This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

Over-the-Air Federated Averaging with Limited Power and Privacy Budgets

Na Yan, Kezhi Wang, Cunhua Pan, Kok Keong Chai, Feng Shu,
and Jiangzhou Wang, *Fellow, IEEE*

***Abstract*—This paper develops an optimal design for device scheduling, alignment coefficient, and aggregation rounds within a differentially private over-the-air federated averaging (DP-OTA-FedAvg) system considering a constrained sum power budget. In DP-OTA-FedAvg, gradients are aligned using an alignment coefficient and then aggregated over the air, utilizing channel noise to ensure participant privacy. This study highlights two critical tradeoffs in aligned over-the-air federated learning (OTA-FL) systems with limited power and privacy budgets. Firstly, it reveals the tradeoff between the number of scheduled devices and the alignment coefficient. Secondly, it investigates the balance between aggregation distortion and local training error while adhering to the sum power constraint. Specifically, we measure privacy using differential privacy (DP) and perform convergence analyses for both convex and non-convex loss functions. These analyses provide insights into how device scheduling, the alignment coefficient, and the number of global aggregations affect both privacy preservation and the learning process. Building on these analytical results, we formulate an optimization problem aimed at minimizing the optimality gap of DP-OTA-FedAvg under power and privacy constraints. By specifying the number of aggregation rounds, we derive a closed-form expression describing the relationship between the alignment coefficient and the number of scheduled devices. We then tackle the problem through iterative optimization of scheduling and aggregation rounds. The effectiveness of the proposed policies is verified through simulations, and the performance advantage is particularly pronounced in scenarios where devices have poor channel conditions and limited sum-power budgets.**

Intex Terms

Federated averaging, differential privacy, device scheduling, and sum-power constraint.

I. Introduction

With the rapid increase in data volume and computing capability of edge devices, artificial intelligence (AI) and Internet of Things (IoT) are well-developed as a result of the unprecedented success of machine learning (ML) techniques, especially deep learning [1]. These systems normally employ highly parameterized models, such as deep neural networks (DNNs), which are trained by the massive data samples generated or collected by edge devices, e.g. smartphones and sensors. The conventional strategy for training these models is to aggregate all these raw data to a central server with high computing capability, where the training is performed [2]. However, such a centralized training paradigm is becoming more and more costly due to the transmission of raw samples with the dramatic growth in data amount. Furthermore, the raw data usually contains some personal information, and thus the users may refuse to share them with the server. All the above reasons inspire the development of federated learning (FL), which is a kind of privacy-preserving distributed ML paradigm [3]–[5].

FL enables the devices to train models collaboratively with the help of a central controller, such as a base station (BS) [3]. Instead of uploading the raw data to the BS, the model parameters and the gradients are exchanged between the devices and the BS. By training models locally, FL not only makes full use of the computing capability of the edge devices, but also effectively reduces the power consumption, latency, and privacy exposure caused by the transmission of the massive datasets. However, despite these promising benefits, FL still involves the following challenges. First, FL suffers from communication bottlenecks due to the high dimension of each local update, especially when a large number of participants try to upload gradients via a resource-limited wireless multiple access channel (MAC). This also leads to considerable upload latency as the bandwidth allocated to each participant decreases with the increased number of devices [3], [6]. Second, although FL offers basic privacy protection, which benefits from the fact that all raw data is processed locally, it is far from sufficiency if some attacks are applied to the exchanged messages, i.e., the gradients [7], [8]. This is because the gradients are obtained based on local data and therefore may contain some information of raw data [9].

One promising countermeasure to jointly overcome the two challenges is differentially private over-the-air FL (DP-OTA-FL) [10]–[14]. On one hand, differential privacy (DP) [15] preserves individual privacy in FL by introducing a controlled amount of random noise into the local update,

Part of this work will be presented in IEEE International Conference on Communications (ICC), 28 May – 01 June 2023, Rome, Italy. This work of Na Yan was supported by China Scholarship Council. *(Corresponding author: Kezhi Wang and Cunhua Pan.).* Na Yan and Kok Keong Chai are with School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: n.yan, michael.chai@qmul.ac.uk). Kezhi Wang is with Department of Computer Science, Brunel University London, Uxbridge, Middlesex, UB8 3PH, U.K. (email: kezhi.wang@brunel.ac.uk). Cunhua Pan is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (email: cpan@seu.edu.cn). Feng Shu is with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China, and also with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: shufeng0101@163.com). Jiangzhou Wang is with the School of Engineering, University of Kent, Canterbury CT2 7NT, U.K. (Email: j.z.wang@kent.ac.uk).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

ensuring that the contribution of any individual data sample to the model remains statistically indistinguishable. On the other hand, over-the-air FL (OTA-FL) [1], [11], [16]–[24] is promising to alleviate the above-mentioned communication bottleneck and reduce the communication and computation latency. It schedules the devices to convey their gradients simultaneously via a shared wireless MAC with analog signals, i.e., without converting the gradients to discrete coded symbols which need to be decoded at the BS. Then, the gradients are directly aggregated "over-the-air" thanks to the waveform-superposition property of a MAC. Furthermore, over-the-air computation (Aircomp) has been demonstrated to offer a surprising advantage in enhancing DP in FL. It achieves this by enabling the superposition of both local artificial noise and channel noise at the server, effectively enhancing privacy protection as shown in [10], where artificial Gaussian noise was injected into its update locally to enhance privacy. Instead of leveraging additional artificial noise, the work of [25] harnessed inherent communication noise to preserve DP by solely reducing transmit power. The authors in [13] proposed transmit power control algorithms that attain accurate FL training while satisfying given DP requirements. The above-mentioned works considered the aligned OTA-FL, where all the gradients are aligned by a constant, referred to as the alignment coefficient. In this way, the impact of the fading channel becomes a constant and can be easily removed by performing an inverse operation of the pre-processing at the BS. However, the alignment coefficient is limited by the device with the worst channel condition due to the peak transmit power constraint. This limitation can lead to an exceptionally low alignment coefficient, resulting in a very poor signal-to-noise ratio (SNR), particularly when all devices are scheduled for training [10], [13], [25]. To improve the alignment coefficient, the authors of [19], [26] set a threshold to schedule the devices with better channel conditions to participate in the training. Nevertheless, the optimal threshold was not given. The authors in [11] optimized the power control at edge devices and the denoising factors at the server to balance the trade-off on the compromised accuracy of federated averaging (FedAvg) and enhanced privacy, without considering the gradient alignment [13]. However, the aforementioned studies typically revolve around peak power constraints. It is also crucial to investigate sum-power limited OTA-FL systems [23], [27], [28] for the practical deployment of FL as these edge devices are power-constrained due to limited battery capacity. In [23], the device selection and power control were jointly optimized to improve the performance of OTA-FedAvg subjecting to the individual and sum uplink transmit power constraints. However, the number of local training rounds of FedAvg was pre-defined. To the best of our knowledge, there has been no prior study addressing the optimization of aggregation rounds under sum-power constraints, aiming to strike a balance between reducing transmission disturbance and increasing local training error. This exploration is essential for guiding the design of device scheduling and aggregation within powe-limited FedAvg systems.

In this paper, a scheme is proposed to jointly design device scheduling, alignment coefficient, and global aggregation for an aligned differentially private OTA-FedAvg (DP-OTA-FedAvg) system with limited sum power and privacy budgets. The device scheduling, alignment coefficient, and global aggregation can affect the performance of DP-OTA-FedAvg in two ways. On one hand, in each communication round, scheduling more devices to participate in the training is beneficial to alleviate the error of the average gradient. However, the alignment coefficient may decrease with the increased number of the scheduled devices as it is more likely to involve the devices with poor channel conditions, which can significantly lower down the alignment coefficient [10], thus degrading the utility of the aggregated gradient. Therefore, there is a tradeoff between the number of scheduled devices and the alignment coefficient. Additionally, scheduling more devices in each aggregation round may consume more power. As a result, the number of aggregation rounds will be reduced due to the limited sum power budget. Then, the number of local training will increase with the reduced number of global aggregation rounds, which leads to a larger local training error. Therefore, it is crucial for DP-OTA-FedAvg systems with limited sum power budget to design the device scheduling, alignment coefficient, and aggregation rounds. The main contributions can be summarized as follows:

- This paper presents an optimal design of device scheduling, alignment coefficient, and aggregation rounds for DP-OTA-FedAvg (O-DP-OTA-FedAvg), which effectively addresses two crucial tradeoffs within sum-power-limited DP-OTA-FedAvg systems. Firstly, we investigate the delicate balance between accommodating a higher number of participants and the resulting reduction in the alignment coefficient. Secondly, we explore the tradeoff between reducing local training error and increasing aggregation distortion through the optimization of aggregation rounds.
- To characterize the impact of the alignment coefficient on the privacy preservation of OTA-FedAvg, we first conduct the privacy analysis. Then, we derive the closed-form expressions of the optimality gap and the average-squared gradient to demonstrate the convergence of DP-OTA-FedAvg in the cases of convex and non-convex loss functions, respectively. These closed-form expressions quantify the impact of analog over-the-air aggregation on the convergence of DP-OTA-FedAvg, characterizing how the design of the alignment coefficient, device scheduling, and the number of aggregation rounds can affect the privacy protection and the performance of DP-OTA-FedAvg.
- Based on these closed-form theoretical results, we formulate an optimization problem to minimize the optimality gap by jointly designing the device scheduling, alignment coefficient, and aggregation rounds considering the limited sum power and privacy budgets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

- The problem is decoupled into two sub-problems. By giving the number of communication rounds, the optimal design of device scheduling and alignment coefficient is studied. We obtain limited potential optimal solution pairs by exploring the relationship between the number of scheduled devices and the alignment coefficient. Thanks to the reduced search space, the optimal solution can be efficiently obtained. Given the optimal device scheduling and alignment coefficient, the optimal number of aggregation rounds can be obtained by searching a limited solution space.

A. Organization

The remainder of this paper is organized as follows. In Section II, we present the system model, aligned OTA-FedAvg, and the definitions of DP. The theoretically analytical results are presented in Section III. We formulate an optimization problem in Section IV. The simulation results are shown in Section V and we conclude the paper in Section VI.

II. System Model and Preliminaries

As shown in Fig. 1, we consider a DP-OTA-FedAvg system consisting of a BS and N edge devices indexed by $\mathcal{N} = \{1, \cdots, N\}$. Assume that each device of index $k \in \mathcal{N}$ stores a local dataset $\mathcal{D}_k$ which contains D_k pairs of training samples $(\boldsymbol{u}, v)$ where $\boldsymbol{u}$ is the raw data for training and v is the corresponding label. For simplicity, we assume that $D_1 = \cdots = D_N$. The BS and these devices collaborate to train an ML model by exchanging the models and gradients without sharing these locally stored raw data, which offers basic protection for users' personal information. However, the BS is assumed to be curious and attempts to probe sensitive information from the received gradients, threatening users' privacy. In this work, the privacy of the scheduled devices can be guaranteed by channel noise by designing the alignment coefficient.

The goal of an FL task is to obtain the optimal model parameterized by $\boldsymbol{m}^*$ by minimizing the average global loss $L(\boldsymbol{m})$, i.e.,

$$\boldsymbol{m}^* = \arg\min_{\boldsymbol{m}} L(\boldsymbol{m}) \triangleq \frac{1}{N}\sum_{k=1}^{N} L_k(\boldsymbol{m}), \tag{1}$$

where $\boldsymbol{m} \in \mathbb{R}^d$ is the model parameter to be optimized. More specifically, the objective function of device k is defined as:

$$L_k(\boldsymbol{m}) = \frac{1}{D_k}\sum_{(\boldsymbol{u},v)\in\mathcal{D}_k} l(\boldsymbol{m};(\boldsymbol{u},v)), \tag{2}$$

where $l(\boldsymbol{m};(\boldsymbol{u},v))$ denotes the loss function, quantifying the error of model $\boldsymbol{m}$ on the input-output data pair $(\boldsymbol{u}, v)$.

A. Over-the-Air Federated Averaging

To solve the problem in (1) while reducing the communication overhead, we employ the classic and widely-adopted FedAvg algorithm, which is implemented in an iterative manner. Generally, it requires a number of global aggregations, i.e., communication rounds, between devices and the BS to achieve the desired accuracy level of the learned global model $\boldsymbol{m}$. Specifically, we assume that T and I are the number of total training rounds and the number of communication rounds, respectively. Consequently, the local training step in each communication round is decided by $E = \frac{T}{I}$, and we assume that T is divisible by I [1]. Specifically, in each communication round $i \in \{0, \cdots, I-1\}$, FedAvg consists of the following steps: *(1) Parameter broadcasting*: At the beginning of communication round i, the BS broadcasts the latest global model parameter $\boldsymbol{m}^i$ to the scheduled devices denoted by $\mathcal{K}$, $\mathcal{K} \in \mathcal{N}$. *(2) Local training*: Each device first performs the initialization of the local model by setting the received global model parameter as the initial local model parameter, i.e., $\boldsymbol{w}_k^{i,0} = \boldsymbol{m}^i, \forall k \in \mathcal{K}$. Then, each device performs E rounds of local training by

$$\boldsymbol{w}_k^{i,\iota+1} = \boldsymbol{w}_k^{i,\iota} - \tau\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right), \iota \in \{0, ..., E-1\}, \tag{3}$$

where τ is the learning rate and

$$\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right) = \frac{1}{D_k}\sum_{(\boldsymbol{u},v)\in\mathcal{D}_k} \nabla l\left(\boldsymbol{w}_k^{i,\iota};(\boldsymbol{u},v)\right). \tag{4}$$

(3) Over-the-air aggregation: Upon completing E times of local training, each scheduled device uploads the accumulative gradients in this current communication round to the BS, i.e.,

$$\boldsymbol{g}_k^i = \frac{1}{\tau}\left(\boldsymbol{w}_k^{i,E} - \boldsymbol{w}_k^{i,0}\right) = \sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right). \tag{5}$$

To further alleviate communication bottlenecks and unbearable upload latency, we adopt analog over-the-air aggregation in this work, which enables the scheduled devices to simultaneously communicate their gradients to the BS via a shared MAC. Taking device k as an example, the gradient is transmitted by a pre-processed signal $\boldsymbol{x}_k^i$:

$$\boldsymbol{x}_k^i = e^{-j\psi_k}\left(\frac{\sqrt{\varphi_k P_k}}{\varpi}\boldsymbol{g}_k^i\right), \tag{6}$$

where $e^{-j\psi_k}$ is the local phase correction performed by the device k. P_k is the maximum transmission power of device k and $\varphi_k \in [0, 1]$ is the power scaling factor. We assume that the upper bound of each gradient's 2-norm is ϖ, i.e., $\left\|\boldsymbol{g}_k^i\right\|_2 \le \varpi$, so that $\mathrm{E}\left[\|\boldsymbol{x}\|_2^2\right] \le P_k$. The scheduled devices upload their local gradients $\boldsymbol{g}_k^i$ via the uncoded form with perfect time synchronization among them. In this way, the gradients can be aggregated over the air thanks to the

[1]Since I and E are in one-to-one correspondences when we have a fixed T, we use E and I exchangeably when we discuss the impact of the communication rounds I in the rest of this paper.

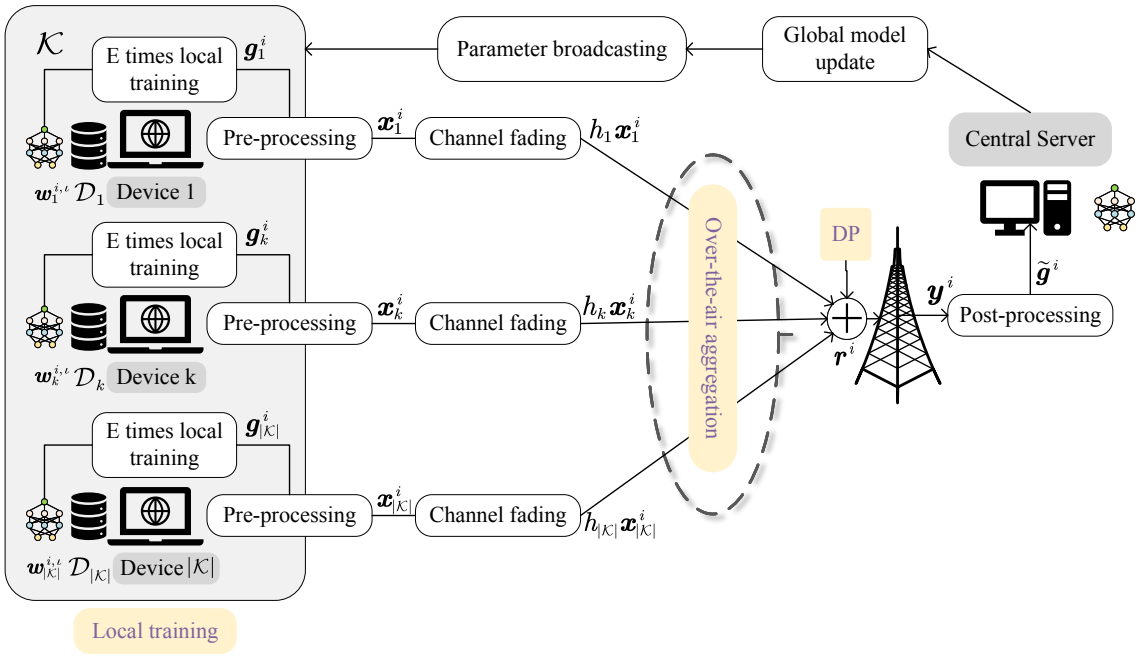


Fig. 1: The procedure of DP-OTA-FL.

superposition property of MAC. Consequently, the received signal at the BS is given by

$$\begin{aligned} \boldsymbol{y}^i &= \sum_{k\in\mathcal{K}} h_k \boldsymbol{x}_k^i + \boldsymbol{r}^i \\ &= \sum_{k\in\mathcal{K}} |h_k| \frac{\sqrt{\varphi_k P_k}}{\varpi} \boldsymbol{g}_k^i + \boldsymbol{r}^i, \end{aligned} \tag{7}$$

where $h_k = |h_k| e^{j\psi_k}$ is the complex-valued time-invariant channel coefficient between device k and the BS. The received noise $\boldsymbol{r}^i \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$ at the BS is employed to prevent privacy leakage in this paper. To recover the desired average gradient $\boldsymbol{g}^i = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \boldsymbol{g}_k^i$ from the received signal, the BS performs the post-processing by

$$\tilde{\boldsymbol{g}}^i = \frac{1}{|\mathcal{K}|\,\nu} \boldsymbol{y}^i = \frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} |h_k| \frac{\sqrt{\varphi_k P_k}}{\nu\varpi} \boldsymbol{g}_k^i + \frac{1}{|\mathcal{K}|\,\nu} \boldsymbol{r}^i, \tag{8}$$

where ν is a post-processing factor, which is referred to as the alignment coefficient. The induced error between the recovered gradient and the desired gradient is derived as:

$$\Delta \boldsymbol{g}_{err}^i = \underbrace{\frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} \left(|h_k| \frac{\sqrt{\varphi_k P_k}}{\nu\varpi} - 1 \right) \boldsymbol{g}_k^i}_{fading\ error} + \underbrace{\frac{1}{|\mathcal{K}|\,\nu} \boldsymbol{r}^i}_{noise\ error}. \tag{9}$$

The estimate gradient recovered from the over-the-air aggregated gradient results in two sources of error, i.e., the misalignment error due to fading and the additive error due to the noise. In order to eliminate the fading-related error, the gradients need to be aligned by the alignment coefficient ν by adjusting the power scaling factor φ_k in pre-precessing as follows,

$$|h_k| \frac{\sqrt{\varphi_k P_k}}{\varpi} = \nu, \forall k \in \mathcal{K}, \tag{10}$$

which is referred to as the aligned OTA-FL and was also studied in [10]. Following such an aligned aggregation scheme, the received signal at the BS in (7) can be simplified as:

$$\boldsymbol{y}^i = \nu \sum_{k\in\mathcal{K}} \boldsymbol{g}_k^i + \boldsymbol{r}^i, \tag{11}$$

and the estimated average gradient is finally given by,

$$\tilde{\boldsymbol{g}}^i = \frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} \boldsymbol{g}_k^i + \frac{1}{|\mathcal{K}|\,\nu} \boldsymbol{r}^i. \tag{12}$$

(4) Model update: The BS updates the global model parameter based on the estimated average gradient as follows:

$$\boldsymbol{m}^{i+1} = \boldsymbol{m}^i - \tau \tilde{\boldsymbol{g}}^i. \tag{13}$$

The above iteration steps are repeated until a certain training termination condition is met.

B. Power Constraints of the DP-OTA-FedAvg System

In this paper, we consider both the peak transmit power constraint of each device and the sum power constraint of the overall DP-OTA-FedAvg system.

1) Peak power constraint: Following (10), we have

$$\varphi_k = \frac{\nu^2 \varpi^2}{|h_k|^2 P_k}, \forall k \in \mathcal{K}. \tag{14}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

To make sure that $\varphi_k \leq 1$, the alignment coefficient ν needs to satisfy:

$$\nu \leq \frac{\min\limits_{s\in\mathcal{K}} \left\{|h_s| \sqrt{P_s}\right.}{\varpi}. \tag{15}$$

From (15), we can learn that the alignment coefficient ν is limited by the device with the worst channel condition among the scheduled devices, i.e., $\min\limits_{s\in\mathcal{K}} \left\{|h_s| \sqrt{P_s}\right\}$. However, a larger ν is expected to mitigate the noise-related error following (9). Since the learning performance will be degraded due to a small ν, i.e., large noise error, the optimal design of device scheduling to improve the alignment coefficient is significant, especially in the FL systems where devices are power-limited and some of the devices suffer from poor channel conditions.

2) Sum power constraint: In each communication round, the power consumption for transmitting gradient of device k is:

$$\varphi_k P_k = \frac{\nu^2 \varpi^2}{|h_k|^2}, \forall k \in \mathcal{K}. \tag{16}$$

Assume that the total power budget for the communication rounds of DP-OTA-FedAvg is P^{tot}. Then, the sum power transmit power constraint is given by,

$$\sum\nolimits_{k\in\mathcal{K}} \frac{\nu^2 \varpi^2}{|h_k|^2} \leq \frac{P^{tot}}{I}. \tag{17}$$

From (17), we can learn that if the number of the communication rounds I is small, the power budget in each communication round for gradient uploading will be large, which means that we can schedule more devices or set a large alignment coefficient. The impact of the number of the communication rounds I, the number of the scheduled devices, and the alignment coefficient ν on the learning performance will be discussed in Section III.

C. Differential Privacy

DP [15] is defined on the conception of the adjacent dataset, which guarantees the probability that any two adjacent datasets output the same result is less than a constant with the help of adding random noise. More specifically, DP quantifies information leakage in FL by measuring the sensitivity of the gradients to the change of a single data point in the input dataset. The basic definition of (ϵ, ξ)-DP is given as follows.

Definition 1. *(ϵ, ξ)-DP [15]: A randomized mechanism $\mathcal{O}$ guarantees (ϵ, ξ)-DP if for two adjacent datasets $\mathcal{D}, \mathcal{D}'$ differing in one sample, and measurable output space $\mathcal{Q}$ of $\mathcal{O}$, it satisfies,*

$$\Pr\left[\mathcal{O}\left(\mathcal{D}\right) \in \mathcal{Q}\right] \leq e^{\epsilon} \Pr\left[\mathcal{O}\left(\mathcal{D}'\right) \in \mathcal{Q}\right] + \xi. \tag{18}$$

The additive term ξ allows for breaching ϵ-DP with the probability ξ while ϵ denotes the protection level and a smaller ϵ means a higher privacy preservation level. Specifically, the Gaussian DP mechanism which guarantees privacy by adding artificial Gaussian noise is introduced as follows.

Definition 2. *Gaussian mechanism [15]: A mechanism $\mathcal{O}$ is called as a Gaussian mechanism, which alters the output of another algorithm $\mathcal{L}: \mathcal{D} \to \mathcal{Q}$ by adding Gaussian noise, i.e.,*

$$\mathcal{O}\left(\mathcal{D}\right) = \mathcal{L}\left(\mathcal{D}\right) + \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right). \tag{19}$$

Gaussian mechanism $\mathcal{O}$ guarantees (ϵ, ξ)-DP with $\epsilon = \frac{\Delta S}{\sigma}\sqrt{2\ln\left(\frac{1.25}{\xi}\right)}$ where $\Delta S \triangleq \max\limits_{\mathcal{D},\mathcal{D}'} \left\|\mathcal{L}\left(\mathcal{D}\right) - \mathcal{L}\left(\mathcal{D}'\right)\right\|_2$ is the sensitivity of the algorithm $\mathcal{L}$ quantifying the sensitivity of the algorithm $\mathcal{L}$ to the change of a single data point.

III. Privacy and Convergence Analysis of DP-OTA-FedAvg

To reveal the impact of over-the-air aggregation on privacy and learning performance, we conduct privacy and convergence analysis in this section. Then, based on these analytical results, we formulate an optimization problem to minimize the optimality gap by optimizing the device scheduling, alignment coefficient, and the number of communication rounds subject to privacy and sum power constraints.

For analysis purposes, we provide the following common assumptions first.

Assumption 1. *The expected squared norm of each gradient is bounded:*

$$\mathbb{E}\left[\left\|\boldsymbol{g}_k^i\right\|_2\right] \leq \varpi. \tag{20}$$

Assumption 2. *Assume that $L\left(\cdot\right)$ is ζ-smooth, i.e., for all $\boldsymbol{\iota}'$ and $\boldsymbol{\iota}$, one has*

$$L\left(\boldsymbol{\iota}'\right) - L\left(\boldsymbol{\iota}\right) \leq \left(\boldsymbol{\iota}' - \boldsymbol{\iota}\right)^{\mathrm{T}} \nabla L\left(\boldsymbol{\iota}\right) + \frac{\zeta}{2}\left\|\boldsymbol{\iota}' - \boldsymbol{\iota}\right\|_2^2. \tag{21}$$

A. Privacy Analysis

We aim to improve the learning performance while achieving a certain level of DP of the participants in the OTA-FedAvg system by designing device scheduling and alignment coefficient. We conduct the privacy analysis based on the Gaussian mechanism of DP in the following. To calculate the privacy leakage according to the Gaussian mechanism, the key point is the sensitivity of the OTA-FedAvg algorithm to the change of a single data point in the input dataset. Taking device m as an example, assume that $\mathcal{D}_m$ and $\mathcal{D}_m^{'}$ are two adjacent datasets differing in one sample, and $\boldsymbol{g}_m^i$ and $\left(\boldsymbol{g}_m^i\right)'$ are the two gradients obtained based on $\mathcal{D}_m$ and $\mathcal{D}_m^{'}$, respectively. The two signals received at the BS corresponding to datasets $\mathcal{D}_m$ and $\mathcal{D}_m^{'}$ are given by

$$\begin{aligned} \boldsymbol{y}^i &= \nu \sum_{k\in\mathcal{K}} \boldsymbol{g}_k^i + \boldsymbol{r}^i, \\ \left(\boldsymbol{y}^i\right)' &= \nu \sum_{k\in\mathcal{K}, k\neq m} \boldsymbol{g}_k^i + \left(\boldsymbol{g}_m^i\right)' + \boldsymbol{r}^i, \end{aligned} \tag{22}$$

which only differ in the gradient from device m. Then, the sensitivity of the OTA-FedAvg is given by $\Delta S_m^i \triangleq \max\limits_{\mathcal{D}_m, \mathcal{D}_m^{'}} \left\|\boldsymbol{y}^i - \left(\boldsymbol{y}^i\right)'\right\|_2$ and we have the following lemma.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

Lemma 1. *Assume that Assumption 1 holds and the set of the scheduled devices is $\mathcal{K}$. For each device $k \in \mathcal{K}$, such a OTA-FedAvg algorithm achieves (ϵ_k, ξ)-DP in each communication round where*

$$\epsilon_k = \frac{2\varpi\nu}{\sigma} \cdot \sqrt{2\ln\frac{1.25}{\xi}}, k \in \mathcal{K}. \tag{23}$$

Proof: Accroding to the definiation of sentivisity and (22), we have

$$\begin{aligned}\Delta S_m^i &\triangleq \max_{\mathcal{D}_m, \mathcal{D}_m'} \left\| \boldsymbol{y}^i - \left(\boldsymbol{y}^i\right)' \right\|_2 = \nu \max_{\mathcal{D}_m, \mathcal{D}_m'} \left\| \boldsymbol{g}_m^i - \left(\boldsymbol{g}_m^i\right)' \right\|_2 \\ &= \nu \left\| \boldsymbol{g}_m^i - \left(\boldsymbol{g}_m^i\right)' \right\|_2 \overset{(a)}{\leq} 2\varpi\nu,\end{aligned} \tag{24}$$

where (a) is from triangular inequality and Assumption 1. Following Gaussian mechanism of DP and replacing m with k, one completes the proof of Lemma 1. ■

Lemma 1 characterizes the impact of the alignment coefficient on privacy protection. More specifically, a smaller alignment coefficient ν leads to less privacy leakage. Physically speaking, a smaller alignment coefficient ν decreases the amplitude of the gradient signal, which enables the gradient more easily hidden in the channel noise. However, it degrades the utility of the gradients, which is validated in the following convergence analysis results.

Remark 1. *Note that when the "=" in (23) is replaced by "≤", it indicates a stronger privacy protection so it still satisfies (ϵ_k, ξ)-DP.*

B. Convergence Analysis

We here present convergence analysis in the cases of convex and non-convex loss functions. We first present the expectation of the gap between the updated global model $\boldsymbol{m}^{i+1}$ and the current global model $\boldsymbol{m}^i$ for the following analysis.

Lemma 2. *Given the learning rate $\tau \leq \frac{1}{\zeta}$, the upper bound of the gap between the updated global model $\boldsymbol{m}^{i+1}$ and the current model $\boldsymbol{m}^i$, i.e., $\mathbb{E}\left[L\left(\boldsymbol{m}^{i+1}\right)\right] - \mathbb{E}\left[L\left(\boldsymbol{m}^i\right)\right]$ is given by*

$$\begin{aligned}&\mathbb{E}\left[L\left(\boldsymbol{m}^{i+1}\right)\right] - \mathbb{E}\left[L\left(\boldsymbol{m}^i\right)\right] \leq -\frac{\tau}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^i\right)\right\|_2^2\right] \\ &+ \tau\varpi^2\left(E-1\right)^2 + 4\tau\varpi^2\left(1 - \frac{|\mathcal{K}|}{N}\right)^2 + \frac{\zeta\tau^2}{2}\frac{d\sigma^2}{|\mathcal{K}|^2\nu^2}.\end{aligned} \tag{25}$$

The expectation is with respect to the randomness of Gaussian noise.

Proof: Please refer to Appendix A. ■

For notation simplicity, we define $\theta = \nu\varpi$ as an equivalent substitution of ν and refer to it as the alignment factor. In the following, we mainly focus on the alignment factor θ instead of ν. Based on Lemma 2, we give the following convergence analysis in both convex and non-convex settings.

1) Convex Setting: We first consider the most benign setting, where the loss function $L(\cdot)$ is assumed to be strongly convex. We formalize a strong convexity assumption as below.

Assumption 3. *Assume that $L(\cdot)$ is strongly convex with a positive parameter ϱ, i.e., for all $\boldsymbol{\iota}'$ and $\boldsymbol{\iota}$, one has*

$$L\left(\boldsymbol{\iota}'\right) - L\left(\boldsymbol{\iota}\right) \geqslant \left(\boldsymbol{\iota}' - \boldsymbol{\iota}\right)^{\mathrm{T}} \nabla L\left(\boldsymbol{\iota}\right) + \frac{\varrho}{2}\left\|\boldsymbol{\iota}' - \boldsymbol{\iota}\right\|_2^2. \tag{26}$$

Under Assumption 3, we could derive a useful result [29] as follows:

$$\left\|\nabla L\left(\boldsymbol{\iota}\right)\right\|_2^2 \geqslant 2\varrho\left[L\left(\boldsymbol{\iota}\right) - L\left(\boldsymbol{\iota}^*\right)\right]. \tag{27}$$

We state the convergence theorem of the DP-OTA-FedAvg, describing its behavior when minimizing a strongly convex objective function with a fixed learning rate in the following.

Theorem 1. *Assume that $\boldsymbol{m}^*$ is the optimal model and $\boldsymbol{m}^I$ is the obtained model after I communication rounds. Assume that the learning rate is $\tau = \frac{1}{\zeta}$, then, the upper bound of the optimality gap $\mathbb{E}\left[L\left(\boldsymbol{m}^I\right) - L\left(\boldsymbol{m}^*\right)\right]$ is given by*

$$\begin{aligned}&\mathbb{E}\left[L\left(\boldsymbol{m}^I\right) - L\left(\boldsymbol{m}^*\right)\right] \leq \eta^I \underbrace{\mathbb{E}\left[L\left(\boldsymbol{m}^0\right) - L\left(\boldsymbol{m}^*\right)\right]}_{Initial\ gap} \\ &+ \frac{\varpi^2}{\varrho}\left(1-\eta^I\right)\left[\underbrace{4\left(1 - \frac{|\mathcal{K}|}{N}\right)^2}_{\mathcal{A}} + \underbrace{\left(E-1\right)^2}_{\mathcal{B}} + \underbrace{\frac{1}{2}\frac{d\sigma^2}{|\mathcal{K}|^2\theta^2}}_{\mathcal{C}}\right],\end{aligned} \tag{28}$$

where $\eta = 1 - \frac{\varrho}{\zeta}$. The expectation is with respect to the randomness of Gaussian noise.

Proof: Please refer to Appendix B. ■

The optimality gap presented in the right-hand side (RHS) of (28) demonstrates the impact of device scheduling $\mathcal{K}$, alignment factor θ, and the local training times E on the learning process. Specifically, term $\mathcal{A}$ is the error caused by partial device participation. A larger $|\mathcal{K}|$ contributes to a smaller optimality gap, i.e., a better learning performance. This can be understood that the channel noise leads to a smaller distortion to the gradient average when more devices are involved. This term decreases as the number of the scheduled devices increases and will be eliminated with full device participation, i.e., $|\mathcal{K}| = N$. The local update error shown in term $\mathcal{B}$ increases with the number of local training times E. If $E = 1$, i.e., the FedAvg becomes the conventional FL algorithm, this term goes to 0. Term $\mathcal{C}$ is the error caused by the channel noise, which can be controlled by designing the device scheduling and the alignment coefficient. From this term, we can learn that a larger number of participants and the alignment coefficient contribute to a smaller noise-related error.

Furthermore, Theorem 1 offers the following important insights: (1) The impact of the number of the communication round I: the first term decreases with the number of communication round I due to the fact that $\eta \leq 1$. When I goes to infinity, the first term approaches zero. For the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

second term, on one hand, a larger I leads to a smaller E, which is beneficial for mitigating the local training error. On the other hand, a larger I increases the weight of the design-related error, i.e., term $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$.

Based on Theorem 1, we can also derive the optimality gap of a conventional FL algorithm where the number of local training times is one with full device participation and a noise-free channel.

Corollary 1. *Given the total training number is T and learning rate $\tau = \frac{1}{\zeta}$, the upper bound of the optimality gap $\mathbb{E}\left[L\left(\boldsymbol{m}^T\right) - L\left(\boldsymbol{m}^*\right)\right]$ of an conventional FL algorithm with one local training round without considering noise and device scheduling is*

$$\mathbb{E}\left[L\left(\boldsymbol{m}^T\right) - L\left(\boldsymbol{m}^*\right)\right] \leq \left(1 - \frac{\varrho}{\zeta}\right)^T \mathbb{E}\left[L\left(\boldsymbol{m}^0\right) - L\left(\boldsymbol{m}^*\right)\right]. \tag{29}$$

Proof: If the FL algorithms with full device participation only perform one local training round in each communication round and communicate through the noiseless channel, we have $E = 1$, $|\mathcal{K}| = N$, and $\sigma = 0$. Hence, $4\left(1 - \frac{|\mathcal{K}|}{N}\right)^2 + (E-1)^2 + \frac{1}{2}\frac{d\sigma^2}{|\mathcal{K}|^2\theta^2} = 0$. Then (29) can be derived based on (28). ■

From Corollary 1, we can observe that, if the communication between the BS and devices is noise-free and there is only one local training in each communication round, the FL algorithm with all device participation will converge to the optimal global FL model without any gaps. This result corresponds to the results in the existing works [30], [31].

2) Non-Convex Setting: Considering that many useful machine learning models, e.g., deep neural networks, lead to non-convex objective functions, we thus investigate the convergence property of DP-OTA-FedAvg in the non-convex setting in the following. Different from the convex case where the expected optimality gap is employed to measure the convergence rate. In the case of non-convex loss function $L(\cdot)$, the algorithm converging to a global minimum cannot in general be guaranteed. A reasonable substitute is to study the convergence to a local minimum, or at the very least, to stationary points [32], [33]. In Theorem 2, we adopt the average norm of the gradients as the convergence indicator, which is widely used in the convergence analysis for non-convex loss [34]–[38]. Note that the FL algorithm achieves an δ-approximation solution if

$$\frac{1}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^i\right)\right\|_2^2\right] \leq \delta. \tag{30}$$

Theorem 2. *Given the learning rate τ and the communication rounds I, the average-squared gradient after I communication rounds is bounded as follows,*

$$\frac{1}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^i\right)\right\|_2^2\right] \leq \frac{2}{\tau I}\left[\mathbb{E}\left[L\left(\boldsymbol{m}^0\right)\right] - \left[L\left(\boldsymbol{m}^*\right)\right]\right] + \varpi^2\left[8\left(1 - \frac{|\mathcal{K}|}{N}\right)^2 + 2\left(E-1\right)^2 + \frac{d\sigma^2}{|\mathcal{K}|^2\theta^2}\right]. \tag{31}$$

The expectation is with respect to the randomness of Gaussian noise.

Proof: Please refer to Appendix C. ■

In Theorem 2, we establish an upper bound of the average-squared gradients of $L(\cdot)$ for a certain communication round number I, which guarantees convergence of the FL algorithm to a stationary point [29], [34]. According to [29], the larger this upper bound is, the more communication rounds are required for convergence. We can boost the convergence rate by optimizing the device scheduling, alignment factor, and aggregation rounds.

IV. DP-OTA-FedAvg with Limited Sum Power Budgets

In order to improve the learning performance of DP-OTA-FedAvg with privacy and power constraints, we formulate the following problem where we take the optimality gap $\mathbb{E}\left[L\left(\boldsymbol{m}^I\right) - L\left(\boldsymbol{m}^*\right)\right]$ as the objective function. To minimize the objective function, we expect a larger $|\mathcal{K}|$ and a larger θ. However, θ is limited by the device in $\mathcal{K}$ with the worst channel condition. We can improve θ by scheduling the devices with better channel conditions to participate in the training, which leads to a smaller $|\mathcal{K}|$. Therefore, there is a tradeoff between the number of the scheduled devices $|\mathcal{K}|$ and the alignment factor θ. On the other hand, the impact of the global aggregation parameter I is significant. A larger I, corresponding to a smaller E, can help mitigate the local update error and reduce the initial gap. However, it may also introduce more transmission distortion. As a result, there exists a balance between reducing the local training error and increasing the aggregation error. Thus, selecting an optimal value for I that effectively balances the trade-off is of paramount importance. Overall, the design of device scheduling, alignment factor, and the number of global aggregations is crucial for enhancing learning performance while maintaining privacy.

A. Problem Formulation

Assume that each device has the same privacy budget (ϵ, ζ), i.e., the maximum value of tolerable privacy leakage. The total training rounds is T and we use $\frac{T}{I}$ to substitute E for simplicity. The number of global aggregations I and local training times E should be an integer. We firstly ignore the integer constraint of E, which will finally be guaranteed by rounding operation. By defining $G = \mathbb{E}\left[L\left(\boldsymbol{m}^0\right)\right] - \left[L\left(\boldsymbol{m}^*\right)\right]$, $\phi = \sqrt{2\ln\frac{1.25}{\xi}}$, and $\mathcal{W}(\mathcal{K}, \theta, I) =$

$\eta^I G + \frac{\varpi^2}{\varrho}\left(1-\eta^I\right)\left[4\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\left(\frac{T}{I}-1\right)^2+\frac{1}{2}\frac{d\sigma^2}{(|\mathcal{K}|\theta)^2}\right]$, the problem is formulated as follows.

$$\mathbf{P1}. \quad \min_{\mathcal{K},\theta,I}\left\{\mathcal{W}\left(\mathcal{K},\theta,I\right)\right\} \tag{32}$$

$$\mathbf{s.t.} \quad \mathcal{K}\subseteq\mathcal{N}, \tag{32a}$$

$$\frac{2\theta}{\sigma}\cdot\phi\leq\epsilon, \tag{32b}$$

$$0\leq\theta\leq\min_{s\in\mathcal{K}}\left\{\left|h_s\right|\sqrt{P_s}\right\}, \tag{32c}$$

$$I\sum_{k\in\mathcal{K}}\frac{\theta^2}{\left|h_k\right|^2}\leq P^{tot}, \tag{32d}$$

$$1\leq I\leq T, I\in\mathcal{Z}, \tag{32e}$$

where $\mathcal{Z}$ denotes the set of natural numbers. Constraint (32a) guarantees that $|\mathcal{K}|\leq N$. Constraint (32b) ensures that the privacy leakage of each device does not exceed the privacy budget. By replacing ν with $=\frac{\theta}{\varpi}$ in (15), we obtain constraint (32c) implying that the alignment coefficient should ensure that $\varphi_k\leq 1$ due to the peak power constraint. Constraint (32d) is the sum power constraint. Constraint (32e) implies that the number of the aggregation rounds should be an integer and no more than T.

P1 is solved in the following way. First, we decouple P1 into two sub-problems. Given the number of aggregation rounds, the potential optimal solution pairs for $\mathcal{K}$ and θ can be obtained by exploring the relationship between the number of scheduled devices and the alignment coefficient. Thanks to the reduced search space, the globally optimal $\mathcal{K}$ and θ can be efficiently found by searching the set of the limited solution pairs. Furthermore, based on the optimal scheduling and alignment factor, the optimal aggregation rounds can be obtained by the one-dimensional search.

B. Optimal Device Scheduling and Alignment Factor

Assume that the number of the aggregation rounds is I. We define $q_{[\mathcal{K}]}=\sqrt{\frac{P^{tot}}{I}}\left(1/\sqrt{\sum_{k\in\mathcal{K}}\left(1/\left|h_k\right|^2\right)}\right)$ and $c_{[\mathcal{K}]}=\min_{s\in\mathcal{K}}\left\{\left|h_s\right|\sqrt{P_s}\right.$ for notation simplicity. Then, the constraints (32b), (32c) and (32d) can be rewritten as $0\leq\theta\leq\min\left\{\frac{\epsilon\sigma}{2\phi},c_{[\mathcal{K}]},q_{[\mathcal{K}]}\right\}$. The problem that optimizes device scheduling $\mathcal{K}$ and alignment factor θ can be decoupled as follows:

$$\mathbf{P2}. \quad \min_{\mathcal{K},\theta}\left\{4\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\frac{1}{2}\frac{d\sigma^2}{|\mathcal{K}|^2\theta^2}\right\} \tag{33}$$

$$\mathbf{s.t.} \quad \mathcal{K}\subseteq\mathcal{N}, \tag{33a}$$

$$0\leq\theta\leq\min\left\{\frac{\epsilon\sigma}{2\phi},c_{[\mathcal{K}]},q_{[\mathcal{K}]}\right\}. \tag{33b}$$

By observing the objective function, we know that larger $|\mathcal{K}|$ and θ yield a better objective function value. However, the upper bound of θ is limited by the scheduling policy $\mathcal{K}$ as shown in constraint (33b). To this end, we first analyze the relationship between the number of scheduled devices $|\mathcal{K}|$ and the alignment factor θ, which can offer us a set of potential optimal solution pairs.

Assume that the devices are sorted in descending order of $|h_k|$, i.e., $|h_1|\geq|h_2|\geq\cdots\geq|h_N|$. In order to specify the relationship between $|\mathcal{K}|$ and θ, we first conclude the relationship between $|\mathcal{K}|$ and the upper bounds of $c_{[\mathcal{K}]}$ and $q_{[\mathcal{K}]}$, which limits the value of θ. For ease of presentation, we introduce the notation $\mathcal{S}\left(\mathcal{N};q\right)$, representing a mechanism that returns a list where the elements in set $\mathcal{N}$ are arranged in a descending order based on the parameter q.

Lemma 3. *Assume that $c_{|\mathcal{K}|}^{max}$ and $q_{|\mathcal{K}|}^{max}$ are the achieveable upper bounds of $c_{[\mathcal{K}]}$ and $q_{[\mathcal{K}]}$ for a given $|\mathcal{K}|$, then, we have*

$$c_{|\mathcal{K}|}^{max}=\left|h_{\mathbf{n}_c[|\mathcal{K}|-1]}\right|\sqrt{P_{\mathbf{n}_c[|\mathcal{K}|-1]}}, \tag{34}$$

when $\mathcal{K}=\mathcal{K}_c=\left\{\mathbf{n}_c\left[k\right]|k<|\mathcal{K}|,k\in\mathcal{Z}\right\}$ and $\mathbf{n}_c=\mathcal{S}\left(\mathcal{N};|h_k|\sqrt{P_k}\right)$. $q_{|\mathcal{K}|}^{max}$ can be achieved when $\mathcal{K}=\mathcal{K}_q=\{k+1|k<|\mathcal{K}|,k\in\mathcal{Z}\}$, and can be given by

$$q_{|\mathcal{K}|}^{max}=\sqrt{\frac{P^{tot}}{I}}\left(1/\sqrt{\sum_{j=1}^{|\mathcal{K}|}\left(1/\left|h_j\right|^2\right)}\right). \tag{35}$$

Proof: To achieve the maximum value of $c_{[\mathcal{K}]}$ for a given $|\mathcal{K}|$, the $|\mathcal{K}|$ devices with largest $|h_k|\sqrt{P_k}$ should be scheduled as $\mathcal{K}$, i.e., $\mathcal{K}=\mathcal{K}_c=\{\mathbf{n}_c\left[k\right]|k<|\mathcal{K},k\in\mathcal{Z}|\}$. At this point, we achieve the upper bound of $c_{[\mathcal{K}]}$, i.e., $c_{|\mathcal{K}|}^{max}=\left|h_{\mathbf{n}_c[|\mathcal{K}|-1]}\right|\sqrt{P_{\mathbf{n}_c[|\mathcal{K}|-1]}}$. Similarly, the devices with larger $|h_k|$ contribute to a larger $q_{[\mathcal{K}]}$. The maximum value of $q_{|\mathcal{K}|}$ is achieved, i.e., $q_{|\mathcal{K}|}^{max}=\sqrt{\frac{P^{tot}}{I}}\left(1/\sqrt{\sum_{j=1}^{|\mathcal{K}|}\left(1/\left|h_j\right|^2\right)}\right)$ when $\mathcal{K}=\mathcal{K}_q=\{k+1|k<|\mathcal{K}|,k\in\mathcal{Z}\}$. ■

Remark 2. *If all the devices are with the same peak transmit power budget P^{dev}, i.e., $P_1=P_2=\cdots=P_N=P^{dev}$, $\mathcal{K}_c=\mathcal{K}_q$.*

Lemma 3 provides valuable insights, indicating that when the value of $|\mathcal{K}|$ is given, it is possible to determine the potential optimal scheduling sets, denoted as $\mathcal{K}_c$ and $\mathcal{K}_q$, which can maximize $c_{[\mathcal{K}]}$ and $q_{[\mathcal{K}]}$, respectively. Consequently, it offers a potential upper bound for θ. In other words, each given $|\mathcal{K}|$ corresponds one potential optimal set, i.e., $\mathcal{K}_c$ or $\mathcal{K}_q$, which can determine the upper bound of θ, i.e., $\min\{\frac{\epsilon\sigma}{2\phi},c_{[\mathcal{K}_c]},q_{[\mathcal{K}_c]}\}$ or $\min\{\frac{\epsilon\sigma}{2\phi},c_{[\mathcal{K}_q]},q_{[\mathcal{K}_q]}\}$. The potential optimal set, denoted as $\mathcal{K}_c$ or $\mathcal{K}_q$, along with the associated upper bound of θ, forms what we refer to as the potential optimal solution pair. These pairs exhibit a one-to-one correspondence with the number of scheduled devices. In other words, each specific number of scheduled devices determines a potential optimal solution pair. The maximum number of potential optimal solution pairs is N, i.e., the value of $|\mathcal{K}|$ is set from 1 to N. However, when θ is fixed, we have the following results.

Corollary 2. *If θ is given, the optimal $\mathcal{K}$ to $\mathbf{P}2$ is obtained by $\mathcal{K}=\{m|\min\{c_m,q_m\}>=\theta,m\in\mathcal{N}\}$, where $c_m=|h_m|\sqrt{P_m}$ and $q_m=\sqrt{\frac{P^{tot}}{I}}\left(1/\sqrt{\sum_{j=1}^{m}\left(1/\left|h_j\right|^2\right)}\right)$.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

Corollary 2 emphasizes that for each given θ, there exists an optimal $\mathcal{K}$, which leads to a optimal value of $|\mathcal{K}|$. Taking into account the constraint that $\theta \leq \frac{\epsilon\sigma}{2\phi}$, we can further refine the set of potential solution pairs as outlined below.

We define $\boldsymbol{c} = [c_1, \cdots c_m, \cdots c_N]$ and $\boldsymbol{q} = [q_1, \cdots q_m, \cdots q_N]$. Since $|h_1| \geq |h_2| \geq \cdots \geq |h_N|$, the elements in $\boldsymbol{q}$ are sorted in the descending order. However, the elements in $\boldsymbol{c}$ may not be. We define $\boldsymbol{c}^s = [c_1^s, \cdots c_m^s, \cdots c_N^s]$ as the list where the elements of $\boldsymbol{c}$ are sorted in the descending order. Therefore, the minimal values of $q_{[\mathcal{K}]}$ and $c_{[\mathcal{K}]}$ are $q_N = \sqrt{\frac{P^{tot}}{I}}\left(1/\sqrt{\sum_{j=1}^{N}\left(1/|h_j|^2\right)}\right)$ and c_N^s, in which case $\mathcal{K} = \mathcal{N}$. For clarity, we give the solutions in two cases: 1) $\frac{\epsilon\sigma}{2\phi} < \min\{c_N^s, q_N\}$; 2) $\min\{c_N^s, q_N\} \leq \frac{\epsilon\sigma}{2\phi}$ as follows.

1) In the case that $\frac{\epsilon\sigma}{2\phi} < \min\{c_N^s, q_N\}$: The constraint of θ is independent of the device scheduling $\mathcal{K}$. Constraint (33b) can be rewritten as $0 \leq \theta \leq \frac{\epsilon\sigma}{2\phi}$. Then, the optimal solution to $\mathbf{P}2$ can be given by the following Lemma.

Lemma 4. *If $\frac{\epsilon\sigma}{2\phi} < \min\{c_N^s, q_N\}$, the optimal solution to $\mathbf{P}2$ is*

$$\theta^* = \frac{\epsilon\sigma}{2\phi}, \quad \mathcal{K}^* = \mathcal{N}, \tag{36}$$

in which case all the devices are scheduled.

Proof: Firstly, to achieve a larger θ, we have $\theta^* = \frac{\epsilon\sigma}{2\phi}$. On the other hand, all the devices satisfying $c_m \geq \frac{\epsilon\sigma}{2\phi}$ and $q_m \geq \frac{\epsilon\sigma}{2\phi}$ should be selected to achieve a larger $|\mathcal{K}|$, i.e., a better value of objective function. Since $\frac{\epsilon\sigma}{2\phi} < c_N^s \leq c_m$ and $\frac{\epsilon\sigma}{2\phi} < q_N \leq q_m, \forall m \in \mathcal{N}$, we have $\mathcal{K}^* = \mathcal{N}$. This completes the proof of Lemma 4. ■

2) In the case that $\min\{c_N^s, q_N\} \leq \frac{\epsilon\sigma}{2\phi}$: The constraint of θ is related to the device scheduling $\mathcal{K}$. Let $w_m = \min\{c_m, q_m\}$ and we assume that $\mathcal{Q} = \mathcal{Q}_c \cup \mathcal{Q}_q$ where $\mathcal{Q}_c = \left\{m | c_m < \frac{\epsilon\sigma}{2\phi}, m \in \mathcal{N}\right\}$ and $\mathcal{Q}_q = \left\{m | q_m < \frac{\epsilon\sigma}{2\phi}, m \in \mathcal{N}\right\}$. The elements in $\mathcal{Q}$ represents the indexes of devices satisfying $w_m < \frac{\epsilon\sigma}{2\phi}$. We define $\mathbf{s} = \mathcal{S}(\mathcal{N}; w_m)$ denoting the list where the elements in set $\mathcal{N}$ are arranged in a descending order based on the parameter w_m, i.e., $w_{\mathbf{s}[0]} \geq \cdots \geq w_{\mathbf{s}[N-|\mathcal{Q}|-1]} \geq \frac{\epsilon\sigma}{2\phi} > w_{\mathbf{s}[N-|\mathcal{Q}|]} \geq \cdots \geq w_{\mathbf{s}[N-1]}$ as shown in Fig. 2. Then, constraint (33b) can be discussed in two cases: (1) $0 \leq \theta \leq w_m, m \in \mathcal{Q}$; (2) $w_{\mathbf{s}[N-|\mathcal{Q}|]} < \theta \leq \frac{\epsilon\sigma}{2\phi}$. Therefore, there are $|\mathcal{Q}|+1$ potential upper bounds of θ. For each upper bound of θ, there is a corresponding optimal $|\mathcal{K}|$ following Corollary 2. To capture the trade-off between the alignment factor θ and the number of the scheduled devices $|\mathcal{K}|$, we have the following results.

Lemma 5. *The minimum value of the potential optimal $|\mathcal{K}|$ is $N - |\mathcal{Q}|$. The relationship between the potential optimal $|\mathcal{K}|$ and θ can be given by*

$$\theta = \begin{cases} w_{\mathbf{s}[|\mathcal{K}|-1]}, if\ |\mathcal{K}| \geq N - |\mathcal{Q}| + 1 \\ \frac{\epsilon\sigma}{2\phi}, if\ |\mathcal{K}| = N - |\mathcal{Q}| \end{cases}. \tag{37}$$

Proof: Firstly, given a value of $|\mathcal{K}| \geq N - |\mathcal{Q}| + 1$, the largest w_m that can be achieved is $w_{\mathbf{s}[|\mathcal{K}|-1]}$. To achieve a larger θ, we have $\theta = w_{\mathbf{s}[|\mathcal{K}|-1]}$. The largest feasible value of θ is $\frac{\epsilon\sigma}{2\phi}$, in which $|\mathcal{K}|$ achieves the minimum value $N - |\mathcal{Q}|$. Then, we complete the proof of Lemma 5. ■

Remark 3. *Although values in the $\{1, \cdots, N - |\mathcal{Q}| - 1\}$ are technically feasible for $|\mathcal{K}|$, it will not lead to the optimal solution as larger $|\mathcal{K}|$ values usually results in smaller objective function values, which aligns with our goal of obtaining optimal solutions.*

Lemma 5 captures a trade-off between the alignment factor θ and the number of the scheduled devices $|\mathcal{K}|$. In particular, the trade-off involves the choice between scheduling more devices to boost the learning process with a lower alignment factor θ, or scheduling fewer devices to attain a larger alignment factor, thereby enhancing the utility of the aggregated gradient. For instance, in the scenario where $|\mathcal{K}| = N$, indicating the participation of all devices in training, θ will achieve the minimal value of the potential optimal solution, i.e., $\theta = w_{\mathbf{s}[N-1]}$. On the contrary, θ can achieve the largest value by scheduling the device with the largest w_m, i.e., $w_{\mathbf{s}[0]}$ when $|\mathcal{K}| = 1$. From a physical perspective, when more devices are involved in training, there is a higher likelihood of including devices with poorer channel conditions, and this also leads to stricter total power constraints. The trade-off is validated by Fig. 3 in Section V-B.

The space of the potential optimal solutions pairs to $\mathbf{P}2$ can be given as follows.

Lemma 6. *There are $|\mathcal{Q}| + 1$ closed-form solution pairs which may be the globally optimal solution. The j-th, $1 \leq j \leq |\mathcal{Q}|$, potential optimal solution pair θ_j^* and $\mathcal{K}_j^*$ is given by*

$$\theta_j^* = w_{\mathbf{s}[N-j]}, \quad \mathcal{K}_j^* = \{\mathbf{s}[m] | m \leq N - j, m \in \mathcal{Z}\}, \tag{38}$$

and the $|\mathcal{Q}| + 1$-th solution pair $\theta_{|\mathcal{Q}|+1}^*$, $\mathcal{K}_{|\mathcal{Q}|+1}^*$ is

$$\theta_{|\mathcal{Q}|+1}^* = \frac{\epsilon\sigma}{2\phi}, \quad \mathcal{K}_{|\mathcal{Q}|+1}^* = \{\mathbf{s}[m] | m \leq N - |\mathcal{Q}| - 1, m \in \mathcal{Z}\}. \tag{39}$$

Proof: Firstly, there are $|\mathcal{Q}|$ elements in $\mathcal{Q}$, which are qualified for the upper bound of θ i.e., w_m. It thus follows from Lemma 5 that there are $|\mathcal{Q}|$ pairs of θ and $|\mathcal{K}|$, i.e., $|\mathcal{Q}|$ potential optimal solution pairs, which may achieve the best performance. Specifically, the j-th solution corresponds to the setting that $\theta_j^* = w_{\mathbf{s}[N-j]}$ and $|\mathcal{K}| = N - j + 1$, in which case, $\mathcal{K}_j^* = \{\mathbf{s}[m] | m \leq N - j, m \in \mathcal{Z}\}$. Additionally, $\theta_{|\mathcal{Q}|+1}^* = \frac{\epsilon\sigma}{2\phi}$ is the $|\mathcal{Q}| + 1$-th solution, the largest $|\mathcal{K}|$, i.e., $|\mathcal{K}| = N - |\mathcal{Q}| - 1$, can be achieved when $\mathcal{K}_{|\mathcal{Q}|+1}^* = \{\mathbf{s}[m] | m \leq N - |\mathcal{Q}| - 1, m \in \mathcal{Z}\}$, which can contribute to the optimal value of objective function. This completes the proof of Lemma 6. ■

The θ_j^* and $\mathcal{K}_j^*$ in Lemma 6 can be interpreted as representing the optimal solution when the number of scheduled devices, denoted as $|\mathcal{K}|$, is set to j. Based on Lemma 6, we

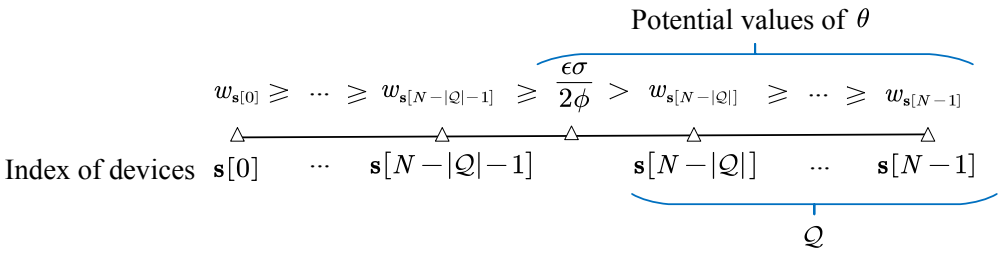


Fig. 2: Illustration of the potential optimal solution space.

can perform the one-dimension search method to obtain the globally optimal solution $\mathcal{K}^*, \theta^* = \mathcal{K}_{j^*}^*, \theta_{j^*}^*$, in which

$$j^* = arg \min_{1 \leq j \leq |\mathcal{Q}|+1} \left\{ \Psi\left(\mathcal{K}_j^*, \theta_j^*\right) \right\}, \tag{40}$$

where $\Psi\left(\mathcal{K}_j^*, \theta_j^*\right) = 4\left(1 - \frac{|\mathcal{K}_j^*|}{N}\right)^2 + \frac{d\sigma^2}{2|\mathcal{K}_j^*|^2(\theta_j^*)2}$. The overall procedure for solving $\mathbf{P}2$ is summarized in Algorithm 1. As the optimal solution for θ and $\mathcal{K}$ is derived through an exhaustive search among $|\mathcal{Q}|+1 (\leq N)$ potential optimal solution pairs, given by a closed-form expression, the complexity of Algorithm 1 is denoted as $\mathcal{O}(N)$.

Algorithm 1 The Procedure for Solving Problem $\mathbf{P2}$

Input: Given N, T, d, σ, (ϵ, ξ), $\boldsymbol{h} = \{|h_1|, \cdots, |h_N|\}$, $\{P_n\}_{n=1}^N$ and P^{tot}. Let $I = T$.
Output: $\mathcal{K}^*, \theta^*$.

```
Calculate ε_σ/(2φ), c and q.
if ε_σ/(2φ) < min{c_N^s, q_N} then
    Obtain the optimal solution K* = N, θ* = ε_σ/(2φ) following Lemma 4.
else
    Calculate Q = Q_c ∪ Q_q where Q_c = {m | c_m < ε_σ/(2φ), m ∈ N} and Q_q = {m | q_m < ε_σ/(2φ), m ∈ N}.
    Obtain |Q| + 1 pairs of potential optimal solution following Lemma 6.
    Obtain the optimal solutions by K*, θ* = K*_{j*}, θ*_{j*} where j* = arg min_{1≤j≤|Q|+1} {Ψ(K*_j, θ*_j)}.
end if
```

From the above analysis, it can also be learned that the aligned DP-OTA-FL with device scheduling will not be worse than that with full devices participation because the full device scheduling is one of the potential optimal solution pairs. We next present which pairs of solutions can achieve better performance than the case of full device participation. Since the optimal solution is the same as full device situation when $\frac{\epsilon\sigma}{2\phi} < \min\{c_N^s, q_N\}$ as shown in Lemma 4, we only consider the case that $\min\{c_N^s, q_N\} \leq \frac{\epsilon\sigma}{2\phi}$.

Lemma 7. *If $\min\{c_N^s, q_N\} \leq \frac{\epsilon\sigma}{2\phi}$, the solution pairs $\mathcal{K}$ and θ that satisfies the following condition will make the aligned DP-OTA-FedAvg perform better than that with full device participation:*

$$|\mathcal{K}|\,\theta \geqslant \frac{1}{\sqrt{\frac{1}{N^2(w^{min})^2} - \frac{8}{d\sigma^2}}}, \tag{41}$$

where $w^{min} = \min\{c_N^s, q_N\}$.

Proof: The aligned DP-OTA-FedAvg with full device participation is equivalent to the solution that $\theta = w^{min}$ and $\mathcal{K} = \mathcal{N}$, in which case, the value of the objective function is $\frac{d\sigma^2}{N^2(w^{min})^2}$. By solving $4 + \frac{d\sigma^2}{2|\mathcal{K}|^2\theta^2} \leq \frac{d\sigma^2}{2N^2(w^{min})^2}$, we complete the proof of Lemma 7. ■

C. Optimal Number of Global Aggregation

By giving $\mathcal{K}$ and θ, the problem of the optimal number of the global aggregations can be formulated by,

$$\mathbf{P}3. \quad \min_{I} \{\mathcal{W}(\mathcal{K}, \theta, I)\} \tag{42}$$

$$\mathbf{s.t.} \quad 1 \leq I \leq \min\left\{\frac{P^{tot}}{\theta^2 \sum_{k\in\mathcal{K}} \frac{1}{|h_k|^2}}, T\right\}, \quad I \in \mathcal{Z}. \tag{42a}$$

Since there are only limited feasible solutions of I, the optimal number of the aggregation rounds I^* can be efficiently obtained by searching the solution space.

D. The Whole Precedure of the Optimal Design for DP-OTA-FedAvg

In this subsection, we present the overall procedure for design the optimal DP-OTA-FedAvg (O-DP-OTA-FedAvg) as shown in Algorithm 2. Suppose that the algorithm of optimal design needs M iterations to converge. In each iteration, it obtains θ, $\mathcal{K}$ with a complexity of $\mathcal{O}(N)$ and I with a complexity of $\mathcal{O}(T)$. Therefore, the overall complexity of the proposed O-DP-OTA-FedAvg is $\mathcal{O}(MN + MT)$.

V. Simulation Results

A. Simulation Setting

We evaluate our proposed scheme by training a convolutional neural network (CNN) on the popular MNIST dataset used for handwritten digit classification. The MNIST dataset consists of 60,000 images for training and 10,000 testing images of the 10 digits. We have the general assumption

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

Algorithm 2 Optimal Design for DP-OTA-FedAvg

Require: Given N, T d, σ, (ϵ, ξ), $\boldsymbol{h} = \{|h_1|, \cdots, |h_N|\}$, $\{P_n\}_{n=1}^N$ and P^{tot}.

1: Set the iteration number $j = 0$ and $I_0 = T$.
2: **repeat**
3: Obtain $\mathcal{K}^*_{(j)}$, $\theta^*_{(j)}$ by Algorithm 1.
4: Compute $I^*_{(j)}$ by solving P3.
5: $j \leftarrow j + 1$.
6: **until** the convergence condition $\left|\mathcal{W}\left(\mathcal{K}^*_{(j)}, \theta^*_{(j)}, I^*_{(j)}\right) - \mathcal{W}\left(\mathcal{K}^*_{(j-1)}, \theta^*_{(j-1)}, I^*_{(j-1)}\right)\right| \leq \varepsilon$ is satisfied.

that there is an equal number of training data samples for each device and no overlap between the local training data sets [19]. We have the common assumption that each device has an equal amount of training data samples and the local training datasets are non-overlapping with each other [19]. We assume that local datasets are IID, where the initial training dataset is randomly divided into N batches and each device is assigned to one batch. In particular, CNN consists of two 5×5 convolution layers with the rectified linear unit (ReLU) activation. The two convolution layers have 10 and 20 channels respectively, and each layer has 2×2 max pooling, a fully-connected layer with 50 units and ReLU activation, and a log-softmax output layer, in which case $d = 21840$. We have a total of $N = 50$ devices, each with a maximum transmission budget of 1 watt. The learning rate is configured with a value of $\eta = 0.1$ and the entire training process consists of $T = 500$ rounds. The privacy requirement is defined as $(\epsilon, \xi) = (10, 0.1)$ and the upper bound of the expected squared norm of each gradient is given as $\varpi = 3$, which can be gaurantted by gradient clipping [39], [40]. The channel noise power is designated as $\sigma = 1$ and we further introduce the notation $h_{\min} = \min_{n \in \mathcal{N}} \{|h_n|\}$ to represent the minimum channel coefficient. This enables us to assess the performance of the proposed scheme across various channel conditions in different systems.

B. The Optimal Alignment Factor and Scheduling

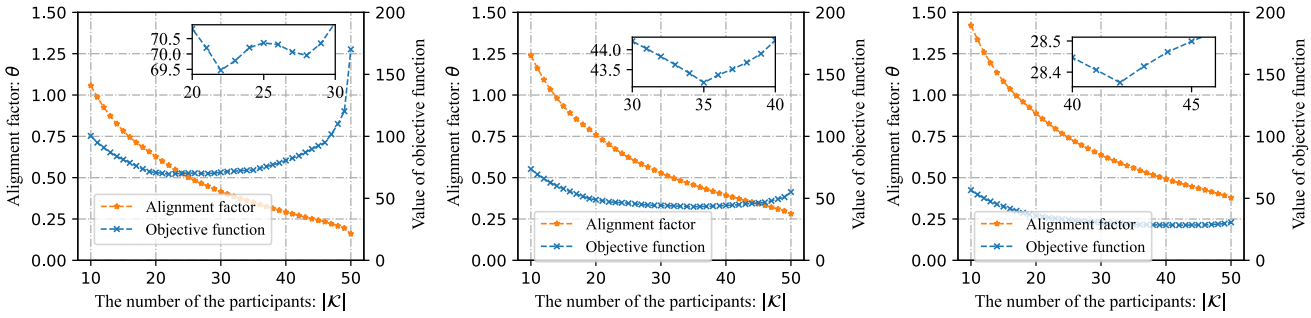
In this section, we first illustrate the trade-off between the alignment factor θ and the number of scheduled devices $|\mathcal{K}|$, and evaluate the effectiveness of the proposed optimal solution for the alignment factor and scheduling while setting $I = T$. The sum-power budget is set to $P^{tot} = 1000W$. Given the one-to-one relationship between the optimal alignment factor and optimal scheduling, for the sake of simplicity, we will refer to the optimal solution for both as "optimal scheduling" in the following discussions.

1) The tradeoff between the potential optimal alignment factor θ and the number of the scheduled devices $|\mathcal{K}|$: In Fig. 3, we plot the alignment factor θ and the value of the objective function in $\mathbf{P}2$ against the number of the number of the scheduled devices $|\mathcal{K}|$ in cases of $h_{\min} = 0.2$, $h_{\min} = 0.6$ and $h_{\min} = 1.0$. It is evident that the alignment factor decreases as the number of participants increases, and there exists an optimal number of scheduled devices that minimizes the objective function's value, for given values of N and I. To elaborate, scheduling more devices accelerates convergence but may result in a lower alignment factor θ, potentially negatively impacting learning performance. The optimal value of $|\mathcal{K}|$ tends to increase with growing $h_{\min}$ as scheduling more devices does not necessarily limit the alignment factor to very small values, as demonstrated in Fig. 3(c).

In Table I, we provide a comprehensive overview of the optimal and full scheduling solutions, along with the corresponding objective function values for $\mathbf{P}2$ across varying $h_{\min}$ values. These solutions are denoted as $(|\mathcal{K}|^*, \theta^*, obj^*)$ and $\left(|\mathcal{K}|^f, \theta^f, obj^f\right)$, respectively. By comparing the optimal and full scheduling solutions for each $h_{\min}$, we observe that the proposed optimal scheduling policy yields a lower value of the objective function. This achievement is realized by scheduling a reduced number of devices to enhance the alignment factor θ. Therefore, this approach is particularly effective when $h_{\min}$ is relatively small. In such cases, the alignment factor θ in the full scheduling policy is limited by the device with the poorest channel condition, denoted as $h_{\min}$, when all the devices are scheduled. As $h_{\min}$ increases, the gaps in the number of the participants $|\mathcal{K}|$, the alignment factor θ, and the objective function value between the optimal and the full scheduling solutions gradually narrow, eventually reaching 0 when $h_{\min} = 2.8$. This is because when the alignment factor is no longer restricted to very small values due to the minimum channel conditions, scheduling more devices becomes advantageous for improving learning performance, which is consistent with the conclusion obtained in Fig. 3.

2) The optimal alignment factor θ and scheduling policy $\mathcal{K}$: In Fig. 4, we plot the learning accuracy with four different scheduling policies: optimal scheduling, full scheduling [10], [13], [25], uniform scheduling (where the number of scheduled devices remains consistent with the optimal scheduling and the participants are uniformly selected from $\mathcal{N}$), and random-θ-based scheduling [26] (where the devices are scheduled based on a random alignment factor which is set to 1). The proposed optimal scheduling policy consistently outperforms full scheduling, uniform scheduling, and random-θ-based scheduling in all scenarios. This performance advantage becomes particularly significant when $h_{\min} = 0.4$, as the alignment factor is severely restricted in the full scheduling scheme. Random-θ-based scheduling attempts to enhance the alignment factor by setting a threshold for scheduling devices but does not effectively strike a balance between the alignment factor and the number of scheduled devices. Despite having an equal number of scheduled devices in both optimal scheduling and uniform scheduling, the optimal scheduling policy achieves a larger value of alignment factor by scheduling the $|\mathcal{K}^*|$ devices with the best conditions. Overall, the results demonstrate the pronounced effectiveness of the proposed optimal scheduling policy, especially in the case of FL systems with poor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications

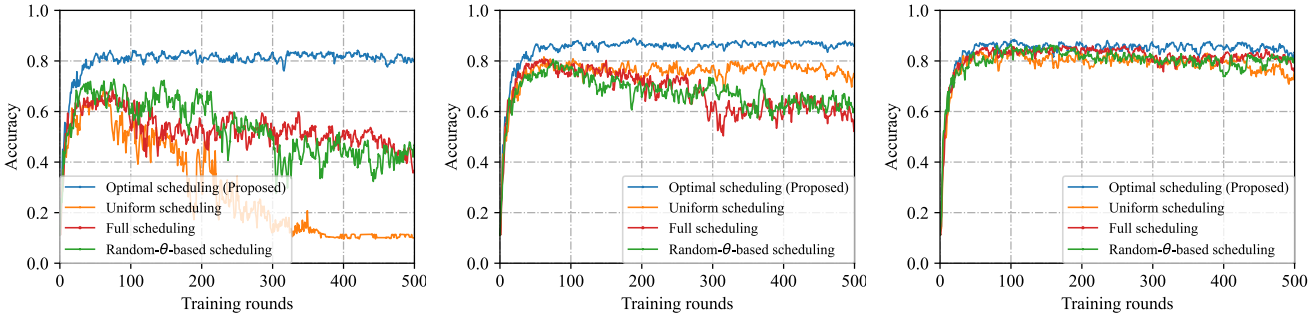


(a) $h_{\min} = 0.2$ (b) $h_{\min} = 0.6$ (c) $h_{\min} = 1.0$

Fig. 3: The alignment factor θ and the value of the objective function against $|\mathcal{K}|$.

TABLE I: The optimal and the full scheduling solutions with different $h_{\min}$.

$h_{\min}$	0.2	0.4	0.6	1.0
$(\lvert\mathcal{K}\rvert^*, \theta^*, obj^*)$	$(22, 0.5751, 69.4774)$	$(35, 0.4041, 54.9602)$	$(35, 0.4563, 43.1706)$	$(42, 0.4680, 28.3718)$
$\left(\lvert\mathcal{K}\rvert^f, \theta^f, obj^f\right)$	$(50, 0.1602, 170.1106)$	$(50, 0.2277, 84.2273)$	$(50, 0.2817, 55.0454)$	$(50, 0.3778, 30.6066)$
$h_{\min}$	1.6	2.0	2.4	2.8
$(\lvert\mathcal{K}\rvert^*, \theta^*, obj^*)$	$(47, 0.5481, 16.4679)$	$(49, 0.6107, 12.1981)$	$(49, 0.6955, 9.4046)$	$(50, 0.7648, 7.4673)$
$\left(\lvert\mathcal{K}\rvert^f, \theta^f, obj^f\right)$	$(50, 0.5111, 16.7196)$	$(50, 0.5969, 12.2595)$	$(50, 0.6813, 9.4098)$	$(50, 0.7648, 7.4673)$



(a) $h_{\min} = 0.4$ (b) $h_{\min} = 0.6$ (c) $h_{\min} = 0.8$

Fig. 4: The learning performance with different scheduling policies

channel conditions.

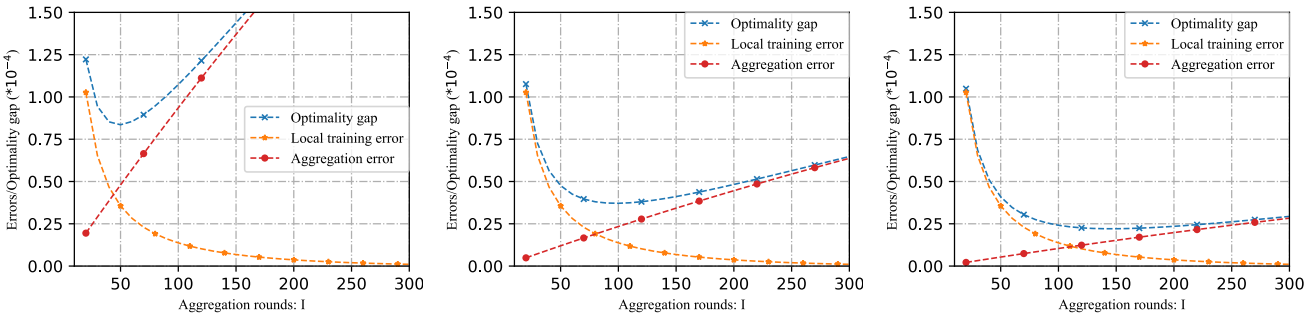
C. The Optimal Aggregation Rounds

In this section, we first illustrate the tradeoff between local training error and the aggregation error and the impact of aggregation rounds on learning performance by fixing the participants as $\mathcal{N}$.

1) The tradeoff between local training error and the aggregation error: In Fig. 5, we plot the local training error and the aggregation error, as well as the value of optimality gap against the number of aggregation rounds in cases of $h_{\min} = 0.2$, $h_{\min} = 0.4$ and $h_{\min} = 0.6$. There is an optimal number of aggregation times I in terms of convergence performance for a given total training rounds T and P^{tot}. In more detail, a larger I leads to a smaller E, which corresponds to a smaller local training error, however, may involve more communication-related errors, including fading distortion and channel noise, and thus may have a negative impact on convergence performance. In this sense, there is a tradeoff in choosing a proper I. The optimal number of aggregation rounds increases as $h_{\min}$ decreases. This is because when $h_{\min}$ is small, the alignment factor tends to be small as well, resulting in significant noise distortion during each aggregation round. Therefore, in such cases, fewer aggregation rounds are preferable for better performance. Conversely, increasing the number of aggregation rounds reduces the local training error and can generally enhance convergence performance, provided that noise distortion is not relatively small (a larger θ).

2) The optimal aggregation rounds: In Fig. 6, we plot the learning accuracy with three different aggregation rounds: the optimal aggregation, per-round aggregation, i.e., $E = 1, I = T$, and the random aggregation rounds. The best

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications



(a) $h_{\min} = 0.2$ (b) $h_{\min} = 0.4$ (c) $h_{\min} = 0.6$

Fig. 5: The local training error, aggregation error and optimality gap against the number of aggregation rounds.

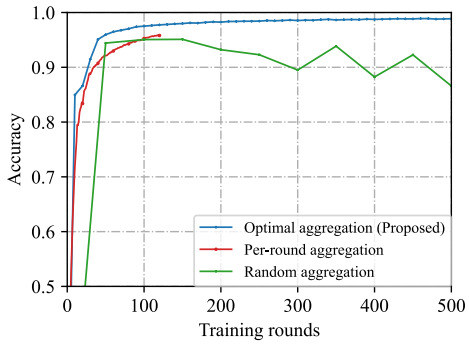


Fig. 6: The learning performance with different aggregation times.

learning performance is achieved when training with the optimal aggregation rounds. In the case of a per-round aggregation scheme, there is a limitation on the number of aggregation rounds due to the constrained sum-power for the given set $\mathcal{K}$. In the random aggregation, where a smaller value of I implies a larger E, local training may converge to the optimal solution of the local objective rather than the global objective. Therefore, it is significant to carefully select the number of aggregation rounds.

D. The Optimal Design for DP-OTA-FedAvg

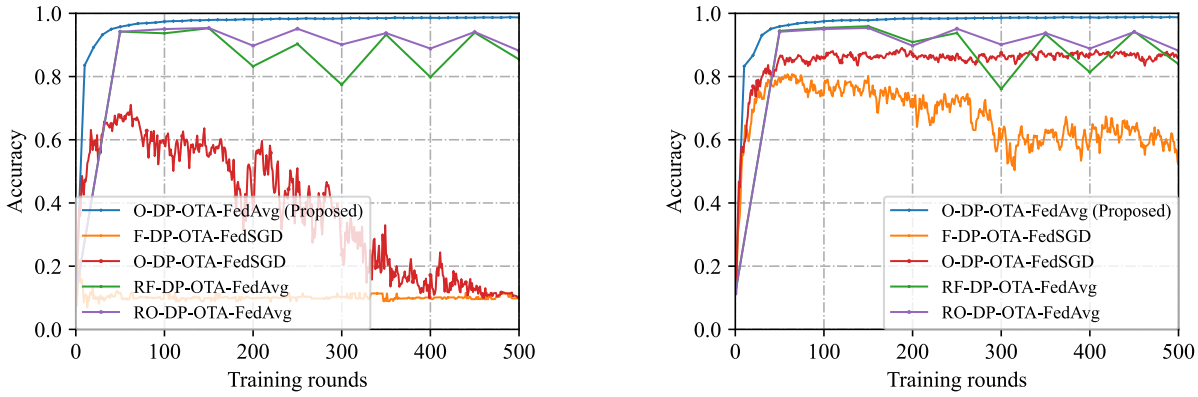
We evaluate the overall performance of O-DP-OTA-FedAvg under different sum power constraints. This assessment involves comparing O-DP-OTA-FedAvg to four benchmark schemes, which include: (1) differentially private over-the-air federated stochastic gradient descent (DP-OTA-FedSGD) with full device participation, denoted as F-DP-OTA-FedSGD, where $I = T$. (2) DP-OTA-FedSGD with optimal scheduling, denoted as O-DP-OTA-FedSGD. (3) DP-OTA-FedAvg with random aggregation rounds and full device participation, denoted as RF-DP-OTA-FedAvg. (4) DP-OTA-FedAvg with random aggregation rounds and optimal device scheduling, denoted as RO-DP-OTA-FedAvg.

Fig. 7 demonstrates that the proposed optimal design can significantly improve the performance of DP-OTA-FL. On the one hand, the optimal design of device scheduling and alignment coefficient enhances the learning performance by improving the utility of the aggregated gradient average in each communication round. On the other hand, O-DP-OTA-FedAvg can set more efficient local training rounds by optimizing the number of the global model aggregations under the constraint of limited sum power and privacy budget. The sum power constraint has a more pronounced impact on DP-OTA-FedSGDs when $E = 1$. In such cases, the power available for transmitting gradients during each communication round is limited, which, in turn, restricts either the alignment factor or the number of scheduled devices. In contrast, due to the reduced aggregation rounds, the sum power constraint has a relatively smaller effect on DP-OTA-FedAvgs, as neither the alignment factor nor the number of scheduled devices is significantly limited by the available power for transmission in each aggregation round. However, it performs worse than O-DP-OTA-FedAvg due to the presence of larger local training errors.

VI. Conclusion

This paper has studied an optimal design for device scheduling, alignment coefficient, and the number of communication rounds of DP-OTA-FedAvg considering constraints on the total available power. It has addressed two fundamental tradeoffs inherent to aligned OTA-FL when confronted with privacy and power limitations. Firstly, it has uncovered the tradeoff between the number of scheduled devices and the alignment coefficient, offering insights into the delicate balance between these factors. Secondly, it has explored the balance between aggregation distortion and local training error while maintaining compliance with the sum power constraint. The proposed optimal schemes can significantly enhance the performance of the privacy and power-limited DP-OTA-FedAvg system, particularly in scenarios characterized by devices with poor channel conditions and limited sum-power budgets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TCOMM.2023.3338738, IEEE Transactions on Communications



(a) $P^{tot} = 500$ (b) $P^{tot} = 1000$

Fig. 7: The learning accuracy with different scheduling and aggregation times

Appendix A
Proof of Lemma 2

Following (5), (12) and 13, we have

$$\begin{aligned}\boldsymbol{m}^{i+1} - \boldsymbol{m}^{i} &= -\tau\tilde{\boldsymbol{g}}^{i} = -\tau \left(\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \boldsymbol{g}_k^i + \frac{1}{|\mathcal{K}|\,\nu}\boldsymbol{r}^i \right) \\ &= -\tau \left(\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1} \nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right) + \frac{1}{|\mathcal{K}|\,\nu}\boldsymbol{r}^i \right).\end{aligned} \tag{43}$$

Then, we have

$$\begin{aligned}&\mathbb{E}\left[L\left(\boldsymbol{m}^{i+1}\right)\right] - \mathbb{E}\left[L\left(\boldsymbol{m}^{i}\right)\right] \\ \overset{(a)}{\leq}& \mathbb{E}\left[\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \boldsymbol{m}^{i+1} - \boldsymbol{m}^{i}\right\rangle\right] + \frac{\zeta}{2}\mathbb{E}\left[\left\|\boldsymbol{m}^{i+1} - \boldsymbol{m}^{i}\right\|_2^2\right] \\ =& -\tau\mathbb{E}\left[\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\boldsymbol{g}_k^i + \frac{1}{|\mathcal{K}|\,\nu}\boldsymbol{r}^i\right\rangle\right] \\ &+ \frac{\zeta\tau^2}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\boldsymbol{g}_k^i + \frac{1}{|\mathcal{K}|\,\nu}\boldsymbol{r}^i\right\|_2^2\right] \\ =& -\tau\mathbb{E}\left[\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\rangle\right] \\ &- \frac{\tau}{|\mathcal{K}|\,\nu}\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \mathbb{E}\left[\boldsymbol{r}^i\right]\right\rangle \\ &+ \frac{\zeta\tau^2}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right] \\ &+ \frac{\zeta\tau^2}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|\,\nu}\boldsymbol{r}^i\right\|_2^2\right] \\ &+ \frac{\zeta\tau^2}{|\mathcal{K}|\,\nu}\left\langle \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right), \mathbb{E}\left[\boldsymbol{r}^i\right]\right\rangle \\ \overset{(b)}{=}& \underbrace{-\tau\mathbb{E}\left[\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\rangle\right]}_{A} \\ &+ \underbrace{\frac{\zeta\tau^2}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right]}_{B} \\ &+ \underbrace{\frac{\zeta\tau^2}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|\,\nu}\boldsymbol{r}^i\right\|_2^2\right]}_{C},\end{aligned} \tag{44}$$

where (a) is from Assumption 2 and (b) is come from the fact that $\mathbb{E}\left[\boldsymbol{r}^i\right] = 0$. To obtain the upper bound of term A, term A is rewritten as follows

$$\begin{aligned}A =& -\tau\mathbb{E}\left[\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \nabla L\left(\boldsymbol{m}^{i}\right)\right\rangle\right] \\ &- \tau\mathbb{E}\left[\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right) - \nabla L\left(\boldsymbol{m}^{i}\right)\right\rangle\right] \\ =& -\tau\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^{i}\right)\right\|_2^2\right] \\ &+ \tau\mathbb{E}\left[\left\langle \nabla L\left(\boldsymbol{m}^{i}\right), \nabla L\left(\boldsymbol{m}^{i}\right) - \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\rangle\right] \\ =& -\tau\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^{i}\right)\right\|_2^2\right] + \frac{\tau}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^{i}\right)\right\|_2^2\right] \\ &+ \frac{\tau}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^{i}\right) - \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right] \\ &- \frac{\tau}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right] \\ =& -\frac{\tau}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^{i}\right)\right\|_2^2\right]\end{aligned}$$

$$+\underbrace{\frac{\tau}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^i\right)-\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right]}_{A_1}$$
$$\underbrace{-\frac{\tau}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right]}_{A_2}.$$

The upper bound of term A_1 is obtained as follows

$$A_1=\frac{\tau}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{k\in\mathcal{N}}\nabla L_k\left(\boldsymbol{m}^i\right)-\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\nabla L_k\left(\boldsymbol{m}^i\right)\right.\right. \tag{45}$$
$$\left.\left.+\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\nabla L_k\left(\boldsymbol{m}^i\right)-\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right]$$
$$=\frac{\tau}{2}\mathbb{E}\left[\left\|\left(\frac{1}{N}-\frac{1}{|\mathcal{K}|}\right)\sum_{k\in\mathcal{K}}\nabla L_k\left(\boldsymbol{m}^i\right)\right.\right. \tag{46}$$
$$\left.+\frac{1}{N}\sum_{k\in\mathcal{N}/\mathcal{K}}\nabla L_k\left(\boldsymbol{m}^i\right)\right\|_2^2$$
$$\left.+\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\left(\nabla L_k\left(\boldsymbol{m}^i\right)-\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right)\right\|_2^2\right]$$
$$\overset{(a)}{\leq}\tau\mathbb{E}\left[\left\|\left(\frac{1}{N}-\frac{1}{|\mathcal{K}|}\right)\sum_{k\in\mathcal{K}}\nabla L_k\left(\boldsymbol{m}^i\right)\right.\right.$$
$$\left.+\frac{1}{N}\sum_{k\in\mathcal{N}/\mathcal{K}}\nabla L_k\left(\boldsymbol{m}^i\right)\right\|_2^2\right]$$
$$+\tau\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\nabla L_k\left(\boldsymbol{m}^i\right)-\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right)\right\|_2^2\right] \tag{47}$$
$$\overset{(b)}{\leq}\tau\mathbb{E}\left[\left(\frac{1}{|\mathcal{K}|}-\frac{1}{N}\right)\sum_{k\in\mathcal{K}}\mathbb{E}\left[\left\|\nabla L_k\left(\boldsymbol{m}^i\right)\right\|_2\right]\right.$$
$$\left.+\frac{1}{N}\sum_{k\in\mathcal{N}/\mathcal{K}}\mathbb{E}\left[\left\|\nabla L_k\left(\boldsymbol{m}^i\right)\right\|_2\right]\right]^2$$
$$+\tau\left(\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=1}^{E-1}\mathbb{E}\left[\left\|\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2\right]\right)^2$$
$$\overset{(c)}{\leq}4\tau\varpi^2\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\tau\varpi^2\left(E-1\right)^2,$$

where (a) is from that $\|a+b\|_2^2\leq 2\|a\|_2^2+2\|b\|_2^2$ and (b) is from $\|a+b+c\|_2^2\leq\left(\|a\|_2+\|b\|_2+\|c\|_2\right)^2$. Inequality (c) comes from Assumption 1. Due to $\tau\leq\frac{1}{\zeta}$, we obtain the upper bound of the sum of term A_2 and term B as follows

$$A_2+B=\frac{\tau}{2}\left(\zeta\tau-1\right)\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\sum_{\iota=0}^{E-1}\nabla L_k\left(\boldsymbol{w}_k^{i,\iota}\right)\right\|_2^2\right]\leq 0. \tag{48}$$

For the last term C, we note that

$$C=\frac{\zeta\tau^2}{2}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{K}|\nu}\boldsymbol{r}^i\right\|_2^2\right]=\frac{\zeta\tau^2}{2}\frac{d\sigma^2}{|\mathcal{K}|^2\nu^2}. \tag{49}$$

By pluggling these upper bounds back into (44), we complete the proof as follows:

$$\mathbb{E}\left[L\left(\boldsymbol{m}^{i+1}\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{m}^i\right)\right]\leq-\frac{\tau}{2}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^i\right)\right\|_2^2\right]$$
$$+4\tau\varpi^2\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\tau\varpi^2\left(E-1\right)^2+\frac{\zeta\tau^2}{2}\frac{d\sigma^2}{|\mathcal{K}|^2\nu^2}. \tag{50}$$

Appendix B
Proof of Theorem 1

Based on Lemma 2 and Assumption 3, we have

$$\mathbb{E}\left[L\left(\boldsymbol{m}^{i+1}\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{m}^*\right)\right] \tag{51}$$
$$\overset{(a)}{\leq}\eta\left[\mathbb{E}\left[L\left(\boldsymbol{m}^i\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{m}^*\right)\right]\right]$$
$$+\frac{\varpi^2}{\zeta}\left[4\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\left(E-1\right)^2+\frac{d\sigma^2}{2|\mathcal{K}|^2\theta^2}\right]$$
$$=\eta^{i+1}\left[\mathbb{E}\left[L\left(\boldsymbol{m}^0\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{m}^*\right)\right]\right]$$
$$+\frac{\varpi^2}{\zeta}\left[4\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\left(E-1\right)^2+\frac{d\sigma^2}{2|\mathcal{K}|^2\theta^2}\right]\sum_{\kappa=0}^{i}\eta^\kappa$$
$$=\eta^{i+1}\left[\mathbb{E}\left[L\left(\boldsymbol{m}^0\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{m}^*\right)\right]\right]$$
$$+\frac{\varpi^2}{\varrho}\left(1-\eta^{i+1}\right)\left[4\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\left(E-1\right)^2+\frac{d\sigma^2}{2|\mathcal{K}|^2\theta^2}\right],$$

where $\eta=1-\frac{\varrho}{\zeta}$ and (a) is from (27). By replacing $i+1$ with I, we complete the proof.

Appendix C
Proof of Theorem 2

Based on Lemma 2, we have

$$\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^i\right)\right\|_2^2\right]\leq\frac{2}{\tau}\left[\mathbb{E}\left[L\left(\boldsymbol{m}^{i+1}\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{m}^i\right)\right]\right.$$
$$\left.+4\tau\varpi^2\left(1-\frac{|\mathcal{K}|}{N}\right)^2+\tau\varpi^2\left(E-1\right)^2+\frac{\zeta\tau^2}{2}\frac{d\sigma^2}{|\mathcal{K}|^2\nu^2}\right]$$
$$=\frac{2}{\tau}\left[\mathbb{E}\left[L\left(\boldsymbol{m}^{i+1}\right)\right]-\mathbb{E}\left[L\left(\boldsymbol{m}^i\right)\right]\right]$$
$$+\varpi^2\left[8\left(1-\frac{|\mathcal{K}|}{N}\right)^2+2\left(E-1\right)^2+\frac{d\sigma^2}{|\mathcal{K}|^2\theta^2}\right],$$

where (a) is from $\tau=\frac{1}{\zeta}$. By summing i from 0 to $I-1$, we complete the proof of Theorem 2 as follows:

$$\frac{1}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[\left\|\nabla L\left(\boldsymbol{m}^i\right)\right\|_2^2\right] \tag{52}$$

$$\leq \frac{2}{\tau I}\left[\mathbb{E}\left[L\left(\boldsymbol{m}^0\right)\right]-\left[L\left(\boldsymbol{m}^I\right)\right]\right] + \varpi^2\left[8\left(1-\frac{|\mathcal{K}|}{N}\right)^2+2\left(E-1\right)^2+\frac{d\sigma^2}{|\mathcal{K}|^2\theta^2}\right]$$

$$\overset{(a)}{\leq} \frac{2}{\tau I}\left[\mathbb{E}\left[L\left(\boldsymbol{m}^0\right)\right]-\left[L\left(\boldsymbol{m}^*\right)\right]\right] + \varpi^2\left[8\left(1-\frac{|\mathcal{K}|}{N}\right)^2+2\left(E-1\right)^2+\frac{d\sigma^2}{|\mathcal{K}|^2\theta^2}\right]$$

where (a) comes from the fact that $L\left(\boldsymbol{m}^*\right) \leq L\left(\boldsymbol{m}^I\right)$.

REFERENCES

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

[2] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[3] 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. Artif. Intell. Stat.*, 2017, pp. 1273–1282.

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

[5] Z. Yang, M. Chen, K.-K. Wong, H. V. Poor, and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," *Engineering*, vol. 8, pp. 33–41, 2022.

[6] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[7] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proc. IEEE Symp. Security Privacy (SP)*, 2019, pp. 691–706.

[8] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning," in *Proc. IEEE Symp. Security Privacy (SP)*, 2018, pp. 1–15.

[9] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 587–601.

[10] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2020, pp. 2604–2609.

[11] J. Jiang, K. Han, Y. Du, G. Zhu, Z. Wang, and S. Cui, "Optimized power control for over-the-air federated averaging with data privacy guarantee," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2728–2733, 2022.

[12] K. Wei, J. Li, C. Ma, M. Ding, C. Chen, S. Jin, Z. Han, and H. V. Poor, "Low-latency federated learning over wireless channels with differential privacy," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 290–307, 2021.

[13] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, 2020.

[14] N. Yan, K. Wang, C. Pan, and K. K. Chai, "Private federated learning with misaligned power allocation via over-the-air computation," *IEEE Commun. Lett.*, vol. 26, no. 9, pp. 1994–1998, 2022.

[15] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[16] X. Cao, Z. Lyu, G. Zhu, J. Xu, L. Xu, and S. Cui, "An overview on over-the-air federated edge learning," *arXiv preprint arXiv:2208.05643*, 2022.

[17] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, 2020.

[18] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4434–4449, 2021.

[19] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2019.

[20] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 342–358, 2021.

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

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

[23] W. Guo, R. Li, C. Huang, X. Qin, K. Shen, and W. Zhang, "Joint device selection and power control for wireless federated learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2395–2410, 2022.

[24] Y. Liu, D. Liu, G. Zhu, Q. Shi, and C. Zhong, "Over-the-air federated edge learning with error-feedback one-bit quantization and power control," *arXiv preprint arXiv:2303.11319*, 2023.

[25] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private aircomp federated learning with power adaptation harnessing receiver noise," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2020, pp. 1–6.

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

[27] Y. Chen, G. Zhang, H. Xu, X. Chen, and R. Li, "Federated learning: sum power constraints optimization design," *Arab. J. Sci. Eng.*, pp. 1–12, 2021.

[28] S. Huang, P. Zhang, Y. Mao, L. Lian, Y. Wu, and Y. Shi, "Wireless federated learning over mimo networks: Joint device scheduling and beamforming design," in *Proc. IEEE ICC Workshops*. IEEE, 2022, pp. 794–799.

[29] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[30] 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 Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2020.

[31] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. A1380–A1405, 2012.

[32] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optimiz.*, vol. 23, no. 4, pp. 2341–2368, 2013.

[33] Y. Drori and O. Shamir, "The complexity of finding stationary points with stochastic gradient descent," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2020, pp. 2658–2667.

[34] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 9709–9758, 2021.

[35] P. Sun, H. Che, Z. Wang, Y. Wang, T. Wang, L. Wu, and H. Shao, "Pain-FL: Personalized privacy-preserving incentive for federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3805–3820, 2021.

[36] J. Zhang, S. Guo, Z. Qu, D. Zeng, Y. Zhan, Q. Liu, and R. Akerkar, "Adaptive federated learning on non-iid data with resource constraint," *IEEE Trans. Comput.*, vol. 71, no. 7, pp. 1655–1667, 2021.

[37] Y. Liu, X. Zhang, Y. Zhao, Y. He, S. Yu, and K. Zhu, "Chronos: Accelerating federated learning with resource aware training volume tuning at network edges," *IEEE Trans. Veh. Technol.*, early access, 2022.

[38] R. Chen, L. Li, K. Xue, C. Zhang, M. Pan, and Y. Fang, "Energy efficient federated learning over heterogeneous mobile devices via joint design of weight quantization and wireless transmission," *IEEE Trans. Mobile Comput.*, early access, 2022.

[39] X. Zhang, X. Chen, M. Hong, Z. S. Wu, and J. Yi, "Understanding clipping for federated learning: Convergence and client-level differential privacy," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022.

[40] C. Wu, F. Wu, Y. Cao, Y. Huang, and X. Xie, "Fedgnn: Federated graph neural network for privacy-preserving recommendation," *arXiv preprint arXiv:2102.04925*, 2021.