Private Federated Learning with Misaligned Power Allocation via Over-The-Air Computation

Na Yan, Kezhi Wang, Cunhua Pan and Kok Keong Chai

***Abstract*—To further preserve the data privacy of federated learning (FL), we propose a differentially private FL (DPFL) scheme with misaligned power allocation (MPA-DPFL). Unlike most existing over-the-air FL studies, in MPA-DPFL, the gradients are aggregated through over-the-air computation (Aircomp) but do not need to be aligned in the transmission. Therefore, MPA-DPFL can avoid the problem that the signal-to-noise ratio (SNR) of the system is limited by the device with the worst channel condition. We formulate an optimization problem to minimize the optimality gap of MPA-DPFL while guaranteeing a certain degree of privacy protection. Additionally, we demonstrate that the MPA-DPFL is more suitable than the DPFL with aligned power allocation (APA-DPFL) when the channel condition of a device in the system is lower than a threshold. The analytical results are validated through simulation.**

Intex Terms - Data privacy, federated learning, over-the-air computation, power allocation.

I. Introduction

Federated learning (FL) [1] is regarded as one of the privacy-preserving distributed machine learning (ML) techniques, which enables devices to train a model cooperatively with the help of a parameter server (PS). Specifically, devices train the model or compute gradients locally and then send the updated model parameters or gradients to PS for aggregation. Therefore, FL can reduce communication costs and prevent privacy from being exposed to the public by avoiding the transmission of raw data. However, there are still some key challenges for deploying FL due to the limitation of resources in wireless networks [2, 3] and the privacy concerns. Some works [4, 5] have shown that exchanging model parameters or gradients between devices and PS can still reveal sensitive information of local data if exchanged messages are attacked.

One of the countermeasures to prevent privacy leakage of FL is differential privacy (DP) [6] which introduces random noise into the disclosed statistics (i.e., gradients or model parameters) to mask the contribution of any individual data point. However, adopting DP to further secure FL could have a negative impact on the learning performance because the noise will lead to a less accurate aggregated gradient at PS, which is a major issue for the application of DP in FL. Therefore, studies worked on differentially private FL (DPFL) are normally devoted to capturing the tradeoff between privacy and convergence performance or proposing strategy to alleviate the adverse impact of noise on training [7]. The typical noises in DPFL are Binomial [8], Gaussian [9] and Laplacian noise [10]. Unlike the above studies that assumed an ideal communication, DPFL via over-the-air computation (Aircomp) has been investigated recently in [11–13], where channel noise was exploited for enhancing DP. In [11], if channel noise was not adequate for privacy guarantees, artificial Gaussian noise was added to each gradient before transmitting. The scale of the artificial noise (AN) was determined by a static power allocation scheme. Instead of introducing AN, the authors in [12] presented a more energy-efficient strategy to guarantee DP by adjusting transmit power. The authors in [13] proposed adaptive power allocation schemes for DPFL in orthogonal multiple access and non-orthogonal multiple access channels. Similar to most of the studies on over-the-air FL, all the above works tried to align the gradient in the transmission by controlling transmit power. Although gradient alignment can ensure an unbiased gradient estimation at PS, the signal-to-noise ratio (SNR) of the system will be limited to a quite low level when some devices suffer poor channel conditions.

Against the above background, a DPFL scheme with misaligned power allocation (MPA-DPFL) is proposed to enhance the data privacy of FL in this paper. The MPA-DPFL can avoid the issue that the SNR of system is limited by the device with the worst channel condition, which exists in DPFL with aligned power allocation (APA-DPFL) [11]. We also theoretically provide the threshold that can be used to evaluate which one is better, MPA-DPFL or APA-DPFL, in a given FL setting.

II. System Model

A. Federated Learning

We consider a single-antenna wireless FL system as shown in Fig. 1, where K edge devices collaboratively train a model with the help of a PS. Assume that each device holds dataset $\mathcal{D}_k \triangleq \{(\boldsymbol{u}_{k,j}, v_{k,j})\}_{j=1}^{D_k}$ of size D_k where $\boldsymbol{u}_{k,j}$ is the j-th data sample and $v_{k,j}$ is the corresponding label. To simplify the process without losing generality, we assume that $D_1 = \cdots = D_K$, and then, the objective of the training can be given as,

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_k(\boldsymbol{\theta}), \tag{1}$$

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 and Information Sciences, Northumbria University, NE2 1XE, Newcastle upon Tyne, U.K. (e-mail: kezhi.wang@northumbria.ac.uk).

Cunhua Pan is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. (email: cpan@seu.edu.cn).

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the model parameter to be optimized and $\mathcal{L}_k(\boldsymbol{\theta})$ is defined as $\mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{D_k}\sum_{(\boldsymbol{u},v)\in\mathcal{D}_k} \ell(\boldsymbol{\theta};(\boldsymbol{u},v))$, where $\ell(\boldsymbol{\theta};(\boldsymbol{u},v))$ is the empirical loss function.

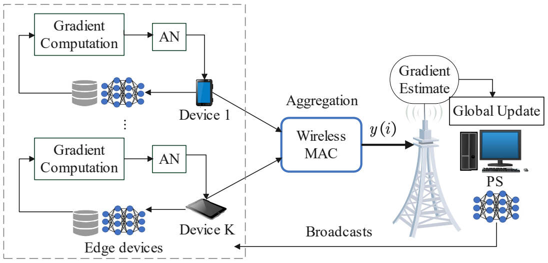

Fig. 1: The differentially private wireless FL.

B. Differential Privacy

Here, we involve DP to further enhance the privacy of the local datasets. The standard DP definition and its Gaussian Mechanism are given as follows.

Definition 1. *(ϵ,ζ)-DP [6]: A randomized mechanism $\mathcal{O}$ guarantees (ϵ,ζ)-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* $\Pr[\mathcal{O}(\mathcal{D})\in\mathcal{Q}] \leqslant e^{\epsilon}\Pr[\mathcal{O}(\mathcal{D}')\in\mathcal{Q}]+\zeta$.

Definition 2. *Gaussian Mechanism (GM) [6] : A Mechanism $\mathcal{O}$ is called as a GM, which alters the output of another algorithm $\mathcal{L}:\mathcal{D}\to\mathcal{Q}$ by adding Gaussian noise, i.e., $\mathcal{O}(\mathcal{D}) = \mathcal{L}(\mathcal{D})+\mathcal{N}\left(0,\sigma^2\mathbf{I}_d\right)$. GM $\mathcal{O}$ guarantees (ϵ,ζ)-DP with $\epsilon = \frac{\Delta S}{\sigma}\sqrt{2\ln\left(\frac{1.25}{\zeta}\right)}$ where $\Delta S \triangleq \max_{\mathcal{D},\mathcal{D}'}\|\mathcal{L}(\mathcal{D})-\mathcal{L}(\mathcal{D}')\|_2$.*

The term ζ allows for breaching ϵ-DP with the probability ζ while ϵ denotes the protection level and a smaller ϵ means a higher privacy preservation level.

C. MPA-DPFL

The details of MPA-DPFL are given as follows by taking round i as an example. At the beginning of round i, PS broadcasts the latest model $\boldsymbol{\theta}^i$ to the devices. Next, each device lets $\boldsymbol{\theta}_k^i = \boldsymbol{\theta}^i$, and then computes the gradient $\boldsymbol{g}_k^i = \nabla\mathcal{L}_k\left(\boldsymbol{\theta}_k^i\right)$, which is assumed to satisfy $\left\|\boldsymbol{g}_k^i\right\|_2 \leqslant I$. According to DP, $\boldsymbol{e}_k^i \sim \mathcal{N}(0,\mathbf{I}_d)$ is added to the gradient before transmitting. In particular, the input signal from device k is given as:

$$\boldsymbol{x}_k^i = \frac{\sqrt{\lambda_k^i P_k}}{I}\boldsymbol{g}_k^i + \sqrt{\frac{\mu_k^i P_k}{d}}\boldsymbol{e}_k^i, \tag{2}$$

where P_k is the maximum transmission power of device k. Power scaling factors $\lambda_k^i \geqslant 0$ and $\mu_k^i \geqslant 0$ indicate the fractions of power dedicated to the gradient and the AN, respectively. Also, $\lambda_k^i + \mu_k^i \leqslant 1$ should hold for satisfying transmit power constraint, i.e., $\mathbb{E}\left[\left\|\boldsymbol{x}_k^i\right\|_2^2\right] \leqslant P_k$. Then, gradients are transmitted via a shared wireless channel and aggregated through Aircomp. The received signal at the PS is given by:

$$\boldsymbol{y}(i) = \sum_{k=1}^{K} h_k^i\left(\frac{\sqrt{\lambda_k^i P_k}}{I}\boldsymbol{g}_k^i + \sqrt{\frac{\mu_k^i P_k}{d}}\boldsymbol{e}_k^i\right) + \boldsymbol{r}^i, \tag{3}$$

where $\boldsymbol{r}^i \sim \mathcal{N}(0, N_0\mathbf{I}_d)$ is channel noise and we assume that $h_k^i \in \mathbb{R}^+$ is the real channel gain coefficient for simplicity. The coefficients are independent across devices and training rounds, but remain constant within one round. The distinction between APA-DPFL [11] and MPA-DPFL is that APA-DPFL requires $h_k^i\sqrt{\lambda_k^i P_k} = \sqrt{\min_j\left(h_j^i\right)^2 P_j} = c$, which is referred to as alignment coefficient.

To obtain the averaging of the aggregated gradient, PS performs post-processing by $\hat{\boldsymbol{g}}^i = \frac{1}{K}\boldsymbol{y}(i)$. It thus follows that the aggregation error caused by the misaligned aggregation is given as $\boldsymbol{\Delta}^i = \hat{\boldsymbol{g}}^i - \boldsymbol{g}^i$, where $\boldsymbol{g}^i = \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right) = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{g}_k^i$ denotes the noise-free aggregated gradient. Finally, PS performs the global update by $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i - \eta\hat{\boldsymbol{g}}^i = \boldsymbol{\theta}^i - \eta\left(\boldsymbol{g}^i + \boldsymbol{\Delta}^i\right)$, where η is the learning rate.

III. Theoretical analysis and power allocation optimization

In this section, we formulate a power allocation optimization problem based on the main analysis results as follows.

A. Privacy analysis and convergence analysis

1) Assumptions: For analysis, we first make the following assumptions.

Assumption 1. *Assume that $\mathcal{L}(\cdot)$ satisfies Polyak-Łojasiewicz inequality, i.e., for all $\boldsymbol{\theta}$, there is a constant $\rho \geqslant 0$ satisfying $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2 \geqslant 2\rho\left(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)\right)$.*

Assumption 2. *Assume that $\mathcal{L}(\cdot)$ is ξ-smooth, i.e., for all $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, one has $\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}') \leqslant \langle\boldsymbol{\theta}-\boldsymbol{\theta}',\nabla\mathcal{L}(\boldsymbol{\theta}')\rangle + \frac{\xi}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2$.*

Assumption 3. *For each device k in round i, the gradient satisfies $\left\|\boldsymbol{g}_k^i\right\|_2 \leqslant I$.*

2) Privacy analysis: We here present the privacy analysis for MPA-DPFL as follows.

Lemma 1. *Assume that Assumption 3 holds and $\mathcal{D}_k,\mathcal{D}_k'$ are two adjacent datasets with only one sample different. Based on the definition given below,*

$$\boldsymbol{y}(i) = \sum_{n=1}^{K} h_n^i\left(\frac{\sqrt{\lambda_n^i P_n}}{I}\boldsymbol{g}_n^i + \sqrt{\frac{\mu_n^i P_n}{d}}\boldsymbol{e}_n^i\right) + \boldsymbol{r}^i, \tag{4}$$

$$\boldsymbol{y}'(i) = \sum_{n=1}^{K} h_n^i\left(\frac{\sqrt{\lambda_n^i P_n}}{I}\left(\boldsymbol{g}_n^i\right)' + \sqrt{\frac{\mu_n^i P_n}{d}}\boldsymbol{e}_n^i\right) + \boldsymbol{r}^i, \tag{5}$$

where

$$\left(\boldsymbol{g}_n^i\right)' = \begin{cases} \boldsymbol{g}_n^i = \frac{1}{D_n}\sum_{(\boldsymbol{u},v)\in\mathcal{D}_n}\nabla\ell\left(\boldsymbol{\theta}_n^i;(\boldsymbol{u},v)\right), n \neq k \\ \left(\boldsymbol{g}_k^i\right)' = \frac{1}{D_k}\sum_{(\boldsymbol{u},v)\in\mathcal{D}_k'}\nabla\ell\left(\boldsymbol{\theta}_k^i;(\boldsymbol{u},v)\right), n = k \end{cases} \tag{6}$$

the bound of the sensitivity of device k in round i is given by:

$$\Delta S_k^i = \max_{\mathcal{D}_k,\mathcal{D}_k'}\|\boldsymbol{y}(i) - \boldsymbol{y}'(i)\|_2 \leqslant 2h_k^i\sqrt{\lambda_k^i P_k}. \tag{7}$$

Proof. See Appendix A. □

One can see that the sensitivity of each device can be controlled by the power scaling factor assigned to the gradients. Based on Lemma 1, we give the privacy leakage of each device as follows.

Lemma 2. *Assume that Assumption 3 holds, MPA-DPFL guarantees $\left(\epsilon_k^i, \zeta\right)$-DP for device k in round i if the following condition is satisfied,*

$$\frac{2h_k^i\sqrt{\lambda_k^i P_k}}{\sqrt{\sum_{j=1}^K \frac{\left(h_j^i\right)^2\mu_j^i P_j}{d}+N_0}}\cdot\sqrt{2\ln\frac{1.25}{\zeta}}=\epsilon_k^i. \tag{8}$$

Proof. Based on Lemma 1 and the GM of DP, we complete the proof of Lemma 2. □

Note that if "=" in (8) is replaced by "$\leqslant$", it indicates a stronger privacy protection, therefore, we regard that it still satisfies $\left(\epsilon_k^i, \zeta\right)$-DP. Since $h_k^i\sqrt{\lambda_k^i P_k}$ is bounded by $\max\limits_k\left\{h_k^i\sqrt{P_k}\right\}$, one can learn that the privacy leakage goes asymptotically to 0 when K approaches infinity.

3) Convergence analysis: Assume that $\boldsymbol{\theta}^*$ is the optimal model and training terminates after T rounds, the optimality gap of MPA-DPFL is given as follows.

Lemma 3. *Assume that Assumption 1 to 3 hold and set $\eta=\frac{1}{\xi}$, the bound of the optimality gap can be given by,*

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}^T\right)\right]-\mathcal{L}\left(\boldsymbol{\theta}^*\right)\leqslant\varphi^T\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}^0\right)\right]-\mathcal{L}\left(\boldsymbol{\theta}^*\right)\\&+\tfrac{1}{2\xi}\textstyle\sum_{i=0}^{T-1}\varphi^{T-1-i}\mathbb{E}\left[\left\|\boldsymbol{\Delta}^i\right\|_2^2\right]_M.\end{aligned} \tag{9}$$

where $\mathbb{E}\left[\left\|\boldsymbol{\Delta}^i\right\|_2^2\right]_M=\frac{1}{K^2}\sum_{k=1}^K\left(h_k^i\right)^2\mu_k^i P_k+\frac{dN_0}{K^2}+\frac{1}{K}\sum_{k=1}^K\left(h_k^i\sqrt{\lambda_k^i P_k}-I\right)^2$ and $\varphi=1-\frac{\rho}{\xi}$.

Proof. See Appendix B. □

The first term on the right hand side of the inequality is the initial optimality gap, and the second term relates to the aggregated error which can be reduced by controlling $\left\{\lambda_k^i\right\},\left\{\mu_k^i\right\}$ for enhancing the learning performance.

B. Power allocation optimization

To minimize the optimality gap while guaranteeing (ϵ,ζ)-DP for each device, we formulate an optimization problem as follows. We define $\varrho=\sqrt{2\ln\frac{1.25}{\zeta}}$ for notation convenience.

$$\min_{\{\lambda_k\},\{\mu_k\}}\left\{\sum_{k=1}^K\left(h_k\sqrt{\lambda_k P_k}-I\right)^2+\sum_{k=1}^K\frac{h_k^2\mu_k P_k}{K}\right\} \tag{10}$$

$$s.t.\frac{2\varrho h_k\sqrt{\lambda_k P_k}}{\sqrt{\sum_{j=1}^K\frac{h_j^2\mu_j P_j}{d}+N_0}}=\epsilon,\quad\forall k \tag{10a}$$

$$\lambda_k+\mu_k\leqslant 1,\quad\forall k \tag{10b}$$

$$\lambda_k\geqslant 0,\mu_k\geqslant 0,\quad\forall k. \tag{10c}$$

In (10), we have simplified the objective function in three ways. Firstly, we ignore the effect of φ on the optimality gap because it has little impact when ρ is quite small compared with ξ, which is equivalent to performing the optimization in each training round. Secondly, we discard the items that are not related to power allocation, i.e., the initial optimal gap. Additionally, we omit the index i of each variable for ease of presentation. (10a) denotes (ϵ,ζ)-DP constraint for each device. The maximum power constraint and nonnegativity constraint are satisfied in (10b) and (10c), respectively. We consider the same privacy constraint and fixed transmit power of each device for simplicity. The MPA-DPFL can be readily extended to the case that devices have distinct privacy constraints. The case of adaptive transmit power will be left for future work.

By defining $H=\sum_{k=1}^K h_k^2 P_k$ and $\varPhi=\sum_{k=1}^K h_k^2\mu_k P_k$, our solution to solve this problem is given as follows.

Lemma 4. *Assume that $H\geqslant\frac{KN_0\epsilon^2}{4\varrho^2}$, our solution to the optimization problem is given as follows:*

$$\lambda_k^*=\min\left\{1,\tfrac{\epsilon^2}{4\varrho^2h_k^2P_k}\left(\tfrac{\varPhi^*}{d}+N_0\right)\right\}, \tag{11}$$

$$\mu_k^*=\min\left\{1-\lambda_k^*,\tfrac{\max\left\{\varPhi^*-\sum_{j=1}^{k-1}\beta_j,0\right\}}{h_k^2P_k}\right\}, \tag{12}$$

where $\beta_j=h_j^2P_j\mu_j^*$ and

- *$\varPhi^*=0$, if $\frac{2I\varrho\epsilon}{\epsilon^2+\frac{4d\varrho^2}{K^2}}\leqslant\sqrt{N_0}$;*
- *$\varPhi^*=\frac{4H\varrho^2-KN_0\epsilon^2}{4\varrho^2+\frac{K\epsilon^2}{d}}$, if $\frac{2I\varrho\epsilon}{\epsilon^2+\frac{4d\varrho^2}{K^2}}\geqslant 2\varrho\sqrt{\frac{H+dN_0}{K\epsilon^2+4d\varrho^2}}$;*
- *$\varPhi^*=d\left(\frac{4I^2\varrho^2\epsilon^2}{\left(\epsilon^2+\frac{4d\varrho^2}{K^2}\right)^2}-N_0\right)$, if $\sqrt{N_0}\leqslant\frac{2I\varrho\epsilon}{\epsilon^2+\frac{4d\varrho^2}{K^2}}\leqslant 2\varrho\sqrt{\frac{H+dN_0}{K\epsilon^2+4d\varrho^2}}$.*

Proof. See Appendix C. □

From the above, one has $0\leqslant\varPhi^*\leqslant\frac{4H\varrho^2-KN_0\epsilon^2}{4\varrho^2+\frac{K\epsilon^2}{d}}=M$ and it thus follows that the upper bound of $\mathbb{E}\left[\left\|\boldsymbol{\Delta}^i\right\|_2^2\right]_M$ is $U=\max_{j=\{0,M\}}\left\{\mathbb{E}\left[\left\|\boldsymbol{\Delta}^i\right\|_2^2\right]_M|_{\varPhi^*=j}\right\}$.

C. MPA-DPFL vs APA-DPFL

Following APA-DPFL [11], the aggregated error is $\boldsymbol{\Delta}_{al}^i=\frac{I}{Kc}\left(\sum_{k=1}^K h_k^i\sqrt{\frac{\mu_{al,k}^iP_k}{d}}\boldsymbol{e}_k^i+\boldsymbol{r}^i\right)$. The privacy leakage can be given by,

$$\epsilon=\frac{2c\varrho}{\sqrt{\sum_{j=1}^K\frac{\left(h_j^i\right)^2\mu_{al,j}^iP_j}{d}+N_0}}. \tag{13}$$

Based on the solution we obtained in Lemma 4, we have the following Theorem.

Theorem 1. *Assume that $H\geqslant\frac{KN_0\epsilon^2}{4\varrho^2}$ and U is the upper bound of $\mathbb{E}\left[\left\|\boldsymbol{\Delta}^i\right\|_2^2\right]_M$. Compared with APA-DPFL, there is a lower optimality gap for MPA-DPFL when c satisfies,*

$$c\leqslant\min\left\{\tfrac{\epsilon\sqrt{N_0}}{2\varrho},\tfrac{I}{K}\sqrt{\tfrac{dN_0}{U}}\right\}. \tag{14}$$

Proof. See Appendix D. □

For any set of feasible solutions to Problem (10), one can obtain the corresponding condition that MPA-DPFL has a

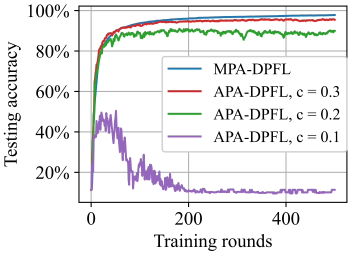

Fig. 2: MPA-DPFL versus APA-DPFL.

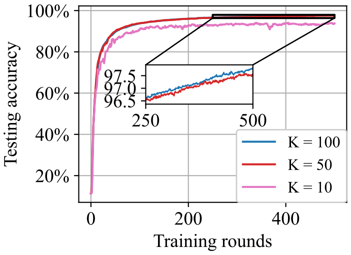

Fig. 3: Accuracy of MPA-DPFL under different K.

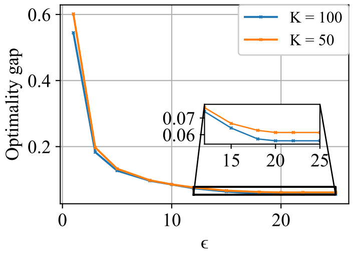

Fig. 4: Optimality gap of MPA-DPFL with different ϵ.

lower optimality gap, which indicates that when the channel gain of one device is worse than a threshold, misaligned power allocation is more suitable than aligned aggregation.

IV. Simulation Results

We evaluate MPA-DPFL through training convolutional neural network (CNN) [14] on the popular MNIST dataset. The learning rate is set as $\eta = 0.1$ and $N_0 = 1$.

In Fig. 2, we plot the testing accuracy of MPA-DPFL and APA-DPFL with different alignment coefficients c, where $K = 100$, $\epsilon = 10$, and $P = 500$. The alignment coefficient c in APA-DPFL is varied by adjusting the worst channel gain of the device. The channel gain coefficients in MPA-DPFL setting are the same as those in $c = 0.1$. The results validate that when the alignment coefficient c of APA-DPFL is smaller than the threshold, MPA-DPFL performs better than APA-DPFL and the superiority is more significant as c decreases. A smaller c results in a lower overall SNR of the system, then, the utility of all gradients will be greatly affected by noise, which leads to a less accurate model. Particularly, when c approaches a quite small level ($c = 0.1$), the noise becomes the main component of the received gradient at PS, and the model fails to converge. By contrast, the power allocation in MPA-DPFL does not force gradient alignment, even though some of the devices have poor channel conditions, the SNR of other devices will not be limited.

We study the impact of the number of devices on MPA-DPFL in Fig. 3, where $\epsilon = 10$, and $P = 500$. The accuracy of the obtained model is observed to increase with K (while keeping the total dataset size constant). When more devices share the noise required for required DP, it means that each gradient suffers less noise distortion, therefore, a more accurate model can be obtained.

In Fig. 4, we plot the optimality gap as a function of the privacy level ϵ, where $P = 500$. The optimality gap decreases with the increase of ϵ until the channel noise is sufficient to meet the privacy requirements, i.e. $\epsilon = 20, 22, 25$.

V. Conclusion

In this paper, we have proposed MPA-DPFL to further secure FL. We have also provided a threshold that can be used to estimate whether MPA-DPFL or APA-DPFL has better performance. To obtain some important analytical comparisons between MPA-DPFL and APA-DPFL, this preliminary work considered simple communication settings for tractability. More practical scenarios will be considered in our future work. For example, more efficient privacy protection is possibly obtained by considering the correlation of the gradients and the channels [15, 16]. Specifically, when the gradients in adjacent training rounds are correlated, we may just send the difference, which may lead to a lower sensitivity, therefore, guaranteeing stronger privacy.

Appendix A
Proof of Lemma 1

The sensitivity of device k in round i is given by,

$$\begin{aligned}&\mathit{\Delta} S_k^i = \max_{\mathcal{D}_k,\mathcal{D}_k'} \left\| \boldsymbol{y}(i) - \boldsymbol{y}'(i) \right\|_2 \\ &= \frac{h_k^i \sqrt{\lambda_k^i P_k}}{I} \max_{\mathcal{D}_k,\mathcal{D}_k'} \left\| \boldsymbol{g}_k^i - \left(\boldsymbol{g}_k^i\right)' \right\|_2 \overset{(a)}{\leqslant} 2 h_k^i \sqrt{\lambda_k^i P_k},\end{aligned} \tag{15}$$

where (a) is obtained by using Triangular Inequality and Assumption 3. □

Appendix B
Proof of Lemma 3

Recall that $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i - \eta\left(\boldsymbol{g}^i + \boldsymbol{\mathit{\Delta}}^i\right)$ and $\boldsymbol{g}^i = \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right)$,

$$\begin{aligned}&\mathcal{L}\left(\boldsymbol{\theta}^{i+1}\right) - \mathcal{L}\left(\boldsymbol{\theta}^i\right) \overset{(a)}{\leqslant} \left\langle \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right), \boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i \right\rangle \\ &+ \tfrac{\xi}{2} \left\| \boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i \right\|_2^2 = -\eta \left\langle \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right), \boldsymbol{g}^i + \boldsymbol{\mathit{\Delta}}^i \right\rangle \\ &+ \tfrac{\xi(\eta)^2}{2} \left\| \boldsymbol{g}^i + \boldsymbol{\mathit{\Delta}}^i \right\|_2^2 = -\eta\left(1 - \tfrac{\xi\eta}{2}\right) \left\| \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right) \right\|_2^2 \\ &+ \tfrac{\xi(\eta)^2}{2} \left\| \boldsymbol{\mathit{\Delta}}^i \right\|_2^2 + \eta\left(\xi\eta - 1\right) \left\langle \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right), \boldsymbol{\mathit{\Delta}}^i \right\rangle,\end{aligned} \tag{16}$$

where (a) comes from Assumption 2. By applying $\langle \boldsymbol{a}, \boldsymbol{b} \rangle \leqslant \frac{\|\boldsymbol{a}\|_2^2}{2} + \frac{\|\boldsymbol{b}\|_2^2}{2}$, one has $\eta\left(\xi\eta - 1\right)\left\langle \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right), \mathbb{E}\left[\boldsymbol{\mathit{\Delta}}^i\right] \right\rangle \leqslant \frac{\eta(\xi\eta-1)}{2} \left\| \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right) \right\|_2^2 + \frac{\eta(\xi\eta-1)}{2} \left\| \mathbb{E}\left[\boldsymbol{\mathit{\Delta}}^i\right] \right\|_2^2$. Then, one has $\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}^{i+1}\right) - \mathcal{L}\left(\boldsymbol{\theta}^i\right)\right] \leqslant -\frac{\eta}{2}\left(3 - 2\xi\eta\right) \left\| \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right) \right\|_2^2 + \frac{\eta(\xi\eta-1)}{2} \left\| \mathbb{E}\left[\boldsymbol{\mathit{\Delta}}^i\right] \right\|_2^2 + \frac{\xi(\eta)^2}{2} \mathbb{E}\left[\left\| \boldsymbol{\mathit{\Delta}}^i \right\|_2^2\right] \overset{(a)}{\leqslant} -\frac{1}{2\xi} \left\| \nabla\mathcal{L}\left(\boldsymbol{\theta}^i\right) \right\|_2^2 + \frac{1}{2\xi} \mathbb{E}\left[\left\| \boldsymbol{\mathit{\Delta}}^i \right\|_2^2\right]$, where (a) follows from the fact that $\eta = \frac{1}{\xi} \leqslant \frac{3}{2\xi}$. It thus follows from Assumption 1 that

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}^{i+1}\right)\right] - \mathcal{L}\left(\boldsymbol{\theta}^*\right) \leqslant \left(1 - \tfrac{\rho}{\xi}\right) \left[\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}^i\right)\right] - \mathcal{L}\left(\boldsymbol{\theta}^*\right)\right] \\ &+ \tfrac{1}{2\xi} \mathbb{E}\left[\left\| \boldsymbol{\mathit{\Delta}}^i \right\|_2^2\right].\end{aligned} \tag{17}$$

Following that $\mathbb{E}\left[\boldsymbol{e}_k^i\right]=\mathbb{E}\left[\boldsymbol{r}^i\right]=0$, one has,

$$\mathbb{E}\left[\left\|\boldsymbol{\Delta}^i\right\|_2^2\right] \overset{(a)}{\leqslant} \tfrac{dN_0}{K^2}+\tfrac{1}{K}\sum_{k=1}^K\left(\tfrac{h_k^i\sqrt{\lambda_k^i P_k}}{I}-1\right)^2\left\|\boldsymbol{g}_k^i\right\|_2^2$$
$$+\tfrac{1}{K^2}\sum_{k=1}^K\left(h_k^i\right)^2\mu_k^i P_k \overset{(b)}{\leqslant} \tfrac{1}{K}\sum_{k=1}^K\left(h_k^i\sqrt{\lambda_k^i P_k}-I\right)^2$$
$$+\tfrac{1}{K^2}\sum_{k=1}^K\left(h_k^i\right)^2\mu_k^i P_k+\tfrac{dN_0}{K^2}, \tag{18}$$

where (a) comes from Jensen's Inequality and $\mathbb{E}\left[\|\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}\|_2^2\right]=\|\boldsymbol{a}\|_2^2+\|\boldsymbol{b}\|_2^2+\|\boldsymbol{c}\|_2^2+2\mathbb{E}\left[\langle\boldsymbol{a},\boldsymbol{b}\rangle\right]+2\mathbb{E}\left[\langle\boldsymbol{a},\boldsymbol{c}\rangle\right]+2\mathbb{E}\left[\langle\boldsymbol{b},\boldsymbol{c}\rangle\right]$, and (b) follows from Assumption 3. Finally, applying recursion on (17) and replacing $i+1$ with T, we complete the proof. □

Appendix C
Proof of Lemma 4

Following (10a), one has $h_k\sqrt{\lambda_k P_k}=\frac{\epsilon}{2\varrho}\sqrt{\frac{\varPhi}{d}+N_0}$. By relaxing (10b) as $\sum_{k=1}^K h_k^2 P_k\left(\lambda_k+\mu_k\right)\leqslant H$ and replacing $\sum_{k=1}^K h_k^2\mu_k P_k$ with $\varPhi$, the original optimization problem can be re-formulated as,

$$\min_{\varPhi}\left\{\frac{K\epsilon^2}{4\varrho^2}\left(\frac{\varPhi}{d}+N_0\right)+\frac{\varPhi}{K}-\frac{KI\epsilon}{\varrho}\sqrt{\frac{\varPhi}{d}+N_0}\right\} \tag{19}$$

$$s.t.\quad 0\leqslant\varPhi\leqslant\frac{4H\varrho^2-KN_0\epsilon^2}{4\varrho^2+\frac{K\epsilon^2}{d}}. \tag{19a}$$

Then, we replace $\varPhi$ with $\omega=\sqrt{\frac{\varPhi}{d}+N_0}$ to convert this problem into a problem of obtaining the minimum value of quadratic function. Due to the limited space, we omit the details here. By applying properties of a quadratic function, one obtains $\varPhi^*$ as follows:

- $\varPhi^*=0$, if $\frac{2I\varrho\epsilon}{\epsilon^2+\frac{4d\varrho^2}{K^2}}\leqslant\sqrt{N_0}$;
- $\varPhi^*=\frac{4H\varrho^2-KN_0\epsilon^2}{4\varrho^2+\frac{K\epsilon^2}{d}}$, if $\frac{2I\varrho\epsilon}{\epsilon^2+\frac{4d\varrho^2}{K^2}}\geqslant 2\varrho\sqrt{\frac{H+dN_0}{K\epsilon^2+4d\varrho^2}}$;
- $\varPhi^*=d\left(\frac{4I^2\varrho^2\epsilon^2}{\left(\epsilon^2+\frac{4d\varrho^2}{K^2}\right)^2}-N_0\right)$, if $\sqrt{N_0}\leqslant\frac{2I\varrho\epsilon}{\epsilon^2+\frac{4d\varrho^2}{K^2}}\leqslant 2\varrho\sqrt{\frac{H+dN_0}{K\epsilon^2+4d\varrho^2}}$.

Then, one has $h_k^2 P_k\lambda_k^*=\frac{\epsilon^2}{4\varrho^2}\left(\frac{\varPhi^*}{d}+N_0\right)$. Since (10b) has been relaxed, we finally use $\lambda_k^*=\min\left\{1,\frac{\epsilon^2}{4\varrho^2 h_k^2 P_k}\left(\frac{\varPhi^*}{d}+N_0\right)\right\}$ to guarantee $\lambda_k\leqslant 1$, which still satisfies (ϵ,ζ)-DP as we mentioned before. For $\{\mu_k\}$, we first rank the leftover powers $1-\lambda_k^*$ of each device and then we decide $\{\mu_k^*\}$ by $\mu_k^*=\min\left(1-\lambda_k^*,\frac{\max\left\{\varPhi^*-\sum_{j=1}^{k-1}\beta_j,0\right\}}{h_k^2 P_k}\right)$ where $\beta_j=h_j^2 P_j\mu_j^*$. Then, we complete the proof. □

Appendix D
Proof of Theorem 1

Accroding to (13), one has $\sum_{k=1}^K\left(h_k^i\right)^2\mu_{al,k}^i P_k=\frac{4dc^2\varrho^2}{\epsilon^2}-dN_0$. Similar to (18), one has $\mathbb{E}\left[\left\|\boldsymbol{\Delta}_{al}^i\right\|_2^2\right]=\frac{I^2}{K^2c^2}\sum_{k=1}^K\left(h_k^i\right)^2\mu_{al,k}^i P_k+\frac{I^2}{K^2c^2}dN_0$. When $c\leqslant\min\left\{\frac{\epsilon\sqrt{N_0}}{2\varrho},\frac{I}{K}\sqrt{\frac{dN_0}{U}}\right\}$, one has $\mu_{al,k}^i=0$ and $\mathbb{E}\left[\left\|\boldsymbol{\Delta}_{al}^i\right\|_2^2\right]\geqslant\mathbb{E}\left[\left\|\boldsymbol{\Delta}^i\right\|_2^2\right]_M$. Therefore, MPA-DPFL has a lower optimality gap than APA-DPFL. □

References

[1] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[2] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2020.

[3] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2020.

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

[5] L. Zhu and S. Han, "Deep leakage from gradients," in *Federated learning*. Springer, 2020, pp. 17–31.

[6] 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.

[7] N. Yan, K. Wang, C. Pan, and K. K. Chai, "Performance analysis for channel-weighted federated learning in oma wireless networks," *IEEE Signal Processing Letters*, vol. 29, pp. 772–776, 2022.

[8] N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. McMahan, "Cpsgd: Communication-efficient and differentially-private distributed sgd," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 7575–7586.

[9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[10] N. Wu, F. Farokhi, D. Smith, and M. A. Kaafar, "The value of collaboration in convex machine learning with differential privacy," in *2020 IEEE Symposium on Security and Privacy (SP)*, 2020, pp. 304–317.

[11] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2604–2609.

[12] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private aircomp federated learning with power adaptation harnessing receiver noise," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

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

[14] Y. Luo, J. Xu, W. Xu, and K. Wang, "Sliding differential evolution scheduling for federated learning in bandwidth-limited networks," *IEEE Communications Letters*, vol. 25, no. 2, pp. 503–507, 2020.

[15] W. Liu, X. Zang, B. Vucetic, and Y. Li, "Over-the-air computation with spatial-and-temporal correlated signals," *IEEE Wireless Communications Letters*, vol. 10, no. 7, pp. 1591–1595, 2021.

[16] M. Frey, I. Bjelaković, and S. Stańczak, "Over-the-air computation in correlated channels," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5739–5755, 2021.