Energy Efficient and Differentially Private Federated Learning via a Piggyback Approach

Rui Chen, *Graduate Student Member, IEEE*, Chenpei Huang, *Student Member, IEEE*, Xiaoqi Qin, *Member, IEEE*, Nan Ma, *Member, IEEE*, Miao Pan, *Senior Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—This artilce aims to develop a differential private federated learning (FL) scheme with the least artificial noises added while minimizing the energy consumption of participating mobile devices. By observing that some communication efficient FL approaches and even the nature of wireless communications contribute to the differential privacy (DP) preservation of training data on mobile devices, in this paper, we propose to jointly leverage gradient compression techniques (i.e., gradient quantization and sparsification) and additive white Gaussian noises (AWGN) in wireless channels to develop a piggyback DP approach for FL over mobile devices. Even with the piggyback DP approach, information distortion caused by gradient compression and noise perturbation may slow down FL convergence, which in turn consumes more energy of mobile devices for local computing and model update communications. Thus, we theoretically analyze FL convergence and formulate an energy efficient FL optimization under piggyback DP, transmission power, and FL convergence constraints. Furthermore, we propose an efficient iterative algorithm where closed-form solutions for artificial DP noise and power control are derived. Extensive simulation and experimental results demonstrate the effectiveness of the proposed scheme in terms of energy efficiency and privacy preservation.**

***Index Terms*—Federated learning over mobile devices, piggyback differential privacy, gradient compression, white Gaussian noises.**

Manuscript received 6 April 2022; revised 22 March 2023; accepted 29 March 2023. Date of publication 30 May 2023; date of current version 6 March 2024. The work of Rui Chen, Chenpei Huang, and Miao Pan was supported in part by the US National Science Foundation under Grant CNS-2029569. The work of Xiaoqi Qin was supported in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2021QNRC001, and in part by the NSFC Project under Grant U22B2003. Recommended for acceptance by Y. Liu. *(Corresponding author: Miao Pan.)*

Rui Chen, Chenpei Huang, and Miao Pan are with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204 USA (e-mail: rchen19@uh.edu; chuang30@uh.edu; mpan2@uh.edu).

Xiaoqi Qin is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: xiaoqiqin@bupt.edu.cn).

Nan Ma is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Department of Broadband Communication, Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: manan@bupt.edu.cn).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMC.2023.3268323, provided by the authors.

Digital Object Identifier 10.1109/TMC.2023.3268323

I. Introduction

MACHINE learning is one of the most disruptive technologies the world has witnessed in the last few years. Federated learning (FL), first introduced by Google [1], is the currently most popular distributed machine learning paradigm, which aims to enable participating devices to collaboratively learn a joint global learning model without sharing their local training data on the devices. In FL, clients conduct local training on their own devices and then periodically transmit local model updates to an FL aggregator. The FL aggregator broadcasts the aggregated global model to the clients for the next round of FL training. The procedure repeats until FL converges. With the fast development of hardware, more and more mobile devices have been equipped with increasing computing capability and large storage, which allow them to conduct on-device learning (e.g., Google Pixel 4a, iPhone 12, Galaxy Note20, iPad Pro, etc.). The "marriage" of FL and mobile devices has potentially promoted numerous applications in various domains such as cardiac event prediction in e-Healthcare, high-definition map construction for autonomous driving in smart transportation, physical hazard detection in smart home, IoT based smart farming in agriculture, environmental monitoring in industrial IoT, etc. [2], [3], [4], [4]. However, to practically execute FL over mobile devices still faces many critical challenges. For example, both on-device local training and wireless transmissions of huge size model updates during FL training are power hungry, which may quickly deplete the energy of battery powered mobile devices. Besides, although FL does avoid the direct privacy leakage by retaining the data on local devices, the intermediate updates exchanged with FL aggregator and FL outputs could still leak private information due to advanced inference attacks [5], [6], [7], e.g., model inversion attacks [8], [9] and membership inference attacks [10], [11].

To alleviate the energy pressure of participating mobile devices and improve the energy efficiency of FL, reducing the energy consumption of wireless communications for model updates is one possible option, which has led to a popular research direction: communication efficient FL. In the wireless community, existing works mainly focus on radio resource allocation [12], [13], [14], [15] or transmission technique improvement [16], [17] to achieve communication efficiency under FL convergence constraints. Most of those studies ignore how to inherently reduce the communication payload from the FL

algorithm itself. By contrast, in the machine learning community, some gradient compression techniques (e.g., stochastic quantization [18] and randomized sparsification [19]) are proposed to reduce the size of local model updates for communication efficiency, which have similar statistical convergence properties with their uncompressed counterparts in FL. A few pioneering efforts [20], [21], [22] jointly consider radio resource allocation and gradient compression to improve the energy efficiency of FL over mobile devices, while potential risks of training data privacy leakage in FL are left uncovered.

As for training data privacy, standard encryption tools such as secure multi-party computation or homomorphic encryption only ensure no secure information leakage during the transmission process, but cannot protect the input data to be inferred from the final learning outputs. Differential privacy (DP), which is a cryptography-inspired rigorous definition of privacy [23], [24], has become the de-facto remedy for training data privacy preservation in FL. Besides, unlike the complex and high-cost encryption based methods, DP offers a simple implementation mechanism to preserve data privacy by injecting artificial noises to perturb the gradients [25] or model outputs [26]. Despite DP's desired properties above, the dark side is that artificially injecting large amount of noises to achieve targeted DP may severely distort the information to transmit, slow down FL convergence and even degrade FL learning accuracy. Consequently, that may incur more computing and communication workload and thereby more energy consumption for mobile devices. Thus, how to guarantee the DP of FL with the least injected noises and energy consumed poses great challenges. Fortunately, some communication efficient FL approaches and even the nature of wireless communications contribute to the DP preservation of training data on mobile devices in FL [26], [27], [28], [29], i.e., both can help to reduce the amount of artificially injected noises while guaranteeing the same DP level. In particular, communication efficiency and privacy protection are not separate but actually closely related with each other in the context of distributed machine learning. Originally designed for communication efficiency purposes, gradient compression techniques, including both gradient quantization [26], [27] and gradient sparsification, also have amplification impacts on DP [28], [29]. Moreover, additive white Gaussian noises (AWGN) in wireless channels can serve as free noise resources for (ϵ, δ)-DP implementation [30]. While all the discussed works have analyzed the trade-off between the learning accuracy and privacy guarantee, the impacts of gradient compression and AWGN on energy consumption in privacy preserving FL over mobile devices were not taken into consideration.

Inspired by those intriguing observations, in this paper, we propose to implement DP in a novel piggyback manner and develop a corresponding DP preserving and energy efficient FL over mobile devices. The proposed piggyback DP approach jointly leverages the amplified DP impacts of gradient compression techniques, including gradient quantization and sparsification, and free AWGN in wireless communication channels to reach the targeted DP guarantee with the least amount of artificially injected noises in FL. Based on the piggyback DP mechanism, we theoretically analyze the FL convergence, formulate the joint transmission power and gradient compression control optimization with the objective of minimizing participating mobile devices' total energy consumption and develop feasible solutions. Our salient contributions are summarized as follows.

- We propose to jointly exploit gradient compression techniques, including gradient quantization and gradient sparsification, AWGN in wireless channels, and transmission power control to piggyback the DP in FL. The goal is to best utilize the DP amplification impacts of gradient compression techniques and free Gaussian noise resources in channels to reach the target DP for mobile devices with the least amount of artificially injected noises in FL.
- Under the proposed piggyback DP framework, we study the optimal position to artificially inject noises, i.e., adding noises before gradient compression or adding noises after gradient compression. Our theoretical analysis shows that for the same DP guarantee (i.e, the same ϵ), "adding noises before gradient compression" incurs fewer noises than "adding noises after gradient compression".
- Based on the piggyback DP approach, we further analyze the FL convergence and formulate the energy efficient and differentially private FL (EE-DP-FL) problem into mixed-integer nonlinear programming (MINLP) optimization, whose goal is to minimize the overall energy consumption of participating mobile devices in FL. An efficient iterative algorithm is proposed with low complexity, in which we derive new closed-form solutions for the power allocation and artificially injected DP noise.
- We conduct extensive experiments to verify the proposed piggyback DP approach in both our lab and dark room. We also employ different datasets to evaluate the performance of the proposed FL scheme via our in-lab FL testbed consisting of RTX 8000 as the FL aggregator, and NVIDIA Jetson TX2s concatenated by Universal Software Radio Peripherals (USRPs) as mobile devices. The results demonstrate its superiority to peer designs in terms of energy efficiency and privacy preservation.

The rest of this paper is organized as follows. The related work is discussed in Section II. In Section III, FL system, wireless model and DP preliminaries are presented. The proposed piggyback approach and main theoretical results are introduced in Section IV. Section V provides the problem formulation and feasible solutions. In Section VI, the performance is evaluated and experimental results are analyzed. Conclusions are made in Section VII.

II. Related Work

Energy Efficiency of FL: Recognizing that training large-scale FL models over mobile devices can be energy intensive, several research efforts have been made to reduce these costs through device scheduling [13], network optimization [17] and resource utilization optimization [14], [20], [22]. In particular, Li et al. [20] utilized gradient sparsification to improve

communication efficiency and device heterogeneity to maximize the energy efficiency of FL. Similarly, Chen et al. [21] investigated gradient quantization as a way to increase the energy efficiency of FL across heterogeneous mobile devices. Our work, however, varies from theirs in that we additionally investigate how to assure DP for sparse and quantized model updates and present a theoretical convergence result for the learning algorithm with energy reduction and privacy protection guarantee.

Differential Privacy in FL: Several works have been proposed for the design of differentially private FL to protect the data privacy of users from model inversion attacks. Mao et al. in [31] and Wang et al. in [32] have focused on the privacy leakage issue of publishing well-trained deep neural network models and proposed differentially private model publishing approaches to address the concern. However, both of them require that the servers have a well-trained DNN model. Different from them, FL server doesn't have any training dataset or well-train models. Hence, most existing DP schemes in FL usually inject artificial DP noise into model parameters continuously during the training phase. Wei et al. [33], for example, investigated local DP mechanisms in which each participating device intentionally adds Gaussian noises before uploading them to the server for aggregation. Because of the high complexity of today's state-of-the-art DL models, it incurs high communication costs. Furthermore, significant amounts of local DP noise are required to provide a strong privacy guarantee, resulting in poor model accuracy. To address these concerns, research works such as [27], [28], [34] aimed to improve both the communication cost and the utility under rigorous privacy guarantees. In particular, Agarwal et al. [27] enabled each device to first quantize the model gradients and then inject binomial noise to ensure DP. However, it made no attempt to analyze if the order in which noise is added is ideal. Moreover, most of them fail to take into account communication channel noise [35].

Recent studies [30], [36], [37], [38] have shown that channel noise can be used as a source of DP guarantee in FL. For example, Liu et al. in [30] investigated the circumstances necessary to provide "free" DP via effective power allocation. The power allocation is tuned to maximize the convergence rate under certain privacy budgets. However, the experiments have demonstrated that it is effective with a significantly large privacy budget that can give only limited privacy protection. Mohamed et al. [36] developed a user sampling strategy in which only a fraction of the participating users were required to inject artificial noises prior to transmission. Thus, the same level of protection can be achieved with less perturbation, hence improving the learning accuracy. All the discussed works mainly focus on improving utility within a given privacy budget and have largely disregarded their impact on energy efficiency in differentially private FL training, which will be addressed in this study.

Differently from the works cited above, we jointly consider gradient quantization, gradient sparsification, and AWGN in wireless channels and propose a piggyback DP approach. Our proposed piggyback DP framework accomplishes two goals concurrently: 1) Participating devices protect their privacy by injecting the fewest artificial noises; 2) Participating devices collaboratively learn the FL model while using the least amount of energy. To this end, we undertake a theoretical study to determine the ideal place for artificially injecting noises. In addition, most existing work on FL are based on simulations, whereas we implement our algorithm in an actual hardware prototype with resource-constrained devices.

III. System Model and Preliminaries

A. System Model

We consider a FL system consisting of one mobile edge server (e.g., base station or gNodeB) as the FL aggregator and $\mathcal{N} = \{1, \ldots, N\}$ mobile devices as FL clients. Each device $i \in \mathcal{N}$ has its private dataset $\mathcal{D}_i$ with $D_i = |\mathcal{D}_i|$ training data samples. All mobile devices collaboratively train a global learning model under the coordination of the mobile edge server via a shared noisy channel. The goal is to learn the optimal $\boldsymbol{w} \in \mathbb{R}^d$ that minimizes the following *global* empirical risk $f(\boldsymbol{w})$ over N mobile devices as

$$f(\boldsymbol{w}^*) = \min_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ f(\boldsymbol{w}) := \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{w}) \right\}, \tag{1}$$

where d represents the model dimension and $f_i(\boldsymbol{w}) := \frac{1}{D_i} \sum_{j=1}^{D_i} f_i^j(\boldsymbol{w}, \theta_j)$ is the *local* loss function of mobile device i. Federated learning algorithms generally solve (1) by following a simple three-step protocol [1]. In the first step, participating mobile devices download the latest global model from the mobile edge server. Next, the devices improve the downloaded global model based on their local datasets. Finally, the mobile edge server periodically aggregates the local model updates from those devices and broadcasts the updated global model back to them.

Let $t \in \{0, \ldots, T-1\}$ be index of training iterations, H be the number of local updates, and $k \in \{1, \ldots, T/H\}$ be index of global communication rounds. Mathematically, the mobile edge server initializes the global model with $\boldsymbol{w}^0$. At each global communication round k (i.e., training iterations index $t = Hk$), the mobile edge server broadcasts the global model $\boldsymbol{w}^t$ to all mobile devices. Then, mobile device $i \in \mathcal{N}$ runs H local iterations of stochastic gradient descent (SGD) method in parallel, computes the differential between the local updated model $\boldsymbol{w}_i^{t+H}$ and the global model $\boldsymbol{w}^t$ as $\boldsymbol{g}_i^t = (\boldsymbol{w}^t - \boldsymbol{w}_i^{t+H})/\gamma$, and transmits $\boldsymbol{g}_i^t$ to the mobile edge server. After that, the mobile edge server aggregates $\boldsymbol{g}_i^t$ to improve the global model $\boldsymbol{w}^{t+H} = \boldsymbol{w}^t - \gamma/N \sum_{i \in \mathcal{N}} \boldsymbol{g}_i^t$. This procedure repeats until FL converges.

Since $\boldsymbol{g}_i^t$ typically has a large size (e.g., millions of model parameters for modern deep neural network models) and should be transmitted frequently from mobile devices to the mobile edge server during FL training, gradient compression techniques are in dire need for communication efficiency and further energy efficiency purposes of mobile devices. Besides, as introduced in Section I, it is desired to have a privacy preserving mechanism to secure the FL system above against various inference attacks without degrading FL performance.

B. Wireless Communication Model

Consider that the local model update transmissions from mobile devices to the mobile edge server follow a general wireless communication model:

$$\boldsymbol{y}_i = h_i \boldsymbol{x}_i + \boldsymbol{n}_i, \tag{2}$$

where $\boldsymbol{x}_i$ is mobile device i's transmitting signal,[1] $\boldsymbol{y}_i$ is the receiving signal at the mobile edge server, and $\boldsymbol{n}_i$ is the channel noise, which is i.i.d. and drawn from Gaussian distribution $\mathcal{N}(0, N_0\mathbf{I})$, and $h_i \geq 0$ is the channel coefficient for mobile device i. We assume a block flat-fading channel, where the channel coefficients remain constant over time and the channel state information about local channels is available[2] for mobile devices and the mobile edge server. Besides, assuming the maximum transmission of mobile device i is $P_i^{\max}$, we have

$$p_i \triangleq \mathbb{E}\left[\|\boldsymbol{x}_i\|^2\right] \leq P_i^{\max}. \tag{3}$$

C. Differential Privacy Preliminaries

Differential privacy (DP) is a compelling paradigm to protect data privacy [23], [25], which is measured by the differences in the output distribution caused by only one sample change in datasets. The (ϵ, δ)-DP is defined as follows.

Definition 1 ((ϵ, δ)-DP [23]): Given any two neighboring datasets X and Y differing in at most one data sample, a randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathbb{R}^d$ satisfies (ϵ, δ)-differential privacy if for the privacy parameters $\epsilon \geq 0$ and $\delta \in (0, 1)$ and any output $S \in range(\mathcal{M})$, we have

$$Pr[\mathcal{M}(X) = S] \leq e^{\epsilon} Pr[\mathcal{M}(Y) = S] + \delta. \tag{4}$$

The above (ϵ, δ)-DP can reduce to ϵ-DP, if $\delta = 0$. The privacy preservation level is controlled by the privacy budget, ϵ. A smaller ϵ indicates stronger privacy, i.e., less possibility for any attacker to distinguish the outputs of the randomized $\mathcal{M}$ with two different inputs. Further, we introduce another generalization of ϵ-DP, called R é nyi differential privacy (RDP) [39], which is comparable to (ϵ, δ) version but supports a tight composition analysis. That makes RDP widely used for privacy protection in distributed machine learning.

Definition 2 ((ρ, λ)-RDP): Given any two neighboring datasets X and Y differing in at most one record, a randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathbb{R}^d$ satisfies (ρ, λ)-R é nyi differential privacy if for $\rho > 1$, $\lambda > 0$ and any output $S \in range(\mathcal{M})$, we have

$$D_\rho[\mathcal{M}(X)||\mathcal{M}(Y)] \triangleq \log \mathbb{E}\left[\left(\frac{\mathcal{M}(X)}{\mathcal{M}(Y)}\right)^{\rho}\right] \leq \lambda, \tag{5}$$

where the expectation is taken over the output of $\mathcal{M}(Y)$.

[1] The relationship between $\boldsymbol{x}_i$ and $\boldsymbol{g}_i$ is characterized in Section IV-C.

[2] The discussion about imperfect channel information is in Appendix F, available online.

The following two lemmas from [39] show that the Gaussian mechanism can achieve RDP, and the connection between the RDP and (ϵ, δ)-DP, respectively.

Lemma 1 (Gaussian Mechanism for RDP): Let $f: \mathcal{D} \to \mathbb{R}^d$ be a function over datasets and Δ_2 be the ℓ_2 sensitivity of f, i.e., $\Delta_2 = \sup_{\mathcal{D} \sim \mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2$. The Gaussian mechanism is defined as: $\mathcal{M}_G(\mathcal{D}, f) = f(\mathcal{D}) + \boldsymbol{g}$, where $\boldsymbol{g}$ is sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$, and satisfies $(\rho, \rho\Delta_2^2/(2\sigma^2)) - \text{RDP}$.

Lemma 2 (From RDP to (ϵ, δ)-DP): If $\mathcal{M}$ is an $(\rho, \lambda) - \text{RDP}$, then it also satisfies $(\lambda + \frac{\log(1/\delta)}{(\rho-1)}, \delta) - \text{DP}$ for any $\delta \in (0, 1)$.

IV. Privacy Preserving FL Over Mobile Devices via the Piggyback DP Approach

In this section, we first present how to leverage gradient compression and AWGN in wireless channels to piggyback DP implementation in FL. With the piggyback DP, we then analyze the convergence of privacy preserving FL. Finally, we provide some insights about the appropriate position in the piggyback DP based FL to inject artificial noises.

A. The Piggyback DP Approach Overview

Piggyback DP approach aims to preserve the training data privacy in FL with the least artificially injected noises. The piggyback DP approach jointly leverages gradient compression (i.e., randomized sparsification and stochastic quantization) and AWGN from wireless channels to achieve this goal. Generally speaking, *gradient compression* allows us to upload part of the original information, which indicates that less information needs to be protected, and correspondingly the injected noises should be smaller. The inherent *AWGN* in wireless channels satisfy Gaussian distribution and can naturally serve as free resources for (ϵ, δ)-DP implementation via proper transmission power control. Given a target privacy budget for FL, both help to reduce the amount of artificially injected noises, and have the potential to improve DP preserving FL's performance in terms of energy efficiency and learning accuracy. The procedure of piggyback DP based FL is outlined in Algorithm 1

B. Piggyback DP From Stochastic Compression

Our gradient compression scheme includes both stochastic quantization and randomized sparsification. The former enables each mobile device to reduce the precision of the gradient vectors to update/transmit by a mapping function to represent each of its elements with a smaller quantization level (e.g., from 32 bits to 16/8/4 bits). The latter sparsifies the gradient vectors to update/transmit via keeping only a subset of elements. By using gradient compression techniques above, the size of local model updates can be efficiently reduced in each communication round during FL training.

Here, we combine a stochastic quantizer with an unbiased randomized sparsifier to compress the mobile device's dense, real-valued gradient vector (i.e., $\boldsymbol{g}$) into a sparse, low precision vector in two steps. First, we randomly select the l dimensions in the parameter vector and set the remaining dimensions to zero.

Algorithm 1: Federated Learning With Piggyback DP.

Input: Initial model $\boldsymbol{w}^0 = \boldsymbol{w}_i^0$, learning rate γ, mini-batch size B, number of training iteration T, number of local updates H
Output: $\boldsymbol{w}^T$

for $t = 0, \ldots, T-1$ **do**
  The FL server broadcasts $\boldsymbol{w}^t$ to the mobile devices.
  for device $i \in \mathcal{N}$ in parallel **do**
    Randomly sample a mini-batch $\{\xi_j\}_{j=1}^B$ from $\mathcal{D}_i$ and compute $\grave{\nabla} f_i(\widetilde{\boldsymbol{w}}_i^t) \triangleq \frac{1}{B}\sum_{j=1}^B \nabla f_i(\widetilde{\boldsymbol{w}}_i^t, \xi_j)$
    $\widetilde{\boldsymbol{w}}_i^{t+\frac{1}{2}} \leftarrow \widetilde{\boldsymbol{w}}_i^t - \gamma \grave{\nabla} f_i(\widetilde{\boldsymbol{w}}_i^t)$
    if $(t+1) \mod H \neq 0$ **then**
      $\widetilde{\boldsymbol{w}}_i^{t+1} \leftarrow \widetilde{\boldsymbol{w}}_i^{t+\frac{1}{2}}$ and $\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t$
    else
      Compute the local differential: $\widetilde{\boldsymbol{g}}_{i,\text{LDP}}^t \leftarrow (\boldsymbol{w}^t - \widetilde{\boldsymbol{w}}_i^{t+\frac{1}{2}})/\gamma + \boldsymbol{\eta}_i^t$ where $\boldsymbol{\eta}_i^t$ is the artificially injected DP noise and obtain compressed message $\grave{\boldsymbol{g}}_{i,\text{LDP}}^t$ via (8)
      Transmit a scaled $\alpha_i \grave{\boldsymbol{g}}_{i,\text{LDP}}^t$ to the FL server as in (2)
      Receive the latest model $\boldsymbol{w}^{t+1}$ and $\widetilde{\boldsymbol{w}}_i^{t+1} \leftarrow \boldsymbol{w}^{t+1}$
    end if
  end for
  if $(t+1) \mod H = 0$ **then**
    The FL server estimates the devices' local differential $\hat{\boldsymbol{g}}_i^t$ in (14) and updates the new global model as $\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - \frac{\gamma}{N}\sum_{i\in\mathcal{N}} \hat{\boldsymbol{g}}_i^t$
  else
    $\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t$
  end if
end for

Let $\mathbf{m} \in \{0,1\}^d$ be a d-dimensional randomized mask vector with l entries of 1 and $\boldsymbol{g} \in \mathbb{R}^d$ (Here, we ignore the indices to these variables for simplicity in the section). The random sparsification operation is defined as [19]

$$\boldsymbol{g}^s = \frac{1}{\theta_s} \cdot (\mathbf{m} \circ \boldsymbol{g}), \tag{6}$$

where $\circ$ denotes element-wise multiplication and $\theta_s = l/d$ is the sparificiation ratio. The stochastic quantization on each nonzero element g_j^s of vector $\boldsymbol{g}^s$ is defined as [18]

$$\mathcal{Q}(g_j^s) = \frac{\|\boldsymbol{g}^s\|_2}{Q} \cdot \begin{cases} \xi_j + 1, & \text{w.p. } \frac{Qg_j^s}{\|\boldsymbol{g}^s\|_2} - \xi_j, \\ \xi_j, & \text{otherwise,} \end{cases} \tag{7}$$

where $\xi_j = \lfloor \frac{Qg_j^s}{\|\boldsymbol{g}^s\|_2} \rfloor$ and Q denotes the number of quantization levels ($1 \leq Q \leq 32$). Here, we consider uniform quantization scheme, i.e., the range of the weight is evenly splited into Q intervals. Note that we leverage the ℓ_2 norm of model differentials as a scaling factor since it can provide the sparsity guarantee and further reduce the dimensionality of $\boldsymbol{g}$. The expected sparsity of $\mathcal{Q}(\cdot)$ is bounded by $Q(Q+\sqrt{d})$ [18].

Given the sparsifier and quantizer defined above, the proposed gradient compression scheme in this work can be written as

$$\grave{\boldsymbol{g}} \triangleq C_{q,s}(\boldsymbol{g}) = \mathcal{Q}\left(\frac{1}{\theta_s} \cdot (\mathbf{m} \circ \boldsymbol{g})\right). \tag{8}$$

Note that the number of non-zero elements of the compressed gradients $C_{q,s}(\boldsymbol{g}^s)$ is $\kappa_{q,s} = Q(Q+\sqrt{l})$ in expectation. The expected sparsity ratio is $\kappa_{q,s}/d$. Formally, we can show:

Lemma 3 (Unbiased Compression): Given a vector $\boldsymbol{g} = [g_1, g_2, \ldots, g_d]$, and the gradient compressor in (8). We have

$$\mathbb{E}_{Q,S}\left[C_{q,s}(\boldsymbol{g})\right] = \boldsymbol{g}, \tag{9}$$

$$\mathbb{E}_{Q,S}\left[\|C_{q,s}(\boldsymbol{g}) - \boldsymbol{g}\|_2^2\right] \leq (\theta_{q,s} - 1)\|\boldsymbol{g}\|_2^2, \tag{10}$$

where we define $\theta_s = l/d$, $\theta_q = \frac{\sqrt{d}}{Q}$, and thus $\theta_{q,s} = (\theta_s^{-1} + \theta_q \theta_s^{-\frac{1}{2}})$.

Proof: This is derived based on the definition of compression scheme. Please refer to the detailed proof in Appendix A in the separate supplemental file, available online.

Intriguingly, $C_{q,s}(\boldsymbol{g})$ not only contributes to communication efficiency but also has amplification impacts on DP. With the compression operator in (8), a subset of elements in gradient vectors are selected. Denote the active set $\pi = \{j : [\mathbf{m}]_j = 1\}$. Then, only the values of active elements are transmitted to the mobile edge server and the rest are manually set to be zero. In this case, the mobile devices keep the private information of inactive elements from the FL server. Therefore, only the active elements in the $\boldsymbol{g}$ are need to be protected. Hence, with gradient compression strategies θ_s and θ_q in $C_{q,s}(\boldsymbol{g})$, we found that the DP impact of artificially injected Gaussian noises σ^2 is amplified proportionally to a factor of $\kappa_{q,s}/d$. Informally, if a DP mechanism $\mathcal{M}$ makes the original $\boldsymbol{g}$ satisfy (ϵ, δ)-DP per round, such a mechanism $\mathcal{M}$ with gradient compression $C_{q,s}(\boldsymbol{g})$ can guarantee $(\mathcal{O}(\kappa_{q,s}/d \cdot \epsilon), \delta)$-DP ($\kappa_{q,s} \leq d$), which is jointly determined by gradient sparisification and quantization strategies. Obviously, $(\mathcal{O}(\kappa_{q,s}/d \cdot \epsilon), \delta)$-DP provides more privacy protection than (ϵ, δ)-DP does. Thus, with proper compression strategies, piggyback DP injects less artificial noises than the DP approachs without gradient compression, to achieve the same DP guarantee.

C. More Piggyback DP From Wireless Channel Noises

After the local model differential, $\{\boldsymbol{g}_i\}_{i=1}^N$, is compressed, they will be transmitted to the mobile edge server via noisy wireless channels. By harnessing the AWGN in wireless channels, we can further improve DP protection for free. Following Algorithm 1, mobile device i will transmit a scaled version of compressed local differential $\grave{\boldsymbol{g}}_{i,\text{LDP}}$ as

$$\boldsymbol{x}_i = \alpha_i \grave{\boldsymbol{g}}_{i,\text{LDP}} = \alpha_i C_{q,s}(\boldsymbol{g}_i + \boldsymbol{\eta}_i). \tag{11}$$

Here, the transmitted $\grave{\boldsymbol{g}}_i$ is first perturbed by the artificially injected Gaussian noises and then compressed by gradient quantization and sparsification, $\boldsymbol{\eta}_i$ is the artificially injected Gaussian noises sampled from $\mathcal{N}(0, \sigma^2\mathbf{I})$, and α_i is a scaling factor for

transmission power control. Following (2), the received signal is given as,

$$\boldsymbol{y}_i = h_i\alpha_i C_{q,s}\left(\boldsymbol{g}_i + \boldsymbol{\eta}_i\right) + \boldsymbol{n}_i, \tag{12}$$

$$= h_i\alpha_i \boldsymbol{g}_i + h_i\alpha_i\left(\boldsymbol{\varepsilon}_i + \boldsymbol{\eta}_i\right) + \boldsymbol{n}_i, \tag{13}$$

where $\boldsymbol{\varepsilon}_i$ represents the noises from the compression operator. We note that the effective noise in the received signal in (13) includes three kinds of noises: artificially injected DP noises $\boldsymbol{\eta}_i$, channel noise N_0 and the compression errors from the compression operation $C_{q,s}$, which have been derived in Lemma 3. Given the channel inversion, the mobile edge server next estimates the global model updates by averaging the local differentials from mobile devices as

$$\hat{\boldsymbol{g}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}_i + \boldsymbol{\varepsilon}_i + \boldsymbol{\eta}_i + \left(h_i\alpha_i\right)^{-1}\boldsymbol{n}_i. \tag{14}$$

Since the adversaries considered in this paper can eavesdrop or access the local differentials and attempt to infer the sensitive training data from the estimation of channel outputs. Therefore, if the DP of channel outputs are preserved, the same privacy guarantee applies to the estimated model updates according to the post-processing property of DP [23]. In the following, we focus on deriving the RDP guarantee for query in (13).

Based on Lemma 1, we need to first calculate the query sensitivity in the case of noiseless channel outputs. More specifically, the sensitivity measures the amount by which a single input data record can change the disclosed function in the worst case. Throughout this section, we make the following common assumption (See [30]). We assume that the mobile edge server knows about parameters $\{\alpha_i, \forall i\}$. Furthermore, we assume that those parameters are fixed and do not reveal any private information about the private local data. We use gradient clipping to control the sensitivity. The clipping is performed on each coordinate j and thus $|g_j| \leq Z/\sqrt{d}$. It also implies that the ℓ_2 norm of the local model updates is bounded by Z^2. With the adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$, ℓ_2 sensitivity of the query $h_i\alpha_i[\boldsymbol{g}_i(\boldsymbol{w},\mathcal{D})]_{\pi_i}$ is denoted as

$$\Delta_{2,i} = \max_{\mathcal{D},\mathcal{D}'} \left\| h_i\alpha_i[\boldsymbol{g}_i(\boldsymbol{w},\mathcal{D})]_{\pi_i} - h_i\alpha_i[\boldsymbol{g}_i(\boldsymbol{w},\mathcal{D}')]_{\pi_i}\right\|_2,$$

$$= h_i\alpha_i \max_{\mathcal{D},\mathcal{D}'} \left\| [\boldsymbol{g}_i(\boldsymbol{w},\mathcal{D})]_{\pi_i} - [\boldsymbol{g}_i(\boldsymbol{w},\mathcal{D}')]_{\pi_i}\right\|_2,$$

$$\leq 2h_i\alpha_i\sqrt{\frac{\kappa_{q,s}}{d}}Z, \tag{15}$$

We observe that the sensitivity $\Delta_{2,i}$ is proportional to the compression ratio $\sqrt{\kappa_{q,s}/d}$, i.e., the ratio of the number of non-zero elements and the original model size. Then, we give the end-to-end DP guarantee in Theorem 1.

Theorem 1: Assume the ℓ_2 norm of the local model updates is bounded by Z^2, for any $\delta \in (0,1)$ the executions of Algorithm 1 with $K = \frac{T}{H}$ global communication rounds satisfies (ϵ_i, δ)-DP for mobile device i with

$$\epsilon_i = \frac{2K\rho(h_i\alpha_i)^2\kappa_{q,s}Z^2/d}{h_i^2\alpha_i^2\sigma_i^2 + N_0} + \frac{\log(1/\delta)}{\rho - 1}, \tag{16}$$

Here, the effective DP noise at the receiving end equals to $\widetilde{\sigma}_i^2 = (h_i\alpha_i\sigma_i)^2 + N_0$.

Proof. Given the Δ_2 sensitivity in (15) and Lemma 1, the transmission scheme achieves $(\rho, \epsilon_i'(\rho))-$RDP per iteration with $\epsilon_i'(\rho) = \frac{2\rho(h_i\alpha_i)^2\kappa_{q,s}Z^2/d}{h_i^2\alpha_i^2\sigma_i^2+N_0}$. Since the algorithm has run K communication rounds, according sequentially composition [40], Algorithm 1 is $(\rho, K\epsilon_i'(\rho))-$RDP. By Lemma 2, Algorithm 1 is (ϵ_i, δ)-DP with $\epsilon_i = \epsilon_i'(\rho) + \frac{\log(1/\delta)}{\rho-1}$ and the proof is completed.

For a fixed value of δ, ϵ is computed numerically by searching an optimal ρ that minimizes ϵ. Theorem 1 shows that the privacy loss is jointly determined by artificially injected noises σ_i^2, gradient compression strategies θ_q and θ_s, noises from wireless channels N_0, and power control scaling factor α_i at mobile device i. It also indicates that the proposed piggyback DP approach has benefits to improve the privacy guarantee ϵ by a factor of $\kappa_{q,s}/d$, where $\kappa_{q,s}/d \leq 1$. Meanwhile, the channel noises N_0 further perturb the information and improve the DP protection. Thus, by jointly considering the DP amplification impacts of gradient compression and the inherent channel noises, we can significantly reduce artificially noises to inject, while guaranteeing the same DP protection.

D. Convergence Analysis of FL With Piggyback DP

In this subsection, we conduct the convergence analysis of the proposed FL algorithm with Piggyback DP. Following previous works [41], [42], we make the common assumptions.

Assumption 1: All the loss functions $f_i, \forall i$, are differentiable and their gradients are L-Lipschitz continuous: for all x and $y \in \mathbb{R}^d$, $\|\nabla f_i(x) - \nabla f_i(y)\|_2 \leq L\|x - y\|_2$.

Assumption 2: The stochastic gradient is unbiased estimator: $\mathbb{E}[\hat{\nabla} f_i(\boldsymbol{w}^t)] = \nabla f_i(\boldsymbol{w}^t)$, its variance with a mini-batch of size B is bounded: $\mathbb{E}\|\hat{\nabla} f_i(\boldsymbol{w}^t) - \nabla f_i(\boldsymbol{w}^t)\|_2^2 \leq G^2/B, \forall i \in \mathcal{N}$.

Note that this is a much weaker assumption compared to the one that has the bounded expected norm of the stochastic gradient used in [28], [34].

Theorem 2: For the proposed FL algorithm with Piggyback DP, under the above assumptions, if learning rates γ satisfies

$$1 - \gamma^2L^2H^2\theta_{q,s} - \gamma L \geq 0, \tag{17}$$

then we have

$$\mathbb{E}\left[\|\nabla f(\check{\boldsymbol{w}}_T)\|_2^2\right] \leq \frac{2(f(\boldsymbol{v}^0) - f^*)}{\gamma T} + \underbrace{2\gamma^2 L^2\frac{H^2\,G^2\theta_{q,s}}{B}}_{e_c} + \underbrace{\frac{2d\gamma^2\,L^2}{N^2}\sum_{i=1}^{N}\grave{\alpha}_i^2}_{e_{dp}}. \tag{18}$$

Here, $\check{\boldsymbol{w}}_T$ is a random variable which samples a previous parameter $\widetilde{\boldsymbol{w}}_i^t$ with probability $1/NT$ and $\grave{\alpha}_i = \sqrt{\kappa_{q,s}\sigma_i^2 + \kappa_{q,s}N_0/h_i^2\alpha_i^2}$.

Proof: Please refer to the detailed proof in Appendix B in the separate supplemental file, available online.

Here, the average expected squared gradient norm characterizes the convergence rate due to the non-convex objective in

modern learning models [42], [43], [44]. In Theorem 2, we have three terms in the convergence error: the first one is the common convergence error. The second one is the error incurred by the gradient compression strategies. The last term is the error jointly determined by the gradient compression, the artificially injected noises, and the channel noises. As compression parameters (θ_q^{-1} and θ_s) get *smaller*, it results in two effects: (1) The error e_c gets *bigger* since $\theta_{q,s} = \theta_s^{-1} + \theta_q \theta_s^{-0.5}$ is a decreasing function of (θ_q^{-1}, θ_s); (2) The error e_{dp} gets *smaller* since both $\kappa_{q,s}\theta_{q,s} = 1 + (1 + 2\theta_q\theta_s^{0.5})/(\theta_q^2\theta_s)$ and $\kappa_{q,s} = d(\theta_q^{-2} + \theta_q^{-1}\theta_s^{0.5})$ are increasing function of (θ_q^{-1}, θ_s). In other words, the smaller θ_q^{-1} and θ_s are, the less information will be released. This allows us to add less DP noise (i.e., smaller DP noise variance) to model updates and use less energy to send the model updates in each communication round.

With a proper learning rate, we can obtain the following result from Theorem 2:

Corollary 1: If we set $\gamma = \frac{\sqrt{N}}{L\sqrt{T}}$ is a constant, then we have the convergence rate for Algorithm 1 as

$$\mathbb{E}\left[\|\nabla f(\check{\boldsymbol{w}}_T)\|_2^2\right] \leq \frac{2L(f(\boldsymbol{w}^0) - f^*)}{\sqrt{NT}} + \frac{2NG^2}{BT} H^2 \theta_{q,s} + \frac{d\grave{\sigma}^2}{T}. \tag{19}$$

where $\grave{\sigma} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \grave{\sigma}_i^2}$. Note that according to (19), if we pick a fixed constant value for γ, in order to achieve e-global model convergence (i.e., satisfying $\mathbb{E}[\|\nabla f(\check{\boldsymbol{w}}_T)\|_2^2] \leq e$), the minimum number of global communication rounds

$$\frac{T_e}{H} = \mathcal{O}(NH\theta_{q,s}) + \mathcal{O}\left(\frac{\grave{\sigma}^2}{H}\right). \tag{20}$$

Proof: Please refer to the detailed proof in Appendix C in the separate supplemental file, available online.

In Corollary 1 indicates that the gradient compression strategies (θ_q^{-1} and θ_s), artificially injected noises σ^2 and AWGN in wireless channels jointly affect the convergence rate and global communication rounds. As the smaller compression (i.e., small values of θ_s and θ_q) and large $N_0/h_i^2\alpha_i^2$ cause more information distortion in the model updates (i.e., larger error introduced by gradient compressors), which requires extra communication rounds to converge. These extra communication rounds in FL may increase overall privacy loss and impair overall energy efficiency.

E. Some Insights Into the Piggyback DP Approach

Note that in Algorithm 1, the artificial noises are injected before the stochastic gradient quantization and sparsification. Is it beneficial (i.e., reducing the noises to inject while providing the target DP guarantee) to implement it in a reverse order, i.e., injecting artificial noises after performing gradient compression on the local differential? To answer this question, we formulate the alternative design as

$$\boldsymbol{y}_i = h_i\alpha_i\left(C_{q,s}\left(\boldsymbol{g}_i\right) + \boldsymbol{\eta}_i\right) + \boldsymbol{n}_i. \tag{21}$$

We want to highlight that after gradient quantization, the model differentials become discrete. If we directly inject artificial DP noises that are real numbers, the benefits of gradient quantization are lost. Hence in the alternative design, the artificially injected DP noises are sampled from a discretization of Gaussian distribution. Moreover, Ding et al. in [26] have shown that discrete Gaussian provides the same RDP as the continuous one. Hence, for this alternative design, we have the following proposition.

Proposition 1: Let the Gaussian masking noises be $\boldsymbol{\eta}_i \in \mathcal{N}_{\mathbb{Z}}(0, \sigma_i^2)$. For $\delta \in (0, 1)$, the alternative design in (21) is $\grave{\epsilon}_i = (\frac{2K\rho(h_i\alpha_i)^2\kappa_{q,s}Z^2(\theta_s^{-1}+\theta_q d^{-0.5})^2}{d(h_i^2\alpha_i^2\sigma_i^2+N_0)} + \frac{\log(1/\delta)}{\rho-1}, \delta) - \text{DP}$.

Proof: Please refer to the detailed proof in Appendix D in the separate supplemental file, available online.

Compared with the alternative design above, our proposed piggyback DP's privacy budget ϵ in Theorem 1 is smaller by $(\theta_s^{-1} + \theta_q d^{-0.5})^2$-factor. This is because the Δ_2-sensitivity in the alternative design is amplified by gradient compression with a factor of $(\theta_s^{-1} + \theta_q d^{-0.5})^2$ (where $(\theta_s^{-1} + \theta_q d^{-0.5})^2 \geq 1$), while in our design, the Δ_2-sensitivity (i.e., (15)) is not affected by the gradient compression. Thus, the amount of DP noises (artificially injected noises + AWGN) in Algorithm 1 can be smaller than this alternative design to achieve the same DP guarantee. So, "injecting noises before gradient compression" in the proposed piggyback DP approach is more favorable than "injecting noises after gradient compression".

V. Energy Minimization of Mobile Devices in Differentially Private FL

The theoretical analysis in Section IV indicates that both gradient compression (θ_q and θ_s) and power control strategies α play critical roles in FL convergence and privacy guarantee. Actually, these strategies also have great impacts on the energy consumption of mobile devices since they affect the total communication rounds and transmission payload per round. In this section, we formulate the energy efficient and differentially private FL (EE-DP-FL) problem. Given the recent findings [20] that the energy consumption of local computing and that of wireless transmissions are comparable in FL over mobile devices, we aim to develop the gradient compression and transmission power control strategies to minimize the overall energy consumption (i.e., local computing + wireless communication) of mobile devices in DP-FL.

A. Energy Models and Problem Formulation

1) Energy Models: For communication energy, we consider each mobile device leverages uncoded transmissions to upload its model differentials. Each device is allowed to compress its local updates into a sparse vector with low bit precision via (8) before transmission. Let t^{cm} be the symbol duration, which is inversely proportional to the channel bandwidth. The transmission time is

$$t^{comm} = (d\theta_s)t^{cm}, \tag{22}$$

within which the model differentials with $(d\theta_s)$ nonzero dimensions is transmitted. Accordingly, the wireless transmission energy is the product of the transmission power and transmission

time, which is given as

$$E_i^{comm} = p_i \cdot t^{comm} = p_i d\theta_s t^{cm}. \tag{23}$$

Since the gradient compression and transmission power control strategies will not affect the local computing energy per round, we simply denote the local computing energy consumption as E_i^{comp} for each global round. While the overall computing energy is affected by total training iterations, which is related to the gradient compression and power allocation strategies.

2) EE-DP-FL Optimization Formulation: The objective is to minimize the overall energy consumption of mobile devices under FL convergence and DP constraints, which can be formulated into the following optimization.

$$\min_{\substack{T,\theta_q,\theta_s,\\ \boldsymbol{\alpha},\boldsymbol{\sigma}}} \quad \frac{T}{H}\sum_{i=1}^{N}(p_i \cdot t^{comm} + E_i^{comp}) \tag{24a}$$

$$\text{s.t.} \quad \mathbb{E}\left[\|\nabla f(\check{\boldsymbol{w}}_T)\|_2^2\right] \le e_{\text{th}}, \tag{24b}$$

$$\min_{\rho} \frac{2T\rho(h_i\alpha_i)^2\kappa_{q,s}Z^2}{d\left(h_i^2\alpha_i^2\sigma_i^2+N_0\right)} + \frac{\log(\frac{1}{\delta})}{\rho-1} \le \epsilon_{\text{th}}, \forall i \in \mathcal{N}, \tag{24c}$$

$$p_i \le P_i^{\max}, \forall i \in \mathcal{N}, \tag{24d}$$

$$T \in \mathbb{Z}^+, \theta_q \in \mathcal{Q}, 0 \le \theta_s \le 1, \tag{24e}$$

$$\alpha_i \ge 0, \sigma_i \ge 0, \forall i \in \mathcal{N}, \tag{24f}$$

where $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^N$ is the power control strategy, $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^N$ is local noise injection strategy, $\theta_q = \sqrt{d}/Q$ denotes the gradient quantization strategy to determine the total number of quantization levels, and $\theta_s = l/d$ represents the sparsification ratio. Constraint (24b) guarantees the model performance where the training loss after T training iterations should be smaller than a pre-set threshold e_{th}. Due to the iterative FL training process, the overall privacy budget should be bounded by a threshold as presented in constraint (24c). Constraints (24e) and (24f) indicate that optimization variables take the values from a set of non-negative integers. We utilize the upper bound in (19) from Corollary 1 to satisfy the convergence constraint (24b). According to (19), we have derived the communication complexity of the FL with piggyback DP, which can be viewed as the lower bound of T/H. By relaxing T into continuous values, which can be rounded back to integer values later, constraint (24b) can be replaced by the following inequality,

$$\frac{T}{H} \ge \beta_1\theta_{q,s} + \frac{\beta_2}{N}\sum_{i=1}^{N}\kappa_{q,s}\left(\theta_{q,s}\sigma_i^2 + \frac{N_0}{h_i^2\alpha_i^2}\right), \tag{25}$$

where β_1 and β_2 are constants used to approximate the big-$\mathcal{O}$ in (20), which can be estimated by using a sampling set of training experimental results. Besides, according to Lemma 3 and gradient clipping, the expected power of each mobile device is upper bounded by $\mathbb{E}[p_i] = \mathbb{E}[\|\alpha_i C_{q,s}(\boldsymbol{g}_i + \boldsymbol{\eta}_i)\|^2] \le \alpha_i^2(\theta_s^{-1} + \theta_q\theta_s^{-\frac{1}{2}})(Z^2 + d\sigma_i^2)$.

Next, we rewrite the constraint (24c) into

$$\min_{\rho} \quad B_i\rho + \frac{\log(1/\delta)}{\rho-1} \le \epsilon_{\text{th}}, \forall i \in \mathcal{N}, \tag{26}$$

where $B_i = \frac{2T(h_i\alpha_i)^2\kappa_{q,s}Z^2}{d(h_i^2\alpha_i^2\sigma_i^2+N_0)} \ge 0$. The optimal ρ can be obtained by setting the following first-order derivative to zero, $\frac{\mathrm{d}}{\mathrm{d}\rho}(B_i\rho + \frac{\log(1/\delta)}{\rho-1}) = B_i - \frac{\log(1/\delta)}{(\rho-1)^2} = 0$, and we have $\rho^* = \sqrt{\frac{\log(1/\delta)}{B_i}} + 1$. We then replace ρ^* into (26), which gives

$$\frac{2T(h_i\alpha_i)^2\kappa_{q,s}Z^2}{d\left(h_i^2\alpha_i^2\sigma_i^2+N_0\right)} \le \frac{(\epsilon_{\text{th}}-1)^2}{4\log(1/\delta)}, \forall i \in \mathcal{N}. \tag{27}$$

By using the upper bound of $\mathbb{E}[p_i]$, substituting (24b) with (25) and substituting (24c) with (27), we obtain,

$$\min_{\substack{T,\theta_q,\theta_s,\\ \boldsymbol{\alpha},\boldsymbol{\sigma}}} \quad \frac{T}{H}\sum_{i=1}^{N}(\alpha_i^2(1+\theta_q\theta_s^{\frac{1}{2}})dt^{cm}(Z^2+d\sigma_i^2) + E_i^{comp}) \tag{28a}$$

$$\text{s.t.} \quad (24e), (24f), \tag{28b}$$

$$\beta_1\theta_{q,s} + \frac{\beta_2}{N}\sum_{i=1}^{N}\kappa_{q,s}\left(\theta_{q,s}\sigma_i^2 + \frac{N_0}{h_i^2\alpha_i^2}\right) \le \frac{T}{H}, \tag{28c}$$

$$\alpha_i^2(\theta_s^{-1} + \theta_q\theta_s^{-\frac{1}{2}})\left(Z^2 + d\sigma_i^2\right) \le P_i^{\max}, \forall i \in \mathcal{N}, \tag{28d}$$

$$\frac{2T(\theta_q^{-2} + \theta_q^{-1}\theta_s^{\frac{1}{2}})Z^2}{(\sigma_i^2 + N_0(h_i\alpha_i)^{-2})} \le \frac{(\epsilon_{\text{th}}-1)^2}{4\log(1/\delta)}, \forall i \in \mathcal{N}. \tag{28e}$$

Here, optimization in (28) is an approximation of the original problem in (24). It is a non-convex non-linear problem. The non-convexity arises from the multiplicative form of θ_q and θ_s, multi-dimensional $\boldsymbol{\sigma}$ and the transmission power control $\boldsymbol{\alpha}$ in both the objective function and constraints, which makes the optimization NP-hard to solve. In the following, we develop an iterative algorithm with low complexity to seek feasible solutions.

B. EE-DP-FL Feasible Solutions

The proposed iterative algorithm divides the original problem (28) into two sub-problems: 1) we first optimize (T, θ_q, θ_s) with fixed $(\boldsymbol{\alpha}, \boldsymbol{\sigma})$; 2) then $(\boldsymbol{\alpha}, \boldsymbol{\sigma})$ is updated based on the obtained (T, θ_q, θ_s) in the previous step. For the first sub-problem, we convert the sub-problem to an equivalent solvable convex problem [45] to drive the feasible solution efficiently. In the second sub-problem, we derive the closed-form solutions for optimal training iterations. The details are presented in the following subsections.

In the first sub-problem, given $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\sigma}})$, problem (28) becomes

$$\min_{T,\theta_q,\theta_s} \quad TH^{-1}\left((1+\theta_q\theta_s^{\frac{1}{2}})dt^{cm}\bar{p} + E^{cp}\right) \tag{29a}$$

$$\text{s.t.} \quad \frac{H\beta_1}{T}(\theta_s^{-1} + \theta_q\theta_s^{-\frac{1}{2}}) + \frac{H\beta_2}{T}d(\theta_q^{-2} + \theta_q^{-1}\theta_s^{\frac{1}{2}})$$

$$\cdot \left((\theta_s^{-1}+\theta_q\theta_s^{-\frac{1}{2}})\bar{\sigma}+\bar{n}\right) \le 1, \tag{29b}$$

$$\frac{\bar{p}_i}{P_i^{\max}}(\theta_s^{-1}+\theta_q\theta_s^{-\frac{1}{2}}) \le 1, \forall i \in \mathcal{N}, \tag{29c}$$

$$\frac{8\,T(h_i\alpha_i)^2(\theta_q^{-2}+\theta_q^{-1}\theta_s^{\frac{1}{2}})Z^2\log(1/\delta)}{(h_i^2\alpha_i^2\sigma_i^2+N_0)\,(\epsilon_{\text{th}}-1)^2} \le 1, \forall i \in \mathcal{N}. \tag{29d}$$

where $\bar{p}_i = \bar{\alpha}_i^2(Z^2+d\bar{\sigma}_i^2)$, $\bar{p} = \sum_{i=1}^N \bar{p}_i$, $E^{cp} = \sum_{i=1}^N E_i^{comp}$, $\bar{\sigma} = \frac{1}{N}\sum_{i=1}^N \bar{\sigma}_i^2$, $\bar{n} = \frac{1}{N}\sum_{i=1}^N N_0/(h_i^2\bar{\alpha}_i^2)$. Note that the objective in this problem can be shown to be a posynomial, which is subject to posynomial upper bound inequality constraints. Therefore, the optimization problem (29) also belongs to the geometric programming (GP).

Define $T = \exp(T')$, $\theta_q = \exp(\theta_q')$, and $\theta_s = \exp(\theta_s')$. The above GP problem can be turned into the following convex form:

$$\min_{T',\theta_q',\theta_s'} \quad \log\big(\exp(T'+\ln(d\bar{p}t^{cm}))+\exp(T'+\theta_q'+0.5\theta_s' + \ln(d\bar{p}t^{cm}))+E^{cp}\exp(T'))-\log(H) \tag{30a}$$

$$\text{s.t.} \quad \log\big(\exp(\ln(H\beta_1)-T'-\theta_s')+\exp(\ln(H\beta_1)+\theta_q' -T'-\theta_s')+\exp(\ln(H\beta_2 d\bar{n})-T'-2\theta_q') + \exp(0.5\theta_s'+\ln(H\beta_2 d\bar{n})-T'-\theta_q')+\exp(-T' + \ln(H\beta_2\, d)-\theta_s'-2\theta_q')+\exp(\ln(2H\beta_2\, d)-0.5\theta_s' -T'-\theta_q')+\exp(\ln(2H\beta_2\, d)-T')\big) \le 0, \tag{30b}$$

$$\log\big(\exp\left(-\theta_s'+\ln(\bar{p}_i)-\ln(P_i^{\max})\right)+\exp\left(\theta_q' -0.5\theta_s'+\ln(\bar{p}_i)-\ln(P_i^{\max})\right)\big) \le 0, \forall i \in \mathcal{N}, \tag{30c}$$

$$\log(\exp(-2\theta_q'+\ln(c_0))+\exp(0.5\theta_s'-\theta_q' + \ln(c_0))) \le 0, \forall i \in \mathcal{N}, \tag{30d}$$

where $c_0 = \frac{8\log(1/\delta)Z^2}{(\epsilon_{\text{th}}-1)^2(\bar{\sigma}^2+N_0/(h_i\bar{\alpha}_i))}$. It is easy to verify that the problem in (30) is a nonlinear and convex problem as the log-sum-exp function is convex [46]. Therefore, its local optimal solution is also global optimal. Hence, it can be efficiently solved optimally through primal dual interior point method [47].

In the second sub-problem, given $(\bar{T}, \bar{\theta}_q, \bar{\theta}_s)$ solved in the first step, problem (28) becomes:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\sigma}} \quad \frac{\bar{T}}{H}\sum_{i=1}^N(1+\bar{\theta}_q\bar{\theta}_s^{\frac{1}{2}})dt^{cm}(\alpha_i^2Z^2+d\alpha_i^2\sigma_i^2) \tag{31a}$$

$$\text{s.t.} \quad \frac{\beta_2}{N}\sum_{i=1}^N\frac{h_i^2\alpha_i^2\sigma_i^2\bar{\theta}_{q,s}+N_0}{h_i^2\alpha_i^2} \le \frac{\bar{T}}{H}-\beta_1\bar{\theta}_{q,s}, \tag{31b}$$

$$\alpha_i^2Z^2+d\alpha_i^2\sigma_i^2 \le P_i^{\max}\bar{\theta}_{q,s}^{-1}, \forall i \in \mathcal{N}, \tag{31c}$$

$$\frac{h_i^2\alpha_i^2\sigma_i^2+N_0}{h_i^2\alpha_i^2} \ge \frac{\bar{T}\bar{\kappa}_{q,s}Z^2(\epsilon_{\text{th}}-1)^2}{2\log(1/\delta)}, \forall i \in \mathcal{N}.. \tag{31d}$$

where $\bar{\kappa}_{q,s} = \bar{\theta}_q^{-2}+\bar{\theta}_q^{-1}\bar{\theta}_s^{0.5}$ and $\bar{\theta}_{q,s} = \bar{\theta}_s^{-1}+\bar{\theta}_q\bar{\theta}_s^{-0.5}$ given $(\bar{\theta}_q, \bar{\theta}_s)$. Here we ignore the computing energy E_i^{comp} since it is a constant value w.r.t the optimal values. Since $\frac{2\bar{T}\rho(\bar{\theta}_q^{-2}+\theta_q^{-1}\bar{\theta}_s^{0.5})Z^2}{\sigma_i^2+N_0/(h_i^2\alpha_i^2)}$ and $\epsilon_{\text{th}}-\frac{\log(1/\delta)}{\rho-1}$ are both non-negetive values, we can equivalently transform (28e) into (31d). To solve problem (31) efficiently, we introduce two auxiliary variables $a_i = \alpha_i^2\sigma_i^2$ and $b_i = \alpha_i^2$. Then we can rewrite the problem into

$$\min_{a_i,b_i} \quad \sum_{i=1}^N c_1(Z^2b_i+da_i) \tag{32a}$$

$$\text{s.t.} \quad \frac{\beta_2}{N}\sum_{i=1}^N\frac{\bar{\theta}_{q,s}h_i^2a_i+N_0}{h_i^2b_i} \le c_2, \tag{32b}$$

$$Z^2b_i+da_i \le c_3, \forall i \in \mathcal{N}, \tag{32c}$$

$$h_i^2a_i-c_4h_i^2b_i \ge -N_0, \forall i \in \mathcal{N}, \tag{32d}$$

$$a_i \ge 0, b_i \ge 0, \forall i \in \mathcal{N}, \tag{32e}$$

where $c_1 = \bar{T}(1+\bar{\theta}_q\bar{\theta}_s^{0.5})dt^{cm}H^{-1}$, $c_2 = \frac{\bar{T}}{H}-\beta_1\bar{\theta}_{q,s}$, $c_3 = \bar{\theta}_{q,s}^{-1}P_i^{\max}$, and $c_4 = \frac{\bar{T}\bar{\kappa}_{q,s}Z^2(\epsilon_{\text{th}}-1)^2}{2\log(1/\delta)}$.

The optimal solution of (32) can be derived using the following theorem.

Theorem 3: The optimal solution $(\boldsymbol{a}^*, \boldsymbol{b}^*)$ of problem (32) satisfies

$$b_i^* = \min\{b_i(\mu), b_i^{\max}\}, \tag{33}$$

and

$$a_i^* = c_4b_i^*-N_0/h_i^2, \tag{34}$$

where $b_i^{\max} = (c_3+dN_0\,h_i^{-2})(Z^2+dc_4)^{-1}$,

$$b_i(\mu) = \sqrt{\frac{\mu\beta_2(\bar{\theta}_{q,s}+1)N_0}{c_1Z^2}}, \tag{35}$$

and μ satisfies

$$\sum_{i=1}^N\frac{\beta_2(\bar{\theta}_{q,s}+1)N_0}{Nh_i^2\min\{b_i(\mu),b)i^{\max}\}} = c_2-\beta_2\bar{\theta}_{q,s}c_4. \tag{36}$$

Proof: Please refer to the detailed proof in Appendix E in the separate supplemental file, available online.

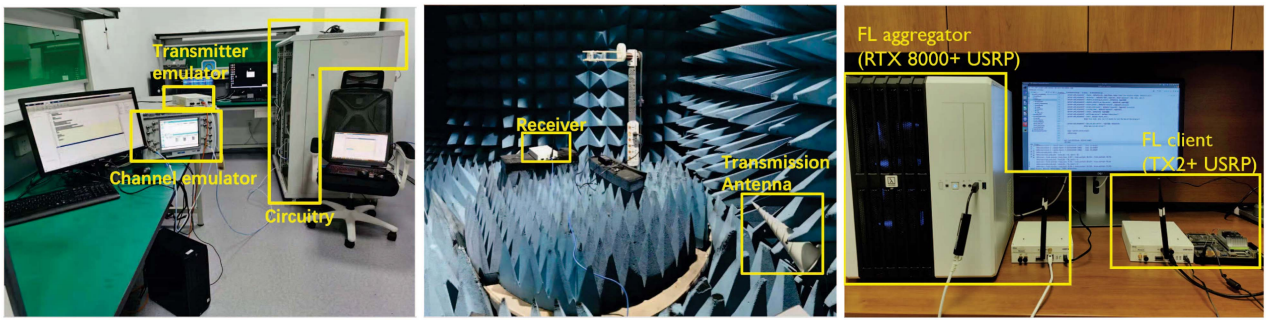
Then we can derive the optimal $\sigma_i^* = \sqrt{a_i^*/b_i^*}$ and $\alpha_i^* = \sqrt{b_i^*}$ for problem (31).

By iteratively solving problem (29) and problem (31), the algorithm that solves problem (28) is given in Algorithm 2. Since the optimal solution of problem (29) or (31) is obtained in each step, the objective value of problem (28) is nonincreasing in each step. Moreover, the objective value of problem (28) is lower bounded by zero. Thus, Algorithm 2 always converges to a feasible solution.

VI. Performance Evaluation

A. Experimental Setup

1) Devices and Platforms: As shown in Fig. 1, we set up experiments to test our proposed designs both in the darkroom and over the FL testbed in the lab. The in-lab FL testbed in Fig. 1(c) consists of the RTX 8,000 as the FL aggregator, and



(a) Darkroom configuration. (b) DF-FL tests in the darkroom. (c) EE-DP-FL testbed in the lab.

Fig. 1. Experimental setup.

Algorithm 2: Iterative Algorithm for EE-DP-FL.

1: **Input:** Initialize $(T(0), \theta_q(0), \theta_s(0), \alpha_i(0), \sigma_i(0))$ of problem (28) and set $\iota = 0$.
2: **Output:** T^*, θ_q^*, θ_s^*, $\boldsymbol{\alpha}^*$, $\boldsymbol{\sigma}^*$
3: **repeat**
4: With given $(\boldsymbol{\alpha}(\iota), \boldsymbol{\sigma}(\iota))$, obtain the optimal $(T(\iota+1), \theta_q(\iota+1), \theta_s(\iota+1))$ via (30).
5: With given $(T(\iota+1), \theta_q(\iota+1), \theta_s(\iota+1))$, obtain the optimal $(\boldsymbol{a}(\iota+1), \boldsymbol{b}(\iota+1))$ via (32).
6: Set $\sigma_i(\iota+1) = \sqrt{\frac{a_i(\iota+1)}{b_i(\iota+1)}}$ and $\alpha_i(\iota+1) = \sqrt{b_i(\iota+1)}$
7: Set $\iota = \iota + 1$
8: **until** objective value (28a) converges
9: Obtain $T^* = \lfloor T(\iota+1) \rceil$, sparsity level $l^* = \lfloor d\theta_s \rceil$, and quantization levels $Q^* = \lfloor \sqrt{d}\theta_q^{-1} \rceil$ with the minimum value of objective value in (28a)

several NVIDIA Jetson TX2s concatenated by Universal Software Radio Peripheral (USRPs) as mobile devices. We employ USRP N210 pairs equipped with WBX 50-2200 MHz Rx/Tx daughterboards as wide bandwidth transceivers that provides up to 100 mW output power. We use TX2s to conduct local computing and measure the computing energy consumption of TX2s via Nvidia profiling tools. We also use the testbeds in the lab in Fig. 1(c) and the darkroom in Fig. 1(b) to test our proposed piggyback DP approach under different wireless channel conditions. The EE-DP-FL optimization is conducted using MATLAB.

2) Learning Models and Datasets: We conduct the experiments over two datasets: MNIST and CIFAR-10. We consider FL training with non-i.i.d case, where each device contains a total number of D/N[3] training samples within 75% of the data belong to a dominant label and the remaining 25% data belong to other labels [48]. For MNIST, we use a CNN model with two 5×5 convolutional layers, a fully connected layer with 50 units and ReLu activation, and a final softmax output layer. For the CIFAR-10 dataset, we use a CNN model that consists of three 3×3 convolution layers, two fully connected layers, and a final softmax output layer. For all experiments, we consider 10 participating mobile devices, which run 20 steps of SGD in parallel. To control the sensitivity of the gradient, we adopt gradient clipping threshold technique, $Clip(g) = \text{sgn}(g)\max\{|g|, C\}$ [26]. Here, we set $C = 5$. For δ in DP, we set $\delta = 10^{-5}$ in all experiments.

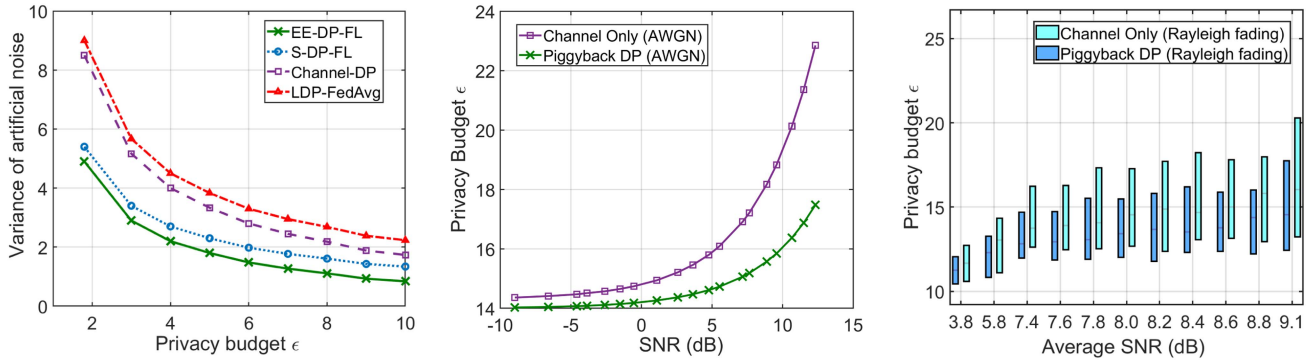
3) Peer Schemes for Comparison: We compare our proposed EE-DP-FL scheme with the following three FL schemes: 1) *LDP-FedAvg [33]*: Mobile devices artificially inject noises on their model differentials locally before transmissions; 2) *S-DP-FL:* Similar to the DP method in [28], mobile devices first perturb their model differentials by artificially injecting noises and then compress the noisy differentials via random sparsification only. Here, we slightly extend the design in [28] to the FL settings; 3) *Channel-DP* [30]: Mobile devices use AWGN from wireless channel and injected noises to perturb their local model updates. The injected DP noises in *Channel-DP* can be obtain by solving a simplified version of the problem (31) without considering gradient compression.

B. Piggyback DP Performance

Fig. 2(a) shows the comparison results of the amount of artificially injected noises, represented by the variance of artificially injected noises, under different DP approaches, when FL reaches convergence. We observe that (i) *Channel-DP* injects less noises than *LDP-FedAvg* since *Channel-DP* leverages wireless channel noises; (ii) *S-DP-FL* and *EE-DP-FL* are better than *LDP-FedAvg* and *Channel-DP* because both *S-DP-FL* and *EE-DP-FL* use gradient compression techniques, which have amplification impacts on the artificially injected noises; (iii) Our proposed *EE-DP-FL* outperforms *S-DP-FL* since the piggyback DP approach considers both the DP amplification impacts from gradient compression and free noise resources from wireless channels. So, among all DP mechanisms above, our piggyback DP approach injects the least artificial noises to achieve the same DP goal (i.e., target ϵ) in FL.

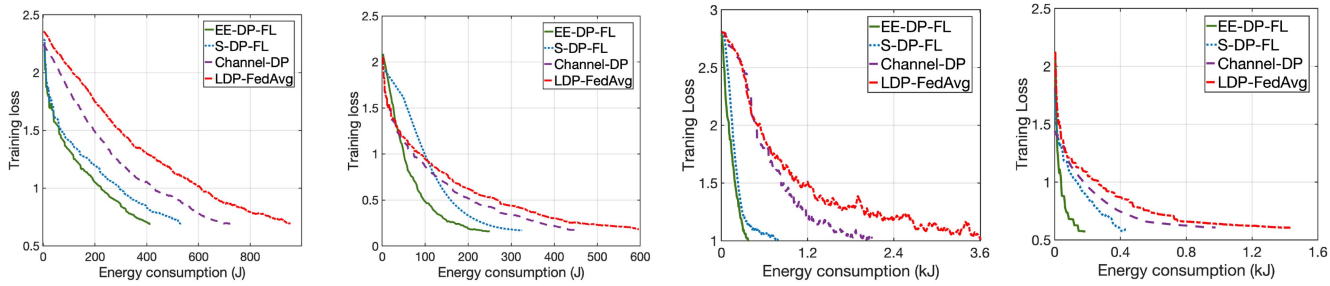
Besides, we evaluate the proposed piggyback DP approach under different channel conditions, i.e., AWGN channel and Rayleigh fading channel, in the darkroom. For the fading channel, we can tune the channel emulator in the darkroom to

[3] D is the total number of training data and N is the number of devices.



(a) Artificially injected noises under different DP approaches. (b) Piggyback DP vs Channel Only (AWGN). (c) Piggyback DP vs Channel Only (Rayleigh).

Fig. 2. Privacy performance comparison among different DP approaches. In Fig. 2(c), the gray line within each bar and the bar length represent the expected privacy budget and the variance of privacy budget under different SNR values, respectively.



(a) ϵ = 1.8, MNIST. (b) ϵ = 5, MNIST. (c) ϵ = 1.8, CIFAR-10. (d) ϵ = 5, CIFAR-10.

Fig. 3. Training Loss and energy consumption under different FL schemes when reaching the target loss e: (a) $e_{\text{th}} = 0.69$; (b) $e_{\text{th}} = 0.11$; (c) $e_{\text{th}} = 1.07$; (d) $e_{\text{th}} = 0.5$.

generate a flat and slow fading channel with 200 kHz bandwidth and 20 km/h mobile velocity. According to Theorem 1, we can compute ϵ from SNR, and achieve different privacy budgets by adjusting the transmission power. Thus, we test the uncoded transmission in the darkroom and record SNR values under both AWGN and Rayleigh fading channel conditions.

As the results shown in Fig. 2(b), for the same target privacy budget, our proposed piggyback DP approach has less SNR requirements than "Channel Only" one, which implement DP purely use channel noises. To put it in another way, the proposed piggyback DP can provide a stronger privacy protection than "Channel Only" one with the same amount of AWGN channel noises. The reason behind is that the piggyback DP approach includes gradient compression techniques, which can amplify the DP protection of the channel noises. We observe the same trend for Rayleigh fading channel as the results shown in Fig. 2(c). However, as a random process, the exact privacy budget cannot be derived, and thus we present the expected privacy budget with standard deviation under different SNR values in Fig. 2(c). In addition, compared with AWGN channel, given the same SNR, an increment of DP protection in fading channels can be observed. The potential reason is that the channel inversion process in (14) may amplify DP of channel noises, when there is a deep fading.

C. Training Loss and Energy Consumption

Fig. 3 shows the training loss and energy consumption comparison of different FL schemes when FL converges. We find that EE-DP-FL and *S-DP-FL* have better energy efficiency than *LDP-FedAvg* and *Channel-DP* due to (i) smaller size of gradients: gradient compression effectively reduces the size of gradients to update and thus saves a lot of energy for wireless communications during FL training; and (ii) smaller injected noises: due to the amplification impacts of gradient compression, EE-DP-FL and *S-DP-FL* inject less amount of artificial noises than *LDP-FedAvg* and *Channel-DP* do, while more injected noises slow down FL convergence, require more training iterations and cause more energy consumption. Furthermore, the proposed EE-DP-FL outperforms *S-DP-FL* since it not only considers gradient sparsification but also integrates gradient quantization and free noises from wireless channels into the design. As shown in Fig. 3, EE-DP-FL consumes 1.5x - 2.5x less energy than the other FL schemes when achieving the targeted

TABLE I
TESTING ACCURACY WITH DIFFERENT FL SCHEMES

Privacy budget	FL Scheme	MNIST	CIFAR
$\epsilon = 1.8$	DP-FedAvg	89.52	54.20
	Channel-DP	89.52	54.20
	S-DP-FL	91.84	63.38
	EE-DP-FL	**94.82**	**67.13**
$\epsilon = 5$	DP-FedAvg	91.12	68.31
	Channel-DP	91.12	68.31
	S-DP-FL	93.60	71.60
	EE-DP-FL	**96.87**	**74.12**

training loss. EE-DP-FL also has better test accuracy than peer FL schemes under different privacy budgets as shown in Table I.

VII. CONCLUSION & FUTURE WORK

In this paper, we have developed energy efficient and differential private FL (EE-DP-FL) via a piggyback DP approach. We have proposed a novel piggyback DP approach that effectively integrates the gradient compression (gradient quantization and sparsification) and wireless channel noises to facilitate DP implementation. The proposed piggyback DP approach can achieve the target DP with the least artificially injected noises. Based on the piggyback DP, we have analyzed the convergence of differentially private FL in the general non-convex case. Guided by those analysis results, we have formulated the energy minimization problem of differentially private FL into a mixed integer nonlinear programming and developed feasible solutions. Through extensive experiments, we have compared the proposed EE-DP-FL with peer FL schemes and demonstrated its superiority in energy efficiency, privacy protection and learning performance.

In the future, we plan to extend our work along the following promising research directions. First, as shown in the experiments, there exhibits privacy amplification in the Rayleigh fading channel. It would be interesting to incorporate the effect of fading channels in the privacy analysis and further reduce noise added for the targeted privacy budget of the piggyback DP approach. Second, besides the gradient quantization and sparsification methods employed in this paper, there are some other compression approaches, such as low rank compression [49] and model pruning [50], which can potentially be exploited to enhance piggyback DP. In particular, model pruning can not only reduce computing complexity and energy consumption by removing the connections with the lowest magnitude weights, but also introduce randomness into the FL training process, hence providing piggyback privacy guarantees [51]. Due to hardware limitations, they are not included in the current EE-DP-FL design. We plan to extend our EE-DP-FL testbed using FPGA based Xilinx Zynq [52] with hardware and software co-designs and to investigate the privacy amplification effects, model accuracies, and energy consumption of different compositions of private quantizers, sparsifiers, and pruning. Finally, we plan to explore the piggyback DP in other FL settings such as personalized FL, in which each device has its own demands for its chosen privacy budget and learning performance, rather than sharing the same global model architecture.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, Ft. Lauderdale, FL, 2017, pp. 1273–1282.

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

[3] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for Internet of Things: A comprehensive survey," 2021, *arXiv:2104.07914*.

[4] D. Ye, R. Yu, M. Pan, and Z. Han, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23920–23935, 2020.

[5] Y. Aono et al., "Privacy-preserving deep learning: Revisited and enhanced," in *Proc. Int. Conf. Appl. Techn. Inf. Secur.*, Singapore: Springer, 2017, pp. 100–110.

[6] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 2512–2520.

[7] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients–How easy is it to break privacy in federated learning?," 2020, *arXiv:2003.14053*.

[8] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1322–1333.

[9] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Secur. Symp.*, Austin, TX: USENIX Association, 2016, pp. 601–618. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer

[10] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 3–18.

[11] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2019, pp. 1–15.

[12] N. H. Tran, W. Bao, A. Zomaya, N. M. NH, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun.*, Paris, France, 2019, pp. 1387–1395.

[13] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun.*, Shanghai, China, 2019, pp. 1–7.

[14] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

[15] R. Chen, D. Shi, X. Qin, D. Liu, M. Pan, and S. Cui, "Service delay minimization for federated learning over mobile devices," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 990–1006, Apr. 2023.

[16] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning (extended version)," 2018, *arXiv:1812.11494*.

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

[18] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, 2017, pp. 1707–1718.

[19] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," 2018, *arXiv:1809.07599*.

[20] L. Li, D. Shi, R. Hou, H. Li, M. Pan, and Z. Han, "To talk or to work: Flexible communication compression for energy efficient federated learning over heterogeneous mobile edge devices," in *Proc. IEEE Conf. Comput. Commun.*, 2021, pp. 1–10.

[21] 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, Oct. 11, 2022, doi: 10.1109/TMC.2022.3213766.

[22] D. Shi, L. Li, R. Chen, P. Prakash, M. Pan, and Y. Fang, "Towards energy efficient federated learning over 5G mobile devices," 2021, *arXiv:2101.04866*.

[23] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*, vol. 9. Boston, MA, USA: Now Foundations Trends, Aug. 2014.
[24] X. Liu, H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang, "Traffic-aware multiple mix zone placement for protecting location privacy," in *Proc. IEEE Conf. Comput. Commun.*, 2012, pp. 972–980.
[25] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318. [Online]. Available: http://dx.doi.org/10.1145/2976749.2978318
[26] J. Ding, G. Liang, J. Bi, and M. Pan, "Differentially private and communication efficient collaborative learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 7219–7227.
[27] N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. Mcmahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," 2018, *arXiv:1805.10559*.
[28] X. Zhang, M. Fang, J. Liu, and Z. Zhu, "Private and communication-efficient edge learning: A sparse differential Gaussian-masking distributed SGD approach," in *Proc. 21st Int. Symp. Theory Algorithmic Found. Protocol Des. Mobile Netw. Mobile Comput.*, 2020, pp. 261–270.
[29] B. Wang, F. Wu, Y. Long, L. Rimanic, C. Zhang, and B. Li, "DataLens: Scalable privacy preserving training via gradient compression and aggregation," 2021, *arXiv:2103.11109*.
[30] 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, Jan. 2021.
[31] Y. Mao, W. Hong, B. Zhu, Z. Zhu, Y. Zhang, and S. Zhong, "Secure deep neural network models publishing against membership inference attacks via training task parallelism," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 11, pp. 3079–3091, Nov. 2022.
[32] J. Wang, W. Bao, L. Sun, X. Zhu, B. Cao, and S. Y. Philip, "Private model compression via knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1190–1197.
[33] K. Wei et al., "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3454–3469, Apr. 2020.
[34] R. Hu, Y. Gong, and Y. Guo, "Federated learning with sparsification-amplified privacy and adaptive optimization," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, 2021, pp. 1463–1469.
[35] X. Wei and C. Shen, "Federated learning over noisy channels: Convergence analysis and design examples," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1253–1268, Jun. 2022.
[36] M. S. E. Mohamed, W.-T. Chang, and R. Tandon, "Privacy amplification for federated learning via user sampling and wireless aggregation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3821–3835, Dec. 2021.
[37] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private aircomp federated learning with power adaptation harnessing receiver noise," in *Proc. IEEE Glob. Commun. Conf.*, 2020, pp. 1–6.
[38] A. Sonee and S. Rini, "Efficient federated learning over multiple access channel with differential privacy constraints," 2020, *arXiv:2005.07776*.
[39] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp.*, Santa Barbara, CA, USA, 2017, pp. 263–275.
[40] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, Xi'an, China, 2008, pp. 1–19.
[41] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Representations*, New Orleans, LA, USA, 2019, pp. 1–12.
[42] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, USA, 2019, Art. no. 698.
[43] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013.
[44] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. Int. Conf. Artif. Intell. Statist.*, PMLR, 2020, pp. 2021–2031.
[45] M. Chiang, *Geometric Programming for Communication Systems*. Boston, MA, USA: Now, 2005.
[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[47] S. Mehrotra, "On the implementation of a primal-dual interior point method," *SIAM J. Optim.*, vol. 2, no. 4, pp. 575–601, 1992.
[48] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 1698–1707.
[49] T. Vogels, S. P. Karimireddy, and M. Jaggi, "PowerSGD: Practical low-rank gradient compression for distributed optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 1278.
[50] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learn. Representations*, San Juan, Puerto Rico, 2016, pp. 1–14.
[51] Y. Huang, Y. Su, S. Ravi, Z. Song, S. Arora, and K. Li, "Privacy-preserving learning via deep net pruning," 2020, *arXiv:2003.01876*.
[52] Xilinx, "Zynq ultrascale MPSoC," Dec. 2021. [Online]. Available: https://www.xilinx.com/products/silicon-devices/soc/zynq-ultrascale-mpsoc.html

Rui Chen (Graduate Student Member, IEEE) received the BS degree from Marine Electrical Engineering College, Dalian Maritime University, Dalian, China, in 2018. She is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, University of Houston, Houston, Texas. Her major research interests include federated learning, data-driven optimization, and differential privacy.

Chenpei Huang (Student Member, IEEE) received the BS degree from the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2019. He is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, University of Houston, Houston, Texas. His research interests include wireless communication and networking, underwater communications, and cybersecurity.

Xiaoqi Qin (Member, IEEE) received the BS, MS, and PhD degrees from Electrical and Computer Engineering, Virginia Tech. She is currently an associate professor of the School of Information and Communication Engineering, Beijing University of Posts and Telecommunication (BUPT). Her research focuses on exploring performance limits of next-generation wireless networks, and developing innovative solutions for intelligent and efficient machine-type communications.

Nan Ma (Member, IEEE) received the BS and PhD degrees from the Beijing University of Posts and Telecommunications (BUPT), China, in 2002 and 2007, respectively. He is currently a professor with the School of Information and Communication Engineering, BUPT, the director of Beijing Key Laboratory of Wireless Communication Testing Technology. His research interests include wireless communication theory, testing technology, and Big Data applications.

Miao Pan (Senior Member, IEEE) received the BSc degree in electrical engineering from the Dalian University of Technology, China, in 2004, the MASc degree in electrical and computer engineering from the Beijing University of Posts and Telecommunications, China, in 2007, and the PhD degree in electrical and computer engineering from the University of Florida, in 2012, respectively. He is now an associate professor with the Department of Electrical and Computer Engineering, University of Houston. He was a recipient of NSF CAREER Award, in 2014. His research interests include Wireless/AI for AI/Wireless, deep learning privacy, cybersecurity, and underwater communications and networking. His work won IEEE TCGCC Best Conference Paper Awards 2019, and Best Paper Awards in ICC 2019, VTC 2018, Globecom 2017 and Globecom 2015, respectively. He is an editor for *IEEE Open Journal of Vehicular Technology* and an associate editor for *IEEE Internet of Things (IoT) Journal*. He has also been serving as a Technical Organizing Committee for several conferences such as TPC co-chair for Mobiquitous 2019, ACM WUWNet 2019. He is a member of AAAI, a member of ACM.

Xuemin (Sherman) Shen (Fellow, IEEE) received the PhD degree in electrical engineering from Rutgers University, New Brunswick, New Jersey, in 1990. He is a University professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular networks. He is a registered professional engineer of Ontario, Canada, an Engineering Institute of Canada fellow, a Canadian Academy of Engineering fellow, a Royal Society of Canada fellow, a Chinese Academy of Engineering Foreign member, and a distinguished lecturer of the IEEE Vehicular Technology Society and Communications Society. He received the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT), in 2021, the R.A. Fessenden Award, in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario), in 2019, James Evans Avant Garde Award, in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award, in 2015 and Education Award, in 2017 from the IEEE Communications Society (ComSoc), and Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He has also received the Excellent Graduate Supervision Award, in 2006 from the University of Waterloo and the Premier's Research Excellence Award (PREA), in 2003 from the Province of Ontario, Canada. He served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom'16, IEEE Infocom'14, IEEE VTC'10 Fall, IEEE Globecom'07, and the Chair for the IEEE ComSoc Technical Committee on Wireless Communications. He is the President of the IEEE ComSoc. He was the vice president for Technical and Educational Activities, vice president for Publications, Member-at-Large on the Board of Governors, chair of the Distinguished Lecturer Selection Committee, and member of IEEE fellow Selection Committee of the ComSoc. He served as the editor-in-chief of the *IEEE IoT Journal*, *IEEE Network*, and *IET Communications*.