $\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\xrightarrow{t}\epsilon$ for*

$$\eta<\min\left(\frac{\mu}{L^2},\frac{1}{L^2}\left(\frac{1}{\mu}-\frac{DG^2\Psi}{DG^2\Psi+2L^2IP\epsilon}\right)\right). \tag{33}$$

Proof: If $\eta<\frac{\mu}{L^2}$, we have $0<b(t')<1$ and

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\xrightarrow{t}\frac{D\eta G^2\Psi}{2I\left(\mu-L^2\eta\right)P}, \tag{34}$$

from Theorem 2, where $\frac{D\eta G^2\Psi}{2I(\mu-L^2\eta)P}<\epsilon$. ■

Remark 1. *In* (27) *and* (29), $\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]$ *is upper bounded by*

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\leq C_1\left(t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\right\|_2^2\right]+C_2, \tag{35}$$

where $C_1\left(t\right)$ is the convergence rate and C_2 is a residual error term. First of all, we observe that the number of channel uses, I, the only communication factor determining $C_1\left(t\right)$ and $C_1\left(t\right)$ decreases as I increases. In particular, $C_1\left(t\right)$ is $1-\eta\mu$ or $1-\eta\left(\mu-L^2\eta\right)$ when $2I=D$. The residual error term, C_2, decreases as I, M, and P increases. Specifically, in (29), *C_2 is caused only by the noise in the uplink communication since the additional term of C_2 in* (27) *is derived to be contained in $C_1(t)$ because of the additionally assumed L-smoothness of the loss function. Thus C_2 in* (29) *decreases to 0 as P increases to ∞.*

Remark 2. *In Theorem 1 and 2, we consider analog FL with local gradient accumulation when PSS with random selection is applied. While the convergence rates are obtained as* (28) *and* (30), *the convergence rate without local gradient accumulation can be readily obtained as $\widehat{A}(t)=\prod_{t'=1}^{t}\widehat{a}(t')$ and $\widehat{B}(t)=\prod_{t'=1}^{t}\widehat{b}(t')$, where*

$$\widehat{a}(t')=1-\eta\mu\frac{2I}{D} \tag{36}$$

$$\widehat{b}(t')=1-\eta\left(\mu-L^2\eta\right)\frac{2I}{D}, \tag{37}$$

while

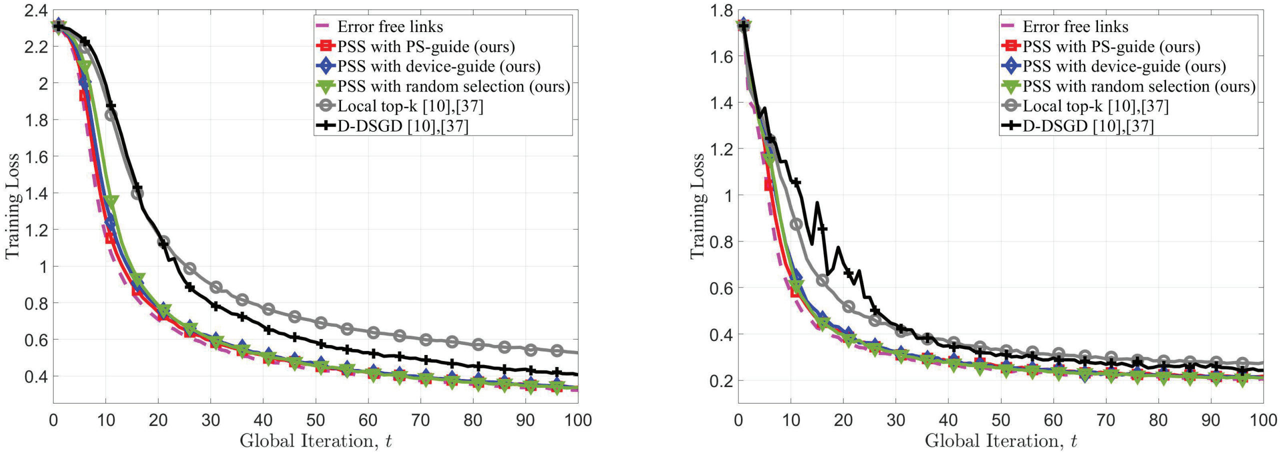
$$a(t')=1-\eta\mu\left(1-\left(1-\frac{2I}{D}\right)^{t'}\right) \tag{38}$$

$$b(t')=1-\eta\left(\mu-L^2\eta\right)\left(1-\left(1-\frac{2I}{D}\right)^{t}\right). \tag{39}$$

in (28) *and* (30). *Without local gradient accumulation, the convergence rate decreases by a constant factor $\widehat{a}(t')$ or $\widehat{b}(t')$. However, with local gradient accumulation, the factor of convergence rate, $a(t')$ and $b(t')$, decreases to $1-\eta\mu$ and $1-\eta\left(\mu-L^2\eta\right)$ starting from $\widehat{a}(t')$ and $\widehat{b}(t')$, respectively.*

V. Numerical Results

To validate the effectiveness of the proposed approach, we consider two examples, reflecting both IID and non-IID data allocation, where each device runs a five-layer Convolutional Neural Network (CNN) that consists of two convolutional layers, one max-pooling layer, and two fully-connected layers for MNIST image classification. For the IID setting, we



(a) Training loss in the case of IID data allocation.

(b) Training loss in the case of non-IID data allocation.

Fig. 7: Training loss of analog FL schemes and D-DSGD considering $I = 5000$, $P = 30$, $M = 50$.

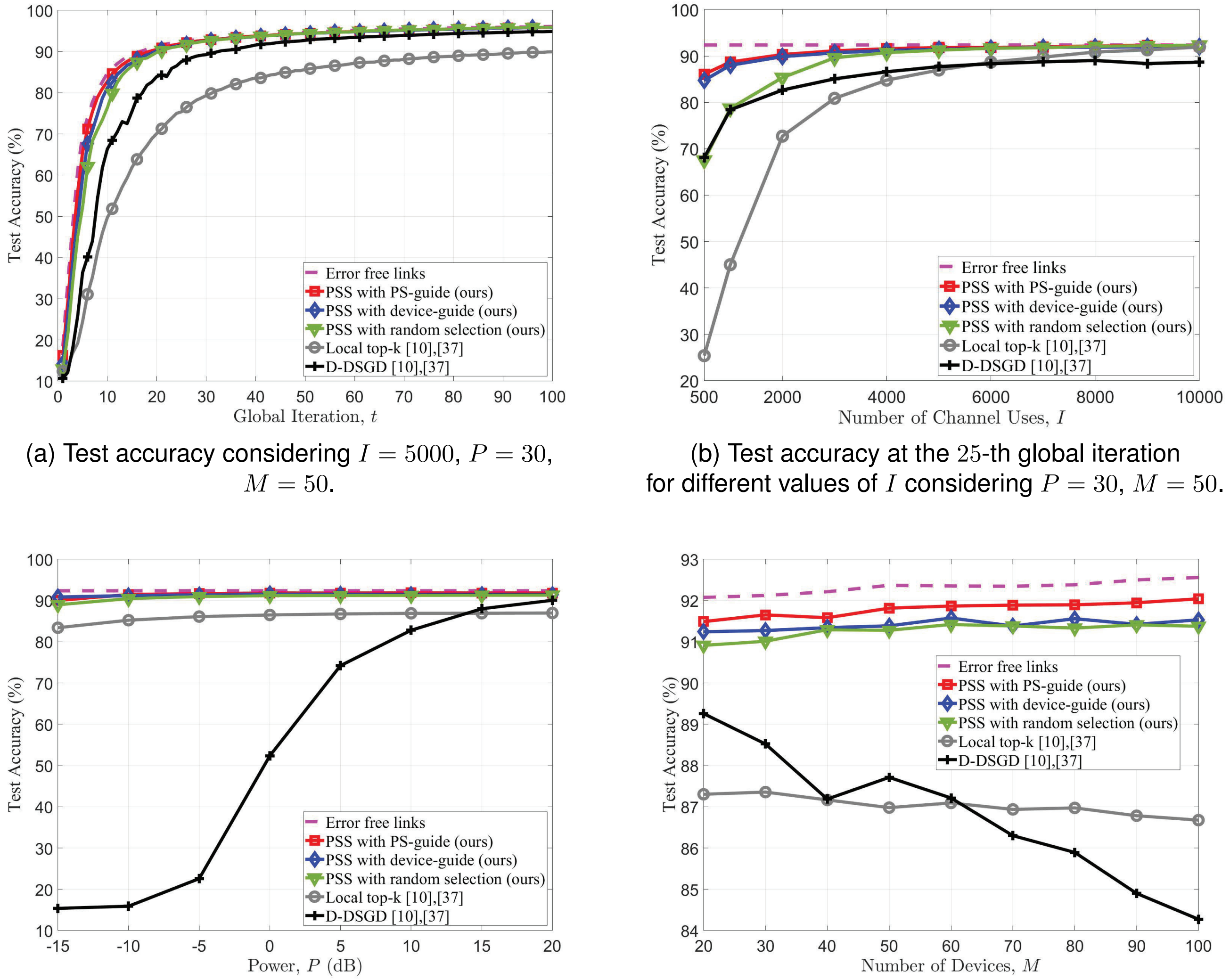
randomly select disjoint sets of 300 samples from the $60,000$ training MNIST examples, and allocate each set to a device. For the non-IID case, we randomly select two labels for each device and allocate 300 training MNIST examples of the selected labels, hence each device has training examples with only two labels. The size of total weight D is 21840; the learning rate η is 0.3; the sparsity level for local top-k is I; the ratio of channel uses in device-guided PSS, ρ, is 0.5; the channel h_m follows a Nakagami distribution with $(m, \Omega) = (3, 1)$. The performance metric is the average test accuracy for all devices measured over $10,000$ randomly selected images from the MNIST dataset. We assume that the PS has the balanced 300 samples from the $60,000$ training MNIST examples, when PSS with PS-guide is considered. Even if the PS can acquire only the imbalanced dataset in the case of non-IID data allocation, the performance of the PSS with PS-guide is lower bounded by the performance of the PSS with device-guide.

In Fig. 7-9, we compare the performance of the proposed PSS schemes and local top-k in various communication environments, considering IID and non-IID dataset allocation respectively. As a benchmark scheme, we show the performance of DSGD with no constraints on the wireless communication, denoted by “error free links”. And we show the performance of D-DSGD [10], [37], which is the state-of-the-art digital scheme. In D-DSGD, the devices set all elements of the gradient estimate to zero except for the elements of largest value. The position and mean value of those elements are sent by the digital transmission. Fig. 7 illustrates the training loss of the schemes which are “error free links”, “PSS with ps-guide”, “PSS with device-guide”, “PSS with random selection”, “local top-k”, and “D-DSGD”. In Fig. 7, it is observed that the proposed schemes converge faster than local top-k and D-DSGD, comparably to the error free links.

In Fig. 8(a) and 9(a), we compare the performance of the PSS schemes with the performance of local top-k, considering $I = 5000$, $P = 30$, and $M = 50$. In Fig. 8(b) and 9(b), we show the test accuracy of PSS schemes and local top-k at the 25-th global iteration for $500 \leq I \leq 10000$, when $P = 30$ and $M = 50$ are considered. In Fig. 8(c) and 9(c), we show the test accuracy at the 25-th global iteration for $-15 \leq P \leq 20$ (dB), when $I = 5000$ and $M = 50$ are considered. In Fig. 8(d) and 9(d), we consider $20 \leq M \leq 100$ for the comparison, when $P = 30$ and $I = 5000$.

In Fig. 8(a) - Fig. 8(d), it is observed that the PSS schemes considerably outperform the local top-k and D-DSGD, and the PSS with PS-guide and PSS with device-guide show better performance than PSS with random selection though the performance gap is relatively small. In Fig. 8(a), it is shown that the performance of schemes converge and the gap between the performance of error free links and that of the PSS schemes is very small. Thus, the assumed environment with $I = 5000$, $P = 30$, and $M = 50$ ensures that PSS schemes have no loss in performance due to the wireless factors. In Fig. 8(b), it is observed that utilizing more channel uses at each global iteration improves the performances, as expected. Besides that, the PSS schemes are robust to the scarce communication bandwidth, while the performance of local top-k significantly deteriorates by reducing the number of channel uses to 500.

In Fig. 8(c), it is shown that the analog FL are robust to the limited power budget, in particular the loss in performance due to power constraints is nearly negligible for $P \geq -15\text{dB}$. The robustness stems from the added Gaussian noise to the gradient estimates which does not seriously degrade the learning performance. In fact, noise injection is a popular technique for mitigating over-fitting. Moreover, it is observed that the convergence rate is not directly related to the power constraint in Theorem 1 and 2. This is in contrast to the performance of D-DSGD which is severely degraded for low power constraints. The degradation is due to the fact that the low power budget reduces the communication bandwidth, unlike the PSS schemes. In Fig. 8(d), the performance of PSS schemes tends not to improve even if M is increased



(a) Test accuracy considering $I = 5000$, $P = 30$, $M = 50$.

(b) Test accuracy at the 25-th global iteration for different values of I considering $P = 30$, $M = 50$.

(c) Test accuracy at the 25-th global iteration for different values of P considering $I = 5000$, $M = 50$.

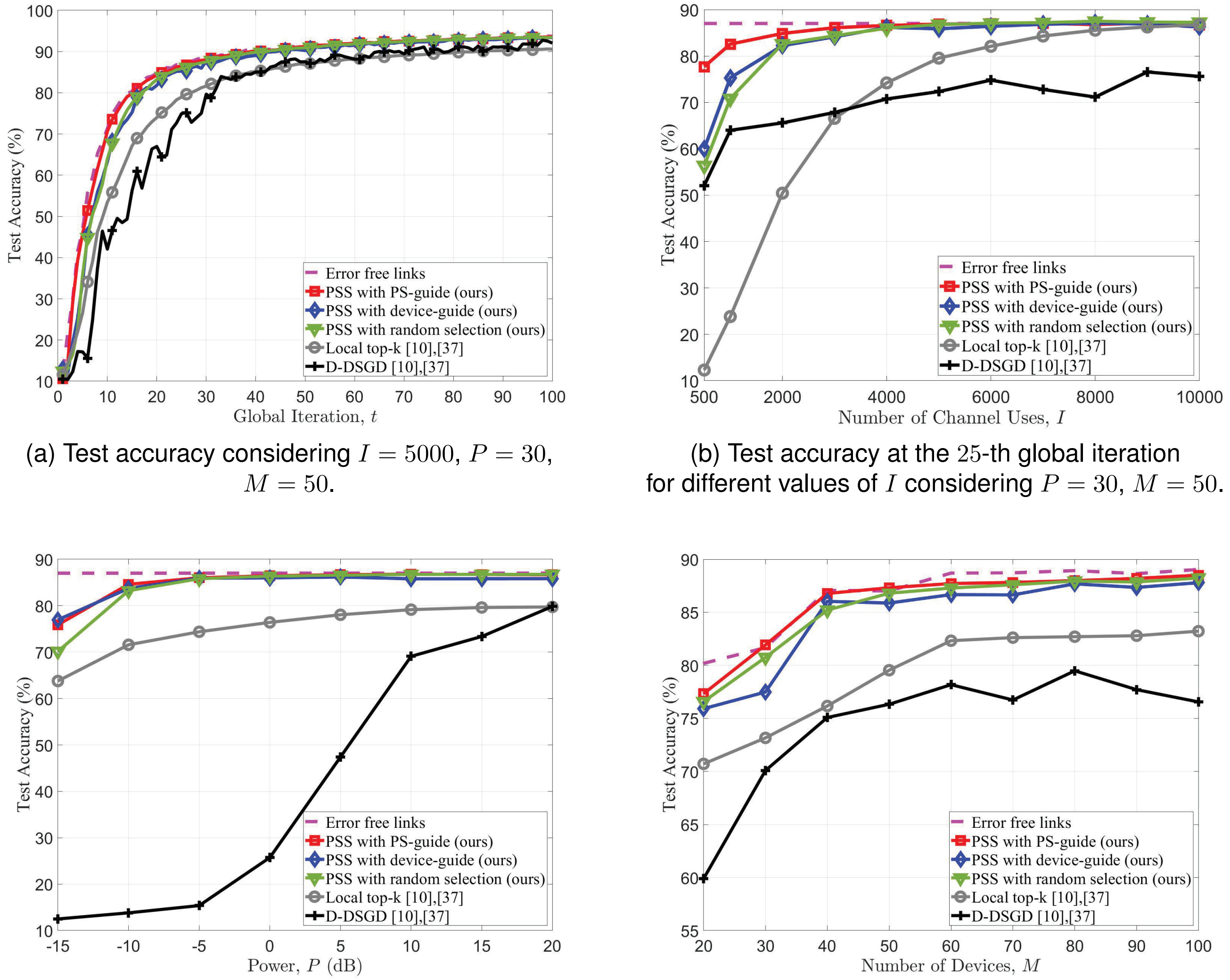
(d) Test accuracy at the 25-th global iteration for different values of M considering $I = 5000$, $P = 30$.

Fig. 8: Classification accuracy of analog FL schemes and D-DSGD in the case of IID data allocation.

over 20, similarly to the performance of error free links. It is remarkable that the performance of local top-k decreases when M is larger than 20 since the non-sparsity of summation of sparsified gradients increases while the gain from the number of devices does not increase when $M \geq 20$, as observed in the error free performance. For D-DSGD, the performance decreases since the devices share the resource of channel uses for upload.

In Fig. 9(a) - Fig. 9(d), the proposed PSS schemes considerably outperform local top-k and D-DSGD in the case of non-IID data allocation. Moreover, the PSS with random selection shows relatively improved performance compared to the other PSS schemes. The relative improvement is because the gain from the guide of the PS and device is less effective compared to the IID counterpart.

In Fig. 9(a), it is shown that the performance of schemes converge even in the case of non-IID data allocation and the gap between the performance of error free links and that of the PSS schemes is very small. In Fig. 9(b) and 9(c), we observe the robustness of PSS schemes to the number of channel uses and constrained power, while the performance of local top-k and D-DSGD deteriorate more significantly under the scarce communication bandwidth and lower power constraint, compared to the IID case. Furthermore, it is observed that the non-IID dataset allocation increases the minimum values of channel uses and constrained power guaranteeing the performance of PSS schemes comparable with the error free baseline. We note that the performance degradation due to wireless communication is relatively severe in the case of non-IID data from Figs. 9(b) and 9(c). In Fig. 9(d), the performance of PSS schemes tends not to change for $M \geq 40$, similarly to the performance of error free links. It is observed that the performance of local top-k does not decrease, unlike the case of IID data allocation. It is because of the gain coming from the number of the device in the case of non-IID data allocation. For D-DSGD, the performance decreases as the case of IID



(a) Test accuracy considering $I = 5000$, $P = 30$, $M = 50$.

(b) Test accuracy at the 25-th global iteration for different values of I considering $P = 30$, $M = 50$.

(c) Test accuracy at the 25-th global iteration for different values of P considering $I = 5000$, $M = 50$.

(d) Test accuracy at the 25-th global iteration for different values of M considering $I = 5000$, $P = 30$.

Fig. 9: Classification accuracy of analog FL schemes and D-DSGD in the case of non-IID data allocation.

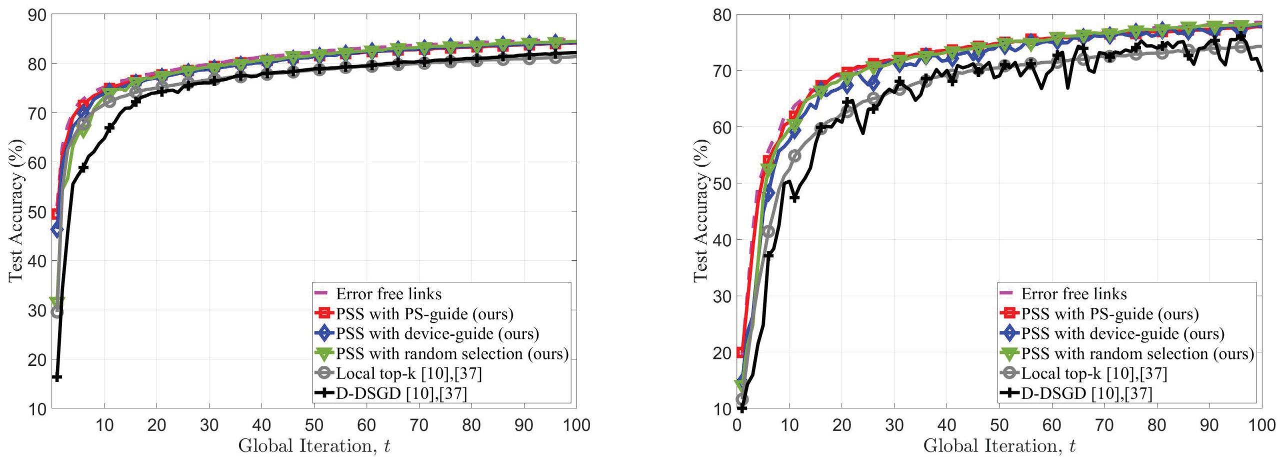
data allocation.

A. Extended results and discussion

In order to demonstrate the effectiveness of the proposed PSS schemes, we show the performance comparison of the schemes in both case of IID allocation and non-IID allocation for Fashion-MNIST dataset [50]. In Fig. 10, it is observed that the proposed PSS schemes outperform the local top-k and D-DSGD as the result for MNIST dataset. For more discussion about the *curse of primal averaging* in the local top-k, we compare the performance of local top-k and the local top-k with error free links. In the local top-k with error free links, each client sends the gradient estimate sparsified by the local top-k without any error. As illustrated in Fig. 11 (a), the local top-k without any error shows the comparable performance with the error free links while the performance of local top-k significantly deteriorates due to the recovery error. In Fig. 11 (b), we show the number of non-zero elements in the sum of sparsified gradient estimates, $\sum_{m=1}^{M} \boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$, with the number of total parameters and the non-zero elements in $\boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$ which are 21840 and 5000, respectively. Fig. 11 (b) illustrates that the $\sum_{m=1}^{M} \boldsymbol{g}_m^{sp}(\boldsymbol{\theta}_t)$ is not sparse in both cases and more dense in the case of non-IID allocation, where it validates the motivation of the proposed PSS schemes.

VI. Conclusion

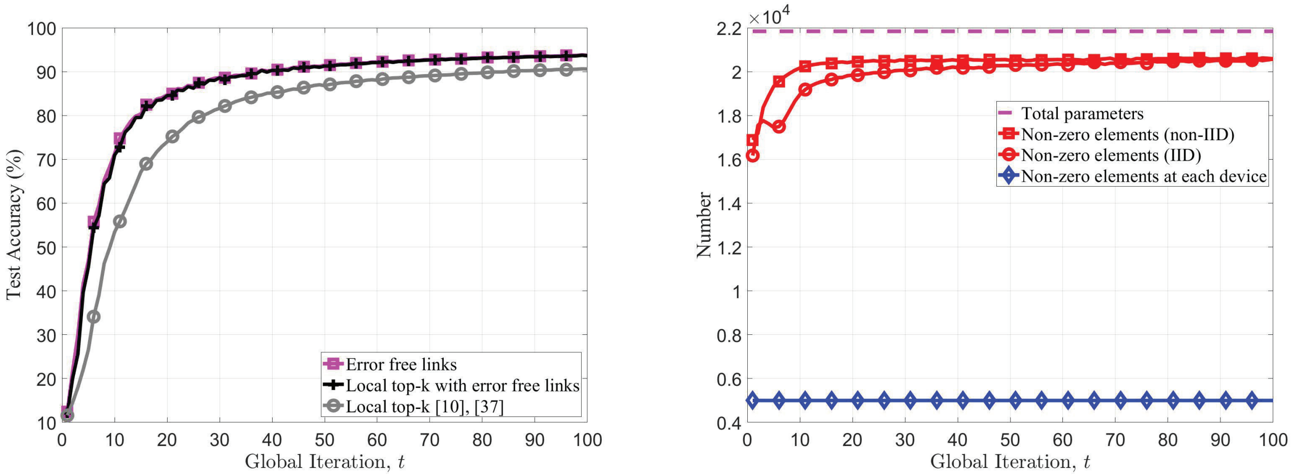
We have investigated the problem of analog federated learning when the available communication bandwidth is scarce calling for model compression strategies. We proposed a pattern shared sparsification strategy by setting the same sparsification pattern of gradient estimates uploaded at each edge device, unlike previous works in which each edge device independently applies a local top-k sparsification. Specifically, we provided several specific schemes for determining sparsification patterns, namely PSS with random selection, device-



(a) Test accuracy in the case of IID data allocation.

(b) Test accuracy in the case of non-IID data allocation.

Fig. 10: Classification accuracy of analog FL schemes and D-DSGD in the case of Fashion-MNIST considering $I = 5000$, $P = 30$, $M = 50$.



(a) Test accuracy in the case of non-IID data allocation of MNIST dataset.

(b) Number of non-zero elements of gradient estimate in the case of MNIST dataset.

Fig. 11: Demonstration of *curse of primal averaging* in the local top-k [10], [37] considering $I = 5000$, $P = 30$, $M = 50$.

guided PSS, and PS-guided PSS. The wireless implementation for all these PSS schemes were developed and validated empirically. In terms of convergence analysis, the model parameter obtained with the PSS schemes is proven to converge to the optimum of model parameter inducing minimal training loss. The proof is done by deriving the upper bound of the expectation of the squared l_2 norm between the obtained parameter and optimal parameter in closed-form. Numerical results show that the proposed PSS scheme consistently outperform current compression strategies in various wireless settings including the non-IID data regime. The PSS schemes show significant improvement in terms of robustness under scarce communication bandwidth and low power budget. In particular, the effect of communication bandwidth and power budget to the performance coincides with the results obtained in the convergence analysis.

Appendix A
Proof of the Theorem 1

When the PSS with random selection is considered, we have

$$\begin{aligned}
&\mathrm{E}\left[\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_2^2\right]\\
&=\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\frac{\eta}{M}\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]\\
&=\mathrm{E}\left[\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\|_2^2\right]-\frac{2\eta}{M}\mathrm{E}\left[\langle\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1},\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\rangle\right]\\
&\quad+\frac{\eta^2}{M^2}\mathrm{E}\left[\|\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}\|_2^2\right], \qquad \text{(A.1)}
\end{aligned}$$

since $\boldsymbol{\theta}_t$ is obtained as

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \frac{\eta}{M}\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}, \tag{A.2}$$

where $\boldsymbol{W}_t$ is defined as

$$\boldsymbol{W}_t \triangleq \sum_{m=1}^{M} \boldsymbol{g}_m\left(\boldsymbol{\theta}_t\right) + \boldsymbol{\Delta}_m^t + \boldsymbol{z}_t, \tag{A.3}$$

and $\boldsymbol{z}_t$ denotes the added noise in the uplink communication at the t-th global iteration. Since the PS scales the received signal with $\frac{1}{\gamma_t}$, $\boldsymbol{z}_t$ is a $d\times 1$ vector with IID $\mathcal{N}\left(0, \frac{1}{2\gamma_t^2}\right)$ entries. For further derivation of (A.1), we introduce two lemmas with proof.

Lemma A.1. *The second term in* (A.1) *is upper bounded by*

$$\begin{aligned}&-\frac{2\eta}{M}\mathrm{E}\left[\langle \boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle\right]\\ &\leq -\frac{2I\eta\mu}{D}\left(\sum_{t'=0}^{t-1}\left(1-\frac{2I}{D}\right)^{t-1-t'}\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t'} - \boldsymbol{\theta}^*\right\|_2^2\right]\right).\end{aligned} \tag{A.4}$$

Proof: First, we have

$$\begin{aligned}&-\frac{2\eta}{M}\mathrm{E}\left[\langle \boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle\right]\\ &\overset{(a)}{=} -\frac{4I\eta}{MD}\mathrm{E}\left[\langle \boldsymbol{W}_{t-1}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle\right] \qquad \text{(A.5)}\\ &\overset{(b)}{=} \frac{4I\eta}{MD}\mathrm{E}\left[\left\langle \sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\rangle\right]\\ &\quad -\frac{4I\eta}{MD}\mathrm{E}\left[\left\langle \sum_{m=1}^{M}\boldsymbol{\Delta}_m^{t-1}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\rangle\right], \qquad \text{(A.6)}\end{aligned}$$

where (a) is because of Lemma 1 and (b) is since the mean of each entry in $\boldsymbol{z}_{t-1}$ is 0. In (A.6), the first term is upper bounded by

$$\begin{aligned}&-\frac{4I\eta}{MD}\mathrm{E}\left[\left\langle \sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{\boldsymbol{t-1}}\right), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\rangle\right]\\ &\overset{(a)}{\leq} \frac{4I\eta}{MD}\sum_{m=1}^{M}\mathrm{E}\left[F_m\left(\boldsymbol{\theta}^*\right) - F_m\left(\boldsymbol{\theta}_{t-1}\right) - \frac{\mu}{2}\left\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\|_2^2\right]\\ &\overset{(b)}{\leq} -\frac{2I\eta\mu}{D}\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\|_2^2\right],\end{aligned} \tag{A.7}$$

where (a) is due to the Assumption 1 and (b) is since $F\left(\boldsymbol{\theta}^*\right) - F\left(\boldsymbol{\theta}_{t-1}\right) \leq 0$. And the second term of (A.6) is

$$\begin{aligned}&-\frac{4I\eta}{MD}\mathrm{E}\left[\left\langle \sum_{m=1}^{M}\boldsymbol{\Delta}_m^{t-1}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\rangle\right]\\ &\overset{(a)}{=} -\frac{4I\eta}{MD}\mathrm{E}\left[\langle\left(\boldsymbol{I}_d - \boldsymbol{S}_{t-2}\right)\boldsymbol{W}_{t-2}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle\right]\\ &\overset{(b)}{=} -\frac{4I\eta}{MD}\mathrm{E}\left[\langle\left(\boldsymbol{I}_d - \boldsymbol{S}_{t-2}\right)\boldsymbol{W}_{t-2}, \boldsymbol{\theta}_{t-2} - \boldsymbol{S}_{t-2}\boldsymbol{W}_{t-2} - \boldsymbol{\theta}^*\rangle\right]\\ &\overset{(c)}{=} -\frac{4I\eta}{MD}\cdot\left(1-\frac{2I}{D}\right)\mathrm{E}\left[\langle\boldsymbol{W}_{t-2}, \boldsymbol{\theta}_{t-2} - \boldsymbol{\theta}^*\rangle\right],\end{aligned} \tag{A.8}$$

where (a) is obtained from the definition of $\boldsymbol{\Delta}_m^{t-1}$; (b) is due to the (A.2); (c) holds because of

$$\langle\left(\boldsymbol{I}_D - \boldsymbol{S}_{t-2}\right)\boldsymbol{W}_{t-2}, \boldsymbol{S}_{t-2}\boldsymbol{W}_{t-2}\rangle = 0 \tag{A.9}$$

and Lemma 1. Hence, (A.6) is upper bounded by

$$\begin{aligned}&-\frac{2I\eta\mu}{D}\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\|_2^2\right] - \frac{4I\eta}{MD}\left(1-\frac{2I}{D}\right)\\ &\cdot\mathrm{E}\left[\langle\boldsymbol{W}_{t-2}, \boldsymbol{\theta}_{t-2} - \boldsymbol{\theta}^*\rangle\right],\end{aligned} \tag{A.10}$$

where we observe the recurrence relation of (A.5) as follows. When we denote (A.5) as X_{t-1}, we have

$$\begin{aligned}X_{t-1} &\leq -\frac{2I\eta\mu}{D}\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\right\|_2^2\right] + \left(1-\frac{2I}{D}\right)X_{t-2} \qquad \text{(A.11)}\\ &\overset{(a)}{\leq} -\frac{2I\eta\mu}{D}\left(\sum_{t'=0}^{t-1}\left(1-\frac{2I}{D}\right)^{t-1-t'}\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t'} - \boldsymbol{\theta}^*\right\|_2^2\right]\right), \qquad \text{(A.12)}\end{aligned}$$

where (a) is obtained by solving the recurrence relation in (A.11). ■

Lemma A.2. *The third term in* (A.1) *is upper bounded by*

$$\frac{\eta^2}{M^2}\mathrm{E}\left[\left\|\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}\right\|_2^2\right] \leq \left(1+\frac{\Psi}{P}\right)G^2\eta^2. \tag{A.13}$$

Proof: First, we have

$$\begin{aligned}&\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{\Delta}_m^t\right\|_2^2\right]\\ &= \mathrm{E}\left[\left\|\left(\boldsymbol{I}_D - \boldsymbol{S}_{t-1}\right)\left(\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right) + \boldsymbol{\Delta}_m^{t-1}\right)\right\|_2^2\right]\\ &\overset{(a)}{=} \left(1-\frac{2I}{D}\right)\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right) + \boldsymbol{\Delta}_m^{t-1}\right\|_2^2\right]\\ &\overset{(b)}{=} \left(1-\frac{2I}{D}\right)\left(\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)\right\|_2^2\right]\right.\\ &\quad \left.+\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{\Delta}_m^{t-1}\right\|_2^2\right]\right) \qquad \text{(A.14)}\\ &\overset{(c)}{\leq} \left(1-\frac{2I}{D}\right)\left(M^2G^2 + \mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{\Delta}_m^{t-1}\right\|_2^2\right]\right), \qquad \text{(A.15)}\end{aligned}$$

where (a) is due to Lemma 1; (b) is since

$$E\left[\left\langle\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right), \sum_{m=1}^{M}\boldsymbol{\Delta}_m^{t-1}\right\rangle\right] = 0, \tag{A.16}$$

when we reasonably assume that $\mathrm{E}\left[\boldsymbol{\Delta}_m^{t-1}\right] = 0$ and $\boldsymbol{\Delta}_m^{t-1}$ is independent with $\boldsymbol{\theta}_{t-1}$; (c) is due to Assumption 2 and Jensen's inequality. In (A.15) we observe the recurrence relation and have

$$\begin{aligned}\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{\Delta}_m^t\right\|_2^2\right] &\le \frac{D-2I}{2I}\left(1-\left(1-\frac{2I}{D}\right)^t\right)M^2G^2\\ &\le \frac{D-2I}{2I}M^2G^2,\end{aligned}\tag{A.17}$$

by solving the recurrence relation. From (A.17), we can derive (A.13) as follows:

$$\begin{aligned}&\frac{\eta^2}{M^2}\mathrm{E}\left[\left\|\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}\right\|_2^2\right]\\ &\overset{(a)}{=}\frac{2I\eta^2}{M^2D}\mathrm{E}\left[\left\|\boldsymbol{W}_{t-1}\right\|_2^2\right]\\ &=\frac{2I\eta^2}{M^2D}\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)+\boldsymbol{\Delta}_m^{t-1}+\boldsymbol{z}_{t-1}\right\|_2^2\right]\\ &\overset{(b)}{\le}\frac{2I\eta^2}{M^2D}\left(M^2G^2+\frac{D-2I}{2I}M^2G^2+\mathrm{E}\left[\left\|\boldsymbol{z}_{t-1}\right\|_2^2\right]\right)\\ &\le G^2\eta^2+\frac{2I\eta^2}{M^2D}\mathrm{E}\left[\left\|\boldsymbol{z}_{t-1}\right\|_2^2\right],\end{aligned}\tag{A.18}$$

where (a) is due to Lemma 1 and (b) is since each entry of $\boldsymbol{z}_{t-1}$ has 0 as its mean. In (A.18), we can derive the upper bound of $\mathrm{E}\left[\left\|\boldsymbol{z}_{t-1}\right\|_2^2\right]$ as follows:

$$\begin{aligned}&\mathrm{E}\left[\left\|\boldsymbol{z}_{t-1}\right\|_2^2\right]\\ &=D\cdot\mathrm{E}\left[\frac{1}{2\gamma_{t-1}^2}\right]\\ &=D\cdot\mathrm{E}\left[\frac{1}{2PI}\max_m\frac{\left\|\boldsymbol{v}_m^{t-1}\right\|_2^2}{\left|h_m^{t-1}\right|^2}\right]\\ &\le D\cdot\mathrm{E}\left[\frac{\max_m\left\|\boldsymbol{S}_{t-1}\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)+\boldsymbol{\Delta}_m^{t-1}\right)\right\|_2^2}{2PI}\right]\\ &\quad\cdot\mathrm{E}\left[\max_m\frac{1}{\left|h_m^{t-1}\right|^2}\right]\\ &\le DM\Psi\cdot\mathrm{E}\left[\sum_{m=1}^{M}\frac{\left\|\boldsymbol{S}_{t-1}\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)+\boldsymbol{\Delta}_m^{t-1}\right)\right\|_2^2}{2PI}\right]\\ &\overset{(a)}{\le}\frac{M\Psi}{P}\mathrm{E}\left[\sum_{m=1}^{M}\left\|\left(\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)+\boldsymbol{\Delta}_m^{t-1}\right)\right\|_2^2\right]\\ &\overset{(b)}{\le}\frac{M\Psi}{P}\left(MG^2+\mathrm{E}\left[\sum_{m=1}^{M}\left\|\boldsymbol{\Delta}_m^{t-1}\right\|_2^2\right]\right)\\ &\overset{(c)}{\le}\frac{M\Psi}{P}\left(MG^2+\frac{D-2I}{2I}MG^2\right)\\ &=\frac{D\Psi}{2PI}M^2G^2,\end{aligned}\tag{A.19}$$

for γ_t defined in (8) and $\Psi=\mathrm{E}\left[\max\limits_m\frac{1}{|h_m|^2}\right]$, where (a) is due to Lemma 1, (b) is due to Assumption 2 and Jensen's inequality, and (c) is since

$$\mathrm{E}\left[\left\|\boldsymbol{\Delta}_m^{t-1}\right\|_2^2\right]\le\frac{D-2I}{2I}MG^2,\tag{A.20}$$

which can be readily derived in a similar manner to (A.17). The combination of (A.18) and (A.19) concludes the proof. ■

By applying Lemma A.1 and A.2, we have

$$\begin{aligned}&\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\\ &\le\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]+\left(1+\frac{\Psi}{P}\right)G^2\eta^2\\ &\quad-\frac{2I\eta\mu}{D}\left(\sum_{t'=0}^{t-1}\left(1-\frac{2I}{D}\right)^{t-1-t'}\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t'}-\boldsymbol{\theta}^*\right\|_2^2\right]\right)\\ &\overset{(a)}{\le}\left(1-\eta\mu\left(1-\left(1-\frac{2I}{D}\right)^t\right)\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]\\ &\quad+\left(1+\frac{\Psi}{P}\right)G^2\eta^2\\ &\le A\left(t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\right\|_2^2\right]\\ &\quad+\left(1+\frac{\Psi}{P}\right)G^2\eta^2\left(1+\sum_{t'=0}^{t-2}\prod_{t''=0}^{t'}a(t-t'')\right)\\ &\overset{(b)}{\le}A\left(t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{D\eta G^2}{2I\mu}\left(1+\frac{\Psi}{P}\right),\end{aligned}\tag{A.21}$$

for $A(t)=\prod_{t'=1}^{t}a(t')$, and

$$a(t')=1-\eta\mu\left(1-\left(1-\frac{2I}{D}\right)^{t'}\right),\tag{A.22}$$

where (a) is since we can reasonably assume that $\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]\le\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t'}-\boldsymbol{\theta}^*\right\|_2^2\right]$ for $t'\le t-1$, and (b) is because $a(t-t'')\le a(1)$ for $\eta<\frac{1}{\mu}$.

APPENDIX B
PROOF OF THE THEOREM 2

With the same argument in the Appendix A, we have

$$\begin{aligned}&\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\\ &=\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]-\frac{2\eta}{M}\mathrm{E}\left[\left\langle\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1},\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\rangle\right]\\ &\quad+\frac{\eta^2}{M^2}\mathrm{E}\left[\left\|\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}\right\|_2^2\right].\end{aligned}\tag{B.1}$$

The third term of (B.1) is obtained in a different way from Lemma A.2 and we introduce the other lemma that handles the third term of (B.1) based on the Assumption 3.

Lemma B.1. *The third term in* (B.1) *is upper bounded by*

$$\begin{aligned}&\frac{\eta^2}{M^2}\mathrm{E}\left[\left\|\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}\right\|_2^2\right]\\ &\le L^2\eta^2\left(1-\left(1-\frac{2I}{D}\right)^t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{\Psi G^2\eta^2}{P}.\end{aligned}\tag{B.2}$$

Proof: In (A.14), we have

$$\begin{aligned}
&\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{\Delta}_m^t\right\|_2^2\right]\\
&=\mathrm{E}\left[\left\|\left(\boldsymbol{I}_D-\boldsymbol{S}_{t-1}\right)\left(\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)+\boldsymbol{\Delta}_m^{t-1}\right)\right\|_2^2\right]\\
&=\left(1-\frac{2I}{D}\right)\left(\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)\right\|_2^2\right]\right.\\
&\qquad\left.+\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{\Delta}_m^{t-1}\right\|_2^2\right]\right), \qquad \text{(B.3)}
\end{aligned}$$

where we observe the recurrence relation and obtain

$$\begin{aligned}
&\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{\Delta}_m^t\right\|_2^2\right]\\
&\leq\frac{D-2I}{2I}\left(1-\left(1-\frac{2I}{D}\right)^t\right)\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)\right\|_2^2\right], \qquad \text{(B.4)}
\end{aligned}$$

by solving the recurrence relation. And we have

$$\begin{aligned}
&\frac{\eta^2}{M^2}\mathrm{E}\left[\left\|\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}\right\|_2^2\right]\\
&=\frac{2I\eta^2}{M^2D}\mathrm{E}\left[\left\|\boldsymbol{W}_{t-1}\right\|_2^2\right]\\
&\overset{(a)}{\leq}\frac{2I\eta^2}{M^2D}\left(\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)\right\|_2^2\right]+\mathrm{E}\left[\left\|\boldsymbol{\Delta}_m^{t-1}\right\|_2^2\right]\right.\\
&\qquad\left.+\mathrm{E}\left[\left\|\boldsymbol{z}_{t-1}\right\|_2^2\right]\right)\\
&\overset{(b)}{\leq}\frac{2I\eta^2}{M^2D}\left(1+\frac{D-2I}{2I}\left(1-\left(1-\frac{2I}{D}\right)^{t-1}\right)\right)\\
&\qquad\cdot\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)\right\|_2^2\right]+\frac{\Psi G^2\eta^2}{P}\\
&=\frac{\eta^2}{M^2}\left(1-\left(1-\frac{2I}{D}\right)^t\right)\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)\right\|_2^2\right]+\frac{\Psi G^2\eta^2}{P}, \qquad \text{(B.5)}
\end{aligned}$$

where (a) is because of the same argument in (A.14) and (A.18), and (b) is due to (B.4) and (A.19). Finally we obtain

$$\begin{aligned}
&\frac{\eta^2}{M^2}\mathrm{E}\left[\left\|\boldsymbol{S}_{t-1}\boldsymbol{W}_{t-1}\right\|_2^2\right]\\
&\leq L^2\eta^2\left(1-\left(1-\frac{2I}{D}\right)^t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{\Psi G^2\eta^2}{P}, \qquad \text{(B.6)}
\end{aligned}$$

since we have

$$\begin{aligned}
&\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)\right\|_2^2\right]\\
&\overset{(a)}{=}\mathrm{E}\left[\left\|\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)-\sum_{m=1}^{M}\boldsymbol{g}_m\left(\boldsymbol{\theta}^*\right)\right\|_2^2\right]\\
&\overset{(b)}{\leq}M\sum_{m=1}^{M}\mathrm{E}\left[\left\|\boldsymbol{g}_m\left(\boldsymbol{\theta}_{t-1}\right)-\boldsymbol{g}_m\left(\boldsymbol{\theta}^*\right)\right\|_2^2\right]\\
&\overset{(c)}{\leq}M^2L^2\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right], \qquad \text{(B.7)}
\end{aligned}$$

where (a) is due to the differentiability of loss and $\boldsymbol{\nabla}F\left(\boldsymbol{\theta}^*\right)=0$, (b) because of Jensen's inequality, and (c) is derived by applying the equivalent definition of L-smoothness in Assumption 3.

■

By applying Lemma A.1 and B.1, we have

$$\begin{aligned}
&\mathrm{E}\left[\left\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right\|_2^2\right]\\
&\leq\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]\\
&-\frac{2I\eta\mu}{D}\left(\sum_{t'=0}^{t-1}\left(1-\frac{2I}{D}\right)^{t-1-t'}\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t'}-\boldsymbol{\theta}^*\right\|_2^2\right]\right)\\
&\quad+L^2\eta^2\left(1-\left(1-\frac{2I}{D}\right)^t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{\Psi G^2\eta^2}{P}\\
&\overset{(a)}{\leq}\left(1-\eta\left(\mu-L^2\eta\right)\left(1-\left(1-\frac{2I}{D}\right)^t\right)\right)\\
&\quad\cdot\mathrm{E}\left[\left\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{\Psi G^2\eta^2}{P}\\
&\leq B\left(t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{\Psi G^2\eta^2}{P}\left(1+\sum_{t'=0}^{t-2}\prod_{t''=0}^{t'}b(t-t'')\right)\\
&\overset{(b)}{\leq}B\left(t\right)\mathrm{E}\left[\left\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\right\|_2^2\right]+\frac{D\eta G^2\Psi}{2I\left(\mu-L^2\eta\right)P} \qquad \text{(B.8)}
\end{aligned}$$

for $B(t)=\prod_{t'=1}^{t}b(t')$, and

$$b(t')=1-\eta\left(\mu-L^2\eta\right)\left(1-\left(1-\frac{2I}{D}\right)^t\right), \qquad \text{(B.9)}$$

where (a) and (b) are obtained from the argument in (A.21) for $\eta<\frac{\mu}{L^2}$ implying $\eta\left(\mu-L^2\eta\right)<1$.

REFERENCES

[1] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[2] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *IEEE Transactions on Communications*, vol. 69, no. 8, pp. 5194-5208, Aug. 2021.

[3] 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. Int. Conf. on AISTATS*, Fort Lauderdale, Florida, Apr. 2017.

[4] 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 Mach. Learn.*, Barcelona, Spain, 2016.

[5] H. B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," [online]. Available. *https://ai.googleblog.com/2017/04/federated-learningcollaborative.html*, Apr. 2017.

[6] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[7] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization for heterogeneous networks," *arXiv:1812.06127 [cs.LG]*, Dec. 2018.

[8] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *arXiv:1903.02891 [cs.LG]*, Mar. 2019.

[9] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FEDAVG on non-IID data," *arXiv:1907.02189 [stat.ML]*, Jul. 2019.

[10] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels ," *IEEE Tr. Wireless Communication*, Feb. 2020.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition" in *International Conference on Learning Representations (ICLR)*, 2015.

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

[14] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," in *International Conference on Artificial Intelligence and Statistics, PMLR*, 2020.

[15] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konecny, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations*, 2021.

[16] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017.

[17] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs," in *INTERSPEECH*, Singapore, Sep. 2014, pp. 1058–1062.

[18] 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. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017.

[19] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proc. Int. Conf. Mach. Learn.*, vol. 80, Stockholm, Sweden, Jul. 2018, pp. 560–569.

[20] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Representations,* Vancouver, Canada, 2018.

[21] F. Sattler, S. Wiedemann, K.-R. Müller, W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," in *Proc. Int. Joint Conf. Neural Networks,* Budapest, Hungary, 2019, pp. 1–8.

[22] D. Alistarh, T. Hoefler, M. Johansson, S. Khirirat, N. Konstantinov, and C. Renggli, "The convergence of sparsified gradient methods," in *Proc. Conf. Neural Inf. Process. Syst.*, Montreal, Canada, 2018.

[23] Y. Tsuzuku, H. Imachi, and T. Akiba, "Variance-based gradient compression for efficient distributed deep learning," *arXiv:1802.06058*, Feb. 2018.

[24] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi "Federated learning with compression: Unified analysis and sharp guarantees," in *International Conference on Artificial Intelligence and Statistics, PMLR*, 2021.

[25] S. M. Shah and V. K. N. Lau, "Model compression for communication efficient federated learning," *IEEE Tr. Neural Networks and Learning Systems*, Dec. 2021. (early access)

[26] N. Ivkin, D. Rothchild, E. Ullah, V. Braverman, I. Stoica, and R. Arora, "Communication-efficient distributed sgd with sketching," in *Conference on Neural Information Processing Systems (NIPS)*, 2019.

[27] A. Malekijoo, M. J. Fadaeieslam, H. Malekijou, M. Homayounfar, F. Alizadeh-Shabdiz, and R. Rawassizadeh, "FEDZIP: A compression framework for communication-efficient federated learning," *arXiv:2102.01593 [cs.LG]*, Feb. 2021.

[28] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor "Convergence of update aware device scheduling for federated learning at the wireless edge," *arXiv:2001.10402 [cs.IT]*, Jan. 2020.

[29] 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 Tr. Wireless Communication*, Oct. 2020.

[30] Z. Yang, M. Chen, W. Saad, C. S. Hong and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Tr. Wireless Communication*, Nov. 2020.

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

[32] 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 *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, Paris, France, 2019.

[33] M. Salehi and E. Hossain, "Federated learning in unreliable and resource-constrained cellular wireless networks," *arXiv:2012.05137 [cs.LG]*, Dec. 2020.

[34] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3498-3516, Oct. 2007.

[35] J.-H. Ahn, O. Simeone, and J. Kang, "Wireless federated distillation for distributed edge learning with heterogeneous data," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Istanbul, Turkey Sep. 2019.

[36] J.-H. Ahn, O. Simeone, and J. Kang, "Cooperative Learning Via Federated Distillation Over Fading Channels," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020.

[37] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Tr. Signal Processing*, Mar. 2020.

[38] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for lowlatency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

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

[40] H. Guo, A. Liu and V. K. N. Lau, "Analog Gradient Aggregation for Federated Learning over Wireless Networks: Customized Design and Convergence Analysis," in *IEEE Internet of Things Journal*, Jun. 2020.

[41] Y. Sun, S. Zhou, and D. Gündüz, "Energy-aware analog aggregation for federated learning with redundant data," in *Proc. Conf. Int. Conf. Commu. (ICC)*, Jun. 2020.

[42] S. Shi, X. Chu, K. C. Cheung, and S. See, "Understanding top-k sparsification in distributed deep learning," *arXiv:1911.08772 [cs.LG]*, Nov. 2019.

[43] S. Shi, Q. Wang, K. Zhao, Z. Tang, Y. Wang, X. Huang, and X. Chu, "A distributed synchronous SGD algorithm with global top-k sparsification for low bandwidth networks," in *ICDCS*, 2019.

[44] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914-18919, Nov. 2009.

[45] H. Yuan, M. Zaheer, and S. Reddi, "Federated composite optimization," in *International Conference on Machine Learning, PMLR*, 2021.

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

[47] W. Liu, X. Zang, Y. Li and B. Vucetic, "Over-the-air computation systems: optimization, analysis and scaling laws," *IEEE Tr. Wireless Communication*, vol. 19, no. 8, pp. 5488-5502, Aug. 2020.

[48] D. Li, and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv:1910.03581*, 2019.

[49] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. Vincent Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," *IEEE Tr. Wireless Communication*, Jan. 2021.

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

This article has been accepted for publication in IEEE Transactions on Green Communications and Networking. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGCN.2022.3186538

Biography Section

Jin-Hyun Ahn received his B.S. and M.S. degree in mathematics in 2013 and 2016, and Ph.D. degree in electrical engineering in 2020, all from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. He is currently working as a research associate in the MGH/BWH Center for Advanced Medical Computing and Analysis, department of radiology, Massachusetts General Hospital and Harvard Medical School. His research interests lie in the field of probability theory, communication theory, and machine learning, with a specific focus on federated learning. From January to September 2019, he was a visiting researcher at King's Communications, Learning and Information Processing lab, King's College London, United Kingdom. He worked as a research associate in KAIST from September 2019 to September 2020. He has served as a reviewer for many journals, including the IEEE Journal on Selected Areas in Communications and the IEEE Transactions on Wireless Communications.

Mehdi Bennis is a full (tenured) Professor at the Centre for Wireless Communications, University of Oulu, Finland and head of the intelligent connectivity and networks/systems group (ICON). His main research interests are in radio resource management, game theory and distributed AI in 5G/6G networks. He has published more than 200 research papers in international conferences, journals and book chapters. He has been the recipient of several prestigious awards including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society, the 2017 EURASIP Best paper Award for the Journal of Wireless Communications and Networks, the all-University of Oulu award for research, the 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award and the 2020 Clarviate Highly Cited Researcher by the Web of Science. Dr Bennis is an editor of IEEE TCOM and Specialty Chief Editor for Data Science for Communications in the Frontiers in Communications and Networks journal. Dr Bennis is an IEEE Fellow.

Joonhyuk Kang received the B.S.E. and M.S.E. degrees from Seoul National University, Seoul, South Korea, in 1991 and 1993, respectively, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin, Austin, in 2002. From 1993 to 1998, he was a Research Staff Member at Sam- sung Electronics, Suwon, South Korea, where he was involved in the development of DSP-based real-time control systems. In 2000, he was with Cwill Telecommunications, Austin, TX, USA, where he participated in the project for multicarrier CDMA systems with antenna array. He was a Visiting Scholar with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA, from 2008 to 2009. He is currently serving as Head of the School of Electrical Engineering (EE), KAIST, Daejeon, South Korea. His research interests include signal processing and machine learning for wireless communication systems. He is a life-member of the Korea Information and Communication Society and the Tau Beta Pi (the Engineering Honor Society). He is a recipient of Qualcomm Innovation Award in 2013 and IEEE VTS Jack Neubauer Memorial Award in 2021 for his paper titled "Mobile Edge Computing via a UAV-Mounted Cloudlet: Optimization of Bit Allocation and Path Planning."