Communication and Energy Efficient Wireless Federated Learning with Intrinsic Privacy

Zhenxiao Zhang, *Student Member, IEEE,* Yuanxiong Guo, *Senior Member, IEEE,*
Yuguang Fang, *Fellow, IEEE,* and Yanmin Gong, *Senior Fellow, IEEE*

Abstract—Federated Learning (FL) is a collaborative learning framework that enables edge devices to collaboratively learn a global model while keeping raw data locally. Although FL avoids leaking direct information from local datasets, sensitive information can still be inferred from the shared models. To address the privacy issue in FL, differential privacy (DP) mechanisms are leveraged to provide formal privacy guarantee. However, when deploying FL at the wireless edge with over-the-air computation, ensuring client-level DP faces significant challenges. In this paper, we propose a novel wireless FL scheme called private federated edge learning with sparsification (PFELS) to provide client-level DP guarantee with intrinsic channel noise while reducing communication and energy overhead and improving model accuracy. The key idea of PFELS is for each device to first compress its model update and then adaptively design the transmit power of the compressed model update according to the wireless channel status without any artificial noise addition. We provide a privacy analysis for PFELS and prove the convergence of PFELS under general non-convex and non-IID settings. Experimental results show that compared with prior work, PFELS can improve the accuracy with the same DP guarantee and save communication and energy costs simultaneously.

Index Terms—Federated learning, over-the-air computation, wireless edge, differential privacy, sparsification.

---✦---

1 INTRODUCTION

FEDERATED Learning (FL) is a distributed machine learning (ML) paradigm in which edge devices collaboratively learn a shared model under the orchestration of a central server while keeping data locally [1]. It has gained significant attention because of its inherent privacy preservation and higher efficiency in comparison with conventional centralized ML, which relies heavily on a trustworthy and powerful central server for model training. Specifically, in each communication round, FL consists of four basic stages: (i) the central server sends the current global model to the selected edge devices; (ii) edge devices update their models using the local training data; (iii) edge devices transmit their model updates back to the central server; and (iv) the central server aggregates devices' updates into a new global model. This process is repeated for multiple communication rounds until the global model converges with a satisfactory accuracy. Although FL only requires the transmission of model updates between edge devices and the server instead of raw data, such model transfer can become a communication bottleneck, especially when dealing with modern deep neural networks (DNNs) that has a huge number of parameters (e.g., on the order of hundreds MB, or even GB) [2], [3]. Additionally, the transmit power of edge devices is often limited in FL. Thus, it is crucial to design transmission protocols that are both communication and energy-efficient in FL.

When deploying FL over wireless edge, wireless FL [4] where edge devices under the coverage of a nearby

access point are coordinated by its co-located edge server to perform FL has been proposed. In wireless FL, the uplink transmissions for model uploading are particularly challenging due to the shared nature of wireless medium among all participating devices. Orthogonal multiple access techniques such as time-division multiple access (TDMA), code-division multiple access (CDMA), and orthogonal frequency-division multiple access (OFDMA) are often used to share the wireless spectrum among the devices. However, as the number of edge devices in wireless FL increases, the spectrum resource that can be allocated to each device decreases proportionally, leading to high latency and low quality of model update transmissions from edge devices to the server. Therefore, enabling scalable FL over wireless edge with a large number of edge devices and limited spectrum resource is very challenging.

Besides the challenge associated with communication, privacy is another core challenge in wireless FL. Although edge devices in wireless FL keep their data locally and only exchange ephemeral model updates that contain less information than raw data, this is not sufficient to guarantee data privacy. For example, by observing the model updates from an edge device, it is possible for the adversary to recover the private dataset in that device using reconstruction attack [5] or infer whether a sample is in the dataset of that device using membership inference attack [6]. Specifically, if the server is not fully trusted, it can easily infer the private information of edge devices from the received model updates during the training by employing existing attack methods. Therefore, how to defend against those advanced privacy attacks and provide rigorous privacy guarantee for each device in wireless FL without a fully trusted edge server is challenging and needs to be addressed. As the state-of-the-art privacy notion, differential privacy (DP) [7] can ensure

- Z. Zhang, Y. Guo, and Y. Gong are with the University of Texas at San Antonio, San Antonio, TX USA 78249 (e-mail: zhenxiao.zhang@utsa.edu, yuanxiong.guo@utsa.edu yanmin.gong@utsa.edu).
- Y. Fang is with the City University of Hong Kong, Kowloon Tong, Hong Kong (e-mail: my.fang@cityu.edu.hk).

formal and rigorous privacy protection in FL by adding random noise to the shared model updates. However, the existing DP mechanisms could severely degrade the model accuracy in FL [8]. How to achieve a strong DP guarantee while maintaining high model accuracy is still challenging in wireless FL.

Over-the-air computation (AirComp) provides a promising solution to addressing both of the aforementioned spectrum and privacy challenges in an integrated manner by achieving scalable and bandwidth-efficient model update aggregation in wireless FL [9]. The basic idea of AirComp is to create and leverage the inter-user interference in the multiple access channel (MAC) to boost throughput. In applying AirComp to wireless FL, devices send their model updates in an uncoded manner by directly mapping each model update parameter to a channel symbol: each device first precodes the transmitted symbols by the inverse of the uplink channel gain and then transmits the precoded symbols to the edge server in an analog manner. All the participating devices transmit simultaneously on the same channel so that their signals are aligned and decoded to obtain desired arithmetic computation results at the edge server. In comparison with the traditional orthogonal multiple access techniques where computing and communications are separately done, AirComp is a joint compute-and-communicate scheme by exploiting the fact that MAC inherently yields an additive superposed signal. Note that by using AirComp in wireless FL, the uplink transmission rate of each device does not degrade as the number of devices increases, making it more scalable. Furthermore, the superposition property of wireless channels in AirComp provides an additional benefit to privacy protection. Specifically, by using AirComp in wireless FL, the edge server will only receive from all participating devices the superposition of transmitted signals computed from their private datasets. From the privacy perspective, this makes it harder for the edge server to infer each device's private dataset since the edge server does not know the transmitted signal from each individual device. Moreover, the channel noise naturally perturbs the received signal at the edge server and contributes to privacy protection.

In this paper, we consider the problem of FL with Air-Comp subject to client-level DP guarantee and aim to design a differentially private wireless FL scheme with high model accuracy and low communication and energy overhead. The first key challenge is the large noise magnitude required for protecting the privacy of each individual model update against attackers, which is proportional to the model dimension and results in a significantly degraded model accuracy. To address it, we propose to first compress the model update for reducing its model dimension and then use the compressed model update to design the transmit signal. As shown later, both model compression and the signal-superposition nature of wireless channel can reduce the noise magnitude required to provide a certain DP guarantee and achieve higher model accuracy. The second challenge is that when adding artificial noise to the local model updates for privacy protection, devices need to consume more energy for transmitting the noisy model updates to the server via AirComp as shown in [10]. If a strong DP guarantee is required, the additional energy consumption

imposes a significant overhead on the energy-constrained edge device. Rather than adding artificial noise to the model updates when designing the transmit signal, we propose to adapt the transmit power and harness the channel noise in AirComp to naturally perturb the received signal at the server and provide intrinsic privacy.

In summary, the contributions of this paper are as follows.

- We propose a new wireless FL scheme to achieve DP with high model accuracy and low communication and energy overhead by harnessing the intrinsic channel noise, signal-superposition nature of wireless channel, and update compression in AirComp.
- We analyze the convergence and privacy properties of the proposed scheme and show that the integrated design of model compression, signal superposition, and channel noise leads to strong DP protection while maintaining high model accuracy.
- We conduct an extensive evaluation of the proposed scheme on benchmark FL datasets by comparing it with several baselines. The empirical results show that the proposed scheme can largely outperform the baselines in terms of model accuracy and energy efficiency given the same level of DP guarantee.

2 RELATED WORK

To improve the communication efficiency over wireless multiple-access channels in FL, AirComp-based FL has been proposed as a promising strategy by simultaneously transmitting model updates from all devices to the server [9], [11], [12], [13], [14]. Specifically, Zhu et al. [11] considered an analog transmission strategy in FL, where each device is scheduled for transmission depending on the channel condition to improve the learning performance in the presence of fading channels. Yang et al. [9] proposed a joint device selection and beamforming design while guaranteeing the mean squared error (MSE) to improve the learning task performance. Amiri et al. [12] proposed an algorithm in which devices pre-process the analog model updates through sparsification and quantization before transmissions. They further modify the analog communication scheme in [13], where the devices first sparsify their gradient estimates, and project the resultant error into a low-dimensional vector and transmit only the important gradient entries while accumulating the error from previous iterations. By taking gradient statistics into account, a gradient aware power control algorithm is developed in [14] to minimize the aggregation error. However, all of the aforementioned works did not consider the privacy issue when the central edge server is not trustworthy in AirComp-based FL.

To mitigate the privacy risks for the model updates shared by the devices in AirComp-based FL, some recent studies [10], [15], [16], [17], [18] adopted AirComp-based FL with DP to prevent the privacy leakage by injecting random noise into the released model updates. In particular, the previous works [10], [16] utilized artificial noise to provide privacy guarantee when the inherent channel noise is not sufficient to satisfy the DP requirement. Accordingly, these strategies may lead to more energy consumption by

TABLE 1
Summary of main notations.

Notation	Definition
i, j	Index for device
s	Index for local iteration
t	Index for global iteration
N	Total number of devices
$[N]$	$\{1, 2, \ldots, N\}$
$\mathcal{S}^t$	Set of selected devices in iteration t
$\boldsymbol{\theta}_i^{t,s}$	Local model of device i
D_i	Local dataset of device i
$f_i(\cdot)$	Local objective function of device i
Δ_i^t	Model updates of device i
$\mathbf{g}_i^{t,s}$	Stochastic gradient of device i
η	Local learning rate
τ	Aggregation period
C	Clipping threshold
$\boldsymbol{h}_i^t$	Channel gain between device i and server
$\mathbf{x}_i^t$	Transmit signal of device i
$\mathbf{z}^t$	Random Gaussian noise
$\mathbf{y}^t$	Received signal by server
P_i	Transmission power limit of device i
β^t	Alignment coefficient
C_1	Bounded gradient coefficient
ζ_i^2	Bounded variance coefficient
γ^2, κ^2	Bounded dissimilarity coefficients
ϵ, δ	differential privacy parameters

transmitting artificial Gaussian noise. In [17], an energy-efficient approach is proposed to scale down the transmit power instead of injecting artificial noise. Inspired by that, an adaptive power allocation scheme was proposed in [18] to achieve DP by utilizing intrinsic channel noise in both orthogonal multiple access (OMA) and nonorthogonal multiple access (NOMA) channels. Nevertheless, both works considered the full gradients transmission. Therefore, the communication cost is high due to the high dimension of DNNs. Meanwhile, to provide an unbiased estimation of model updates in AirComp-based FL, the updates need to be aligned by a coefficient related to the channel and power conditions before transmission [10], [17], [18]. Since both works [17] and [18] considered the full device participation, their power alignment coefficients are limited by the device with the worst channel and power condition. Furthermore, aforementioned works [10], [15], [16], [18] consider record-level DP, which protects the presence of a single data sample. As elaborated in [8], [19], it makes more sense to consider the notion of client-level DP in FL, which protects the presence of a device's entire dataset and is more challenging to achieve. Our work aims to achieve client-level DP while ensuring a high level of model accuracy in DNNs by jointly exploiting the opportunities of model compression and adaptive power control.

3 PRELIMINARIES

3.1 Federated Learning

As shown in Fig. 1, a typical FL system enables N edge devices and a server to collaboratively solve an optimization problem of the form:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} f(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{\theta}), \qquad (1)$$
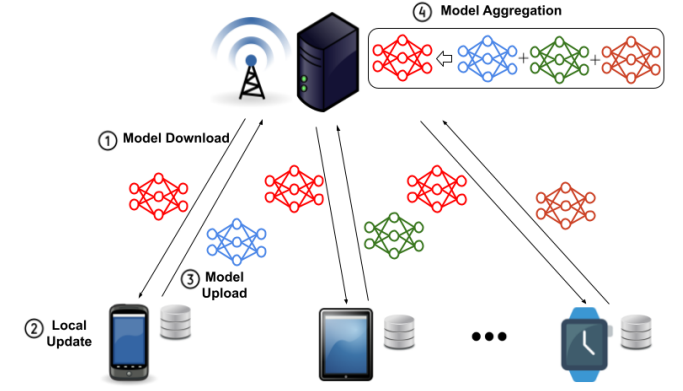
Fig. 1. FL system model.

where $f_i(\boldsymbol{\theta}) := (1/|D_i|) \sum_{\xi \in D_i} l_i(\boldsymbol{\theta}; \xi)$ is the local objective function of device i, and D_i is the local dataset of device i. Here l_i is the (possibly non-convex) loss function defined by the learning task, and ξ represents a data sample from D_i. The main notations used in the paper are summarized in Table 1.

As a classic approach in FL optimization, Federated Averaging (FedAvg) [1] was proposed to solve (1) by performing multiple steps of stochastic gradient descent (SGD) on each edge device and averaging the model updates periodically on the edge server. FedAvg can save communication rounds compared with distributed SGD and has been found to be simple and effective in many convex and non-convex settings empirically. More specifically, in the t-th FL round, FedAvg first sends the global model $\boldsymbol{\theta}^t$ to a subset of edge devices $\mathcal{S}^t$ with $r = |\mathcal{S}^t|$. Then each edge device $i \in \mathcal{S}^t$ computes a local model based on its local dataset by performing τ steps of SGD:

$$\boldsymbol{\theta}_i^{t,0} = \boldsymbol{\theta}^t, \qquad (2)$$

$$\boldsymbol{\theta}_i^{t,s} = \boldsymbol{\theta}_i^{t,s-1} - \eta \mathbf{g}_i^{t,s-1}, \quad \forall s = 1, \ldots, \tau \qquad (3)$$

where $\mathbf{g}_i^{t,s-1}$ is the stochastic gradient computed on a mini-batch z_i randomly sampled from the local dataset D_i, and $\boldsymbol{\theta}_i^{t,s}$ is the local model of device i at the s-th local iteration of round t. Next, devices upload their models $\boldsymbol{\theta}_i^{t,\tau}$ or model updates $\Delta_i^t := \boldsymbol{\theta}_i^{t,\tau} - \boldsymbol{\theta}^t$ to the server, which will then aggregate them to obtain a new global model:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \frac{1}{r} \sum_{i \in \mathcal{S}^t} \Delta_i^t. \qquad (4)$$

3.2 Differential Privacy

DP provides a rigorous privacy concept to prevent the privacy leakage and has become the de-facto standard for measuring privacy risk [7]. Informally, DP mandates that the output distribution of an algorithm remains roughly the same when an individual's data is added or removed from the dataset. This ensures that an adversary with access to the statistics of the dataset cannot distinguish whether the target individual is present in the dataset or not. The formal (ϵ, δ)-DP is defined as follows.

Definition 1 ((ϵ, δ)-DP [7]). *Given privacy parameters $\epsilon > 0$ and $0 \leq \delta < 1$, a random mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if*

for all neighboring datasets $D \simeq D'$ and any subset of outputs $\mathcal{O} \subseteq range(\mathcal{M})$, we have

$$\Pr[\mathcal{M}(D) \in \mathcal{O}] \leq e^\epsilon \Pr[\mathcal{M}(D') \in \mathcal{O}] + \delta. \qquad (5)$$

When $\delta = 0$, we have ϵ-DP, or Pure DP.

The parameter ϵ is known as the *privacy budget* that controls the trade-off between privacy and utility of the mechanism $\mathcal{M}$. A smaller ϵ provides a stronger privacy guarantee but typically results in a lower utility. The parameter δ is usually set to a small value to account for a probability that the upper bound ϵ fails. The details of definition of neighboring datasets $D \simeq D'$ will be discussed in the next subsection.

The Gaussian mechanism is commonly used to achieve (ϵ, δ)-DP by injecting zero-mean Gaussian noise to the query output, the scale of which depends on the ℓ_2-sensitivity of the query function. The definition of sensitivity is given as follows.

Definition 2 (ℓ_2-sensitivity [7]). *Let $h : \mathcal{D} \to \mathbb{R}^d$ be a query function over a dataset. The ℓ_2-sensitivity of h is defined as*

$$\psi(h) := \max_{D \simeq D'} \|h(D) - h(D')\|_2 \qquad (6)$$

where D and D' are two neighboring datasets.

Theorem 1 (Gaussian Mechanism [7]). *Let $h : \mathcal{D} \to \mathbb{R}^d$ be a query function with ℓ_2-sensitivity $\psi(h)$. The Gaussian mechanism $\mathcal{M}(D) = h(D) + \mathcal{N}\left(0, \sigma^2 \mathbf{I}_d\right)$ with $\sigma^2 \geq 2\ln(1.25/\delta)\psi^2(h)/\epsilon^2$ satisfies (ϵ, δ)-DP for any $\delta \in (0, 1)$ and $\epsilon > 0$.*

In DP mechanisms, the privacy amplification property of DP [20] allows us to improve the privacy guarantees of DP algorithms without increasing the added noise. Specifically, by running a DP mechanism on a random subset of a dataset, it can provide stronger privacy than that of running on the entire dataset. The formal privacy amplification by subsampling is given as follows.

Theorem 2 (Privacy Amplification by Subsampling [20]). *Suppose a mechanism $\mathcal{M}$ is (ϵ, δ)-DP over a dataset D of size m. Consider the subsampling mechanism that given the set D outputs a sample from the uniform distribution over all subsets $D_s \subseteq D$ of size n $(n \leq m)$. Executing $\mathcal{M}$ on the subset D_s guarantees $(\epsilon', n\delta/m)$-DP, where $\epsilon' = \log(1 + n(e^\epsilon - 1)/m)$.*

3.3 Record-level and Client-level DP in FL

Depending on how the neighboring datasets are defined, the DP definition can be applied to different granularities. Some prior works [21], [22], [23], [24], [25] on differentially private FL deal with record-level DP that is defined as follows:

Definition 3 (Record-level neighboring datasets). *Two datasets $D = \{D_i\}_{i \in [N]}$ and $D' = \{D'_i\}_{i \in [N]}$ are neighboring datasets if a device j's dataset D'_j is constructed by adding or removing a single training data example ξ from D_j. i.e., $D'_j = D_j \cup \{\xi\}$ or $D'_j = D_j \setminus \{\xi\}$, and $D'_i = D_i$ for all $i \neq j$.*

Accordingly, only one training example's privacy is protected. Protecting individual examples is insufficient in many FL setting because one user may contribute many examples to the training dataset. The same example might

---

Algorithm 1 DP-FedAvg [8]

Input: Initial server model $\boldsymbol{\theta}^0$, aggregation period τ, total rounds T, sample size r, clipping threshold C, noise magnitude σ, and learning rate η.
Output: Final global model $\boldsymbol{\theta}^T$

1: **for** $t = 0, \ldots, T - 1$ **do**
2:     Uniformly sample a set $\mathcal{S}^t \subseteq [N]$ with $r = |\mathcal{S}^t|$
3:     Broadcast $\boldsymbol{\theta}^t$ to all clients in $\mathcal{S}^t$
4:     **for** each device $i \in \mathcal{S}^t$ **in parallel do**
5:         $\boldsymbol{\theta}_i^{t,0} \leftarrow \boldsymbol{\theta}^t$
6:         **for** $s = 1, \ldots, \tau$ **do**
7:             Compute a stochastic gradient $\mathbf{g}_i^{t,s-1}$ over a minibatch ξ_i sampled from D_i
8:             $\boldsymbol{\theta}_i^{t,s} \leftarrow \boldsymbol{\theta}_i^{t,s-1} - \eta \mathbf{g}_i^{t,s-1}$
9:         **end for**
10:         $\Delta_i^t \leftarrow \boldsymbol{\theta}_i^{t,\tau} - \boldsymbol{\theta}^t$
11:         $\tilde{\Delta}_i^t \leftarrow \Delta_i^t / \max\left(1, \frac{\|\Delta_i^t\|_2}{C}\right) + \mathcal{N}(0, C^2\sigma^2\mathbf{I}_d/r)$
12:     **end for**
13:     $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \frac{1}{r}\sum_{i \in \mathcal{S}^t} \tilde{\Delta}_i^t$
14: **end for**

---

be contributed multiple times by an individual user, but it should still be protected [8]. In this paper, we consider the client-level DP to protect the privacy of entire examples of a client in the training dataset. The client-level neighboring datasets in FL can be defined as follows:

Definition 4 (Client-level neighboring datasets). *Two datasets D and D' are neighboring datasets if D' is constructed by adding or removing all of the examples associated with a single client j from D, i.e., $D = \{D_i\}_{i \in [N]}$, $D' = D \cup D_j$ or $D' = D \setminus D_j$. This implies that D and D' only differ in one client's dataset.*

Intuitively, client-level DP implies that a single client's contribution would not have a significant impact on the output distribution of the DP algorithm.

3.4 DP-FedAvg: Achieving Client-level DP in FL

To provide client-level DP in FL under a "honest-but-curious" edge server, DP-FedAvg [8] can be adapted to this setting by perturbing the model updates locally before uploading them to the server. Specifically, as shown in Algorithm 1, DP-FedAvg consists of the following steps in each FL round t:

1) Server sends the global model to a randomly sampled subset of devices (lines 2-3);
2) Each device initializes its local model to be the received global model (line 5), performs τ steps of SGD (lines 6-9) and computes its local model update (line 10);
3) Each device clips the norm of model updates Δ_i^t by a threshold C and adds Gaussian noise to its bounded local model update (line 11);
4) Server aggregates the perturbed local model updates received from edge devices to update the global model (line 13).

Although DP-FedAvg ensures client-level DP, it suffers from large model accuracy degradation and high energy
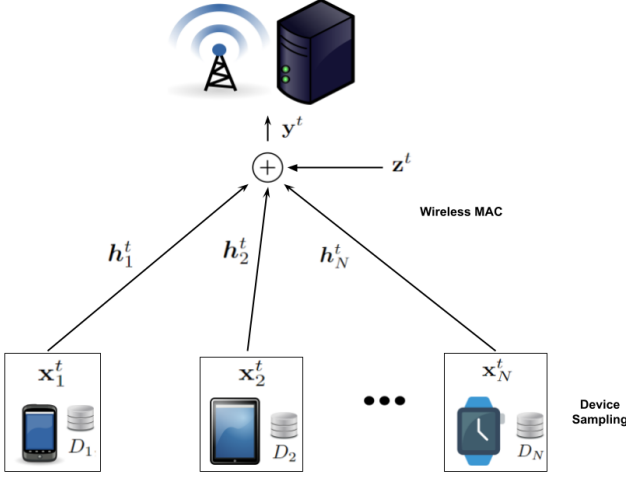
Fig. 2. Wireless communication model in AirComp.

consumption of edge devices. In this paper, we propose to integrate compression and intrinsic channel noise to achieve client-level DP while achieving high communication and energy efficiency and model accuracy in wireless FL.

4 MODEL AND PRIVACY GOAL

In this subsection, we first introduce the wireless communication model in AirComp-based FL and then describe the threat model and privacy goal.

4.1 Wireless Communication Model

As shown in Fig. 2, the edge server communicates with all devices via a wireless flat-fading channel. To simplify the analysis, we assume the download link is ideal so that the download signal is not distorted since the edge server has a much stronger transmission power compared to the devices. The channel gain between device i and the edge server at round t is denoted as a complex value $\boldsymbol{h}_i^t = |h_i^t|e^{j\phi_i^t}$. In this paper, we assume the channel state information (CSI) to be constant within each FL round, but may vary over rounds. We also assume that the CSI is perfectly known at the server and edge devices as done in the literature [17] and leave the case of imperfect CSI for future work.

We consider that the channel is split into K subcarriers under OFDMA, and each device can simultaneously transmit at most K elements of local signal $\mathbf{x}_i^t$ using the subcarriers to the server with analog transmissions in one time slot. For instance, in 5G systems, a 50 MHz channel with 15 KHz carrier spacing provides 3300 subcarriers during 10 ms [26]. Provided that we have accurate channel-gain precoding for phase compensation and strict synchronization among the participating devices [27], in the model upload stage of the t-th round, each device i first designs the transmitted signal as $e^{-j\phi_i^t}\mathbf{x}_i^t$, where $e^{-j\phi_i^t}$ is the local phase correction performed by the device i. Then, each coordinate of the local model update is assigned to a specific subcarrier and then transmitted via a wireless MAC simultaneously. By exploiting the signal-superposition nature of wireless channel, the edge server then receives the coordinates transmitted by different devices over the same set of subcarriers in the form of an aggregated sum. Specifically, the received signal $\mathbf{y}^t$ over K subcarriers at the edge server in the t-th FL round is

$$\mathbf{y}^t = \sum_{i \in \mathcal{S}^t} \boldsymbol{h}_i^t e^{-j\phi_i^t} \mathbf{x}_i^t + \mathbf{z}^t$$
$$= \sum_{i \in \mathcal{S}^t} |h_i^t| \mathbf{x}_i^t + \mathbf{z}^t, \qquad (7)$$

where $\mathbf{z}^t \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_K)$ is the random Gaussian noise with variance σ_0^2 over K subcarriers.

Since edge devices participating in FL are resource-constrained, we have a transmission power limit for each edge device i:

$$(\text{Power Limit}) \quad \mathbb{E}\|\mathbf{x}_i^t\|_2^2 \leq P_i, \quad \forall t. \qquad (8)$$

4.2 Adversarial Model and Privacy Goal

In this paper, we consider the edge server to be "honest-but-curious". That is, the edge server is curious about a specific device's local dataset and intends to infer information from the shared messages, but honestly follows the protocols involving the training process. Moreover, there may exist a third party that can observe the global model broadcasted by the server in each round. The privacy goal of this paper is to ensure that the edge server or third party cannot infer much about a device's local dataset by observing the received global model update in each round.

5 PFELS: PRIVATE FEDERATED EDGE LEARNING WITH SPARSIFICATION

In this section, we propose a new wireless FL scheme named Private Federated Edge Learning with Sparsification (PFELS) to achieve client-level DP while keeping high model accuracy and improving communication and energy efficiency at wireless edge. We also formulate the optimal power control design in PFELS as a constrained optimization problem.

5.1 Proposed Learning Scheme

To reduce communication cost and energy consumption, the key idea of PFELS is that each device sparsifies the model update first before transmitting it to the server via wireless MAC. Moreover, rather than adding artificial noise to the local model updates for privacy protection, which will consume extra transmission energy, PFELS relies on the inherent channel noise and adaptive power scaling to achieve DP in FL.

The pseudo-code for the proposed PFELS is provided in Algorithm 2. At each round t, the server randomly samples a set $\mathcal{S}^t$ of r devices with uniform probability at random without replacement (line 2). After that, a rand_k projection matrix $\mathbf{A}^t \in \{0,1\}^{k \times d}$ is generated by the server according to (9) where $k < d$ (line 3). Note that the error compensation technique [28], [29], [30] has been proposed to accelerate model convergence under compression and can be integrated into PFELS to further improve accuracy. Here, $\mathbf{A}^t$ is utilized to project a d-dimensional vector into a k-dimensional vector such that $\mathbf{A}^t \mathbf{V} \in \mathbb{R}^k$ for any $\mathbf{V} \in \mathbb{R}^d$. In particular, let $\Omega_k = \binom{[d]}{k}$ denote the set of all k-element

Algorithm 2 PFELS Algorithm.

Input: Initial server model $\boldsymbol{\theta}^0$, local update period τ, number of sampled clients r, number of total rounds T, and learning rate η.

Output: Final global model $\boldsymbol{\theta}^T$

1: **for** $t = 0, \ldots, T-1$ **do**
2:     Uniformly sample a set $\mathcal{S}^t \subseteq [N]$ with $r = |\mathcal{S}^t|$ without replacement
3:     Generate a rand_k projection matrix $\mathbf{A}^t \in \{0, 1\}^{k \times d}$
4:     Broadcast $\boldsymbol{\theta}^t$ and $\mathbf{A}^t$ to all devices in $\mathcal{S}^t$
5:     **for** each device $i \in \mathcal{S}^t$ **in parallel do**
6:         $\boldsymbol{\theta}_i^{t,0} \leftarrow \boldsymbol{\theta}^t$
7:         **for** $s = 1, \ldots, \tau$ **do**
8:             Compute a stochastic gradient $\mathbf{g}_i^{t,s-1}$ over a mini-batch ξ_i sampled from D_i
9:             $\boldsymbol{\theta}_i^{t,s} \leftarrow \boldsymbol{\theta}_i^{t,s-1} - \eta \mathbf{g}_i^{t,s-1}$
10:         **end for**
11:         $\Delta_i^t \leftarrow \boldsymbol{\theta}_i^{t,\tau} - \boldsymbol{\theta}^t$
12:         $\mathbf{x}_i^t \leftarrow \alpha_i^t \mathbf{A}^t \Delta_i^t$
13:     **end for**
14:     $\mathbf{y}^t \leftarrow \sum_{i \in \mathcal{S}^t} |h_i^t| \mathbf{x}_i^t + \mathbf{z}^t$
15:     $\hat{\Delta}^t \leftarrow \frac{(\mathbf{A}^t)^\intercal}{r\beta^t} \mathbf{y}^t$
16:     $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \hat{\Delta}^t$
17: **end for**

---

subsets of $[d]$, and a subset $\omega = \{\omega_1, \ldots, \omega_k\}$ is chosen uniformly and randomly from Ω_k. Then, the rand_k projection matrix $\mathbf{A}^t \in \mathbb{R}^{k \times d}$ is defined as follows:

$$[\mathbf{A}^t]_{m,n} := \begin{cases} 1, & \text{if } n = \omega_m \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Then, the global model $\boldsymbol{\theta}^t$ and projection matrix $\mathbf{A}^t$ are shared with all selected clients (line 4). In practice, to avoid an extra communication cost of $\mathbf{A}^t$, the coordinate set ω can be selected from Ω_k distributively by the server and clients via pseudo-random generators with the same seed.

After receiving the global model $\boldsymbol{\theta}^t$ and random projection matrix $\mathbf{A}^t$ from the server, each device $i \in \mathcal{S}^t$ initializes its local model to $\boldsymbol{\theta}^t$ and performs τ step of SGD to update its local model in parallel (lines 6-9). Then, each device i calculates its local model update Δ_i^t (lines 11) and designs the transmit signal $\mathbf{x}_i^t \in \mathbb{R}^k$ by sparsifying its local model update using $\mathbf{A}^t$ and then scaling it (lines 12) as

$$\mathbf{x}_i^t = \alpha_i^t \mathbf{A}^t \Delta_i^t, \tag{10}$$

where $\alpha_i^t \in \mathbb{R}$ is the power scaling factor to be optimized later.

After that, all devices in $\mathcal{S}^t$ send their transmit signals to the server via a channel with additive Gaussian noise $\mathbf{z}^t$. Next, according to the channel model (7) and input signals (10), the signal received at the edge server becomes

$$\mathbf{y}^t = \sum_{i \in \mathcal{S}^t} |h_i^t| \mathbf{x}_i^t + \mathbf{z}^t, \tag{11}$$

where $\mathbf{z}^t \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_k)$ is the additive Gaussian channel noise at the edge server (line 14). In order to align transmitted local model updates, we have the following power alignment constraints:

$$\text{(Power Alignment)} \qquad |h_i^t| \alpha_i^t = \beta^t, \quad \forall i, t, \tag{12}$$

where the alignment coefficient β^t is to be properly designed later in Section 7. Finally, the server can estimate a d-dimensional aggregated model update $\hat{\Delta}^t$ from the received k-dimensional signal $\mathbf{y}^t$ (line 15) as

$$\hat{\Delta}^t = \frac{(\mathbf{A}^t)^\intercal}{r\beta^t} \mathbf{y}^t = \frac{1}{r} \sum_{i \in \mathcal{S}^t} (\mathbf{A}^t)^\intercal \mathbf{A}^t \Delta_i^t + \frac{(\mathbf{A}^t)^\intercal \mathbf{z}^t}{r\beta^t}, \tag{13}$$

where $(\mathbf{A}^t)^\intercal \mathbf{A}^t \Delta_i^t \in \mathbb{R}^d$ is the sparsified local model update by only keeping the k coordinates of the orignal model update for device i. Then, the edge server uses the estimated aggregated model update to compute its global model for the next round (line 16).

We show that an unbiased estimate of the true aggregated model update $\sum_{i \in \mathcal{S}^t} \Delta_i^t / r$ in each round t can be recovered via $\hat{\Delta}^t$ as follows:

Lemma 1. *Given a parameter $k \in [d]$ and a rand_k projection matrix $\mathbf{A}^t \in \mathcal{R}^{k \times d}$ generated from the active subset ω, it holds that*

$$\mathbb{E}_\omega[\hat{\Delta}^t] = \frac{k}{d} \sum_{i \in \mathcal{S}^t} \frac{\Delta_i^t}{r}. \tag{14}$$

Proof: The proof is provided in Appendix A in the supplementary. $\square$

5.2 Optimal Power Control Design in PFELS

The design goal of power control in PFELS is to minimize the training loss while satisfying privacy and energy constraints with adaptive power control $\{\beta^t\}_{t \in [T-1]}$. We observe that the effective noise added to the model update depends on $\{\beta^t\}_{t \in [T-1]}$ according to (13) and will determine the loss $f(\boldsymbol{\theta}^T)$ after T rounds. Therefore, our goal can be represented as solving the following optimization problem:

$$\mathbf{P1} \qquad \min_{\{\beta^t\}_{t \in [T-1]}} \mathbb{E}[f(\boldsymbol{\theta}^T)] \tag{15}$$

while satisfying privacy constraint, power limit constraint (8), and power alignment constraint (12).

Solving the aforementioned problem **P1** faces several challenges. First, we need to find out how the control decisions $\{\beta^t\}_{t \in [T-1]}$ affect the loss $f(\boldsymbol{\theta}^T)$. Generally, there is no explicit mathematical expression to capture the relationship between $\{\beta^t\}_{t \in [T-1]}$ and $f(\boldsymbol{\theta}^T)$. To address it, we propose to analyze the convergence properties of PFELS in the next section and use the convergence error bound to substitute the loss function as the surrogate objective in the optimization problem. Second, we need to carefully design $\{\beta^t\}_{t \in [T-1]}$ to guarantee a certain level of client-level DP. Therefore, we provide the privacy analysis in the next section and model the privacy constraint as a function of β^t.

6 PRIVACY AND CONVERGENCE ANALYSIS

In this section, we provide the client-level DP guarantee and convergence properties of PFELS under the general non-convex and non-IID setting. Before stating our theoretical results, we make the following assumptions:

Assumption 1 (Bounded Gradient). *The stochastic gradient in Algorithm 2 for any client is bounded by a constant $C_1 > 0$, i.e., $\left\|\mathbf{g}_i^{t,s}\right\|_2 \leq C_1$ for any i, t, and s.*

Assumption 2 (Smoothness). *Each local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth for all $i \in [N]$, i.e.,*

$$\|\nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}')\|_2^2 \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \ \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d. \quad (16)$$

Assumption 3 (Unbiased Gradient and Bounded Variance). *The local mini-batch stochastic gradient is an unbiased estimator of the local gradient: $\mathbb{E}[\mathbf{g}_i(\boldsymbol{\theta})] = \nabla f_i(\boldsymbol{\theta})$ and has bounded variance: $\mathbb{E}[\|\mathbf{g}_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta})\|_2^2] \leq \zeta_i^2, \forall \boldsymbol{\theta} \in \mathbb{R}^d, i \in [N]$, where the expectation is over all the local mini-batches. We also denote $\bar{\zeta}^2 := (1/N)\sum_{i=1}^N \zeta_i^2$ for convenience.*

Assumption 4 (Lower Bounded). *There exists a constant $f_{\inf}$ such that*

$$f(\boldsymbol{\theta}) \geq f_{\inf}, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

Assumption 5 (Bounded Dissimilarity). *There exist constants $\gamma^2 \geq 1, \kappa^2 \geq 0$ such that $(1/N)\sum_{i=1}^N \|\nabla f_i(\boldsymbol{\theta})\|_2^2 \leq \gamma^2\left\|(1/N)\sum_{i=1}^N \nabla f_i(\boldsymbol{\theta})\right\|_2^2 + \kappa^2$. If the data across all devices are IID, then $\gamma^2 = 1$ and $\kappa^2 = 0$.*

Note that Assumption 1 has been commonly made in differentially private ML literature [8], [21], [31], [32], which can be ensured by the gradient clipping [21]. Assumptions 2–4 are standard in the analysis of SGD [33]. Assumption 5 is commonly used in the FL literature [34], [35] to capture the dissimilarities of local objectives due to data heterogeneity.

6.1 Privacy Analysis

In this subsection, we provide the client-level DP analysis of PFELS. Theorem 1 indicates that the DP guarantee (ϵ, δ) depends on the sensitivity of the query function on the private dataset. In PFELS, according to (10), (11) and (12), the received signal at the edge server can be rewritten as:

$$\mathbf{y}^t = \sum_{i \in \mathcal{S}^t} \beta^t \mathbf{A}^t \Delta_i^t + \mathbf{z}^t. \quad (17)$$

Thus, the edge server only knows the sum of local model updates $\sum_{i \in \mathcal{S}^t} \beta^t \mathbf{A}^t \Delta_i^t$, which depends on the private dataset. Therefore, we need to analyze the sensitivity of the sum of local model updates. Assume the client-level neighboring datasets $D = \{D_i\}_{i \in \mathcal{S}^t}$ and $D' = D \cup D_e$ or $D' = D \backslash D_e$ that differ in client e's dataset at round t, and the corresponding device datasets are $\mathcal{S}^t$ and $(\mathcal{S}^t)'$, respectively. According to Definition 2, we have the following lemma:

Lemma 2 (Sensitivity in PFELS). *Under Assumption 1, the ℓ_2-norm sensitivity of the sum of local model updates is*

$$\psi_\Delta = \max_{D \simeq D'} \|\sum_{i \in \mathcal{S}^t} \beta^t \mathbf{A}^t \Delta_i^t - \sum_{i' \in (\mathcal{S}^t)'} \beta^t \mathbf{A}^t \Delta_{i'}^t\|_2 \leq \beta^t \eta \tau C_1.$$

Proof: According to the local model updating rule (2) and (3), each device i initializes its model with $\boldsymbol{\theta}^t$ and performs τ local SGD to compute the final local model $\boldsymbol{\theta}_i^{t,\tau}$. Therefore, we have:

$$\boldsymbol{\theta}_i^{t,\tau} = \boldsymbol{\theta}^t - \eta \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1}. \quad (18)$$

Then, for the difference of the received aggregate by the server between neighboring sets $\mathcal{S}^t$ and $(\mathcal{S}^t)'$ that differ in one client index e, its ℓ_2-norm is:

$$\|\sum_{i \in \mathcal{S}^t} \beta^t \mathbf{A}^t \Delta_i^t - \sum_{i' \in (\mathcal{S}^t)'} \beta^t \mathbf{A}^t \Delta_{i'}^t\|_2 = \|\beta^t \mathbf{A}^t \Delta_e^t\|_2 \quad (19a)$$

$$= \beta^t \|\mathbf{A}^t(\boldsymbol{\theta}_e^{t,\tau} - \boldsymbol{\theta}^t)\|_2 \quad (19b)$$

$$= \beta^t \eta \|\mathbf{A}^t \sum_{s=1}^\tau \mathbf{g}_e^{t,s-1}\|_2 \quad (19c)$$

$$\leq \beta^t \eta \sum_{s=1}^\tau \|\mathbf{A}^t \mathbf{g}_e^{t,s-1}\|_2 \quad (19d)$$

where (19c) follows from (18). $\mathbf{A}^t \mathbf{g}_e^{t,s-1}$ means randomly keeping k coordinates of $\mathbf{g}_e^{t,s-1}$, and thus its upper bound depends on the upper bound of $\mathbf{g}_e^{t,s-1}$. Thus, from Assumption 1, we have the final result. $\qquad\square$

Next, we analyze the client-level DP guarantee of PFELS. Informally, according to Theorem 1, the Gaussian mechanism can provide (ϵ, δ)-DP. Based on that, considering the local model updates are performed on a random set of devices in PFELS, we can employ Theorem 2 to provide a tighter (ϵ, δ)-DP bound for PFELS. Formally, we have the following theorem:

Theorem 3. *Suppose that the client set $\mathcal{S}^t$ of size r is uniformly sampled at random without replacement from $[N]$. If the power alignment parameter β^t satisfies:*

$$C_2 \beta^t \leq \epsilon, \quad (20)$$

where

$$C_2 = \frac{2\sqrt{2}\eta\tau C_1 r\sqrt{\log\left(1.25r/(N\delta)\right)}}{N\sigma_0}, \quad (21)$$

then each round of the PFELS algorithm guarantees (ϵ, δ)-DP for any $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$.

Proof: In each round of PFELS, the intrinsic channel noise $\mathbf{z}^t$ follows the distribution of $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_k)$, and r clients are uniformly sampled at random without replacement from $[N]$. According to Theorem 1 and Theorem 2, the sensitivity $\psi_\Delta^2 \leq \sigma_0^2 \epsilon_0^2/(2\log\left(1.25/\delta_0\right))$ can guarantee $(\log(1 + (e^{\epsilon_0} - 1)r/N), \delta_0 r/N)$-DP. For $\epsilon_0 \in (0, 1)$, we have the following inequality:

$$\log\left(1 + \frac{r}{N}(e^{\epsilon_0} - 1)\right) < \frac{r}{N}(e^{\epsilon_0} - 1) < \frac{2r}{N}\epsilon_0, \quad (22)$$

where (22) follows from the facts that $\log(1 + x) < x$ for any $x > 0$ and $e^x - 1 < 2x$ for any $x \in (0, 1)$. Therefore, the sensitivity $\psi_\Delta^2 \leq \sigma_0^2 \epsilon_0^2/(2\log(1.25/\delta_0))$ guarantees $(2r\epsilon_0/N, \delta_0 r/N)$-DP. In other words, to guarantee (ϵ, δ)-DP where $\epsilon = 2r\epsilon_0/N$ and $\delta = \delta_0 r/N$, the sensitivity need to satisfy

$$\psi_\Delta^2 \leq \frac{\sigma_0^2 N^2 \epsilon^2}{8r^2 \log\left(1.25r/(N\delta)\right)}.$$

By replacing the ℓ_2-norm sensitivity ψ_Δ by Lemma 2, we have:

$$(\beta^t)^2 \eta^2 \tau^2 C_1^2 \leq \frac{\sigma_0^2 N^2 \epsilon^2}{8r^2 \log\left(1.25r/(N\delta)\right)}.$$

Let C_2 denote $2\sqrt{2}\eta\tau C_1 r\sqrt{\log\left(1.25r/(N\delta)\right)}/(N\sigma_0)$, we arrive at the conclusion. $\qquad\square$

6.2 Convergence Analysis

In this subsection, we provide the convergence analysis of PFELS under the non-convex and non-iid setting in Theorem 4. We only provide the proof sketches here and put the detailed proofs to the supplementary.

Lemma 3 (Iterate Decomposition). *Under Assumption 2, the iterate of Algorithm 2 satisfies:*

$$\mathbb{E}_t[f(\boldsymbol{\theta}^{t+1} - f(\boldsymbol{\theta}^t)] \leq L\eta^2\tau^2(\lambda_k+1)(\bar{\zeta}^2 + 2\kappa^2) + \frac{\eta\tau\lambda_k\bar{\zeta}^2}{2\pi}$$

$$+ \left(\frac{\eta L^2}{2} + 2L^3\eta^2\tau(\lambda_k+1)\right)\frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1}\right\|^2$$

$$+ \left(-\frac{\eta\tau(1-\pi)}{2} + 2L\eta^2\tau^2\gamma^2(\lambda_k+1)\right)\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2$$

$$+ \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2},$$

where $\lambda_k := 1 - k/d$, $\pi > 0$ is a constant.

   Proof: The proof is provided in Appendix B in the supplementary. □

Lemma 3 provides the composition of per-round loss drift bound. The local parameter divergence $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1}\|^2$, $i \in [N]$ provides a hint on how to derive the convergence property of PFELS. Next, we will provide an upper bound on the local parameter divergence.

Lemma 4 (Bounded Local Divergence). *Under Assumptions 2, 3 and 5, the local model difference at round t is bounded as follows:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1}\right\|^2 \leq 16\eta_l^2\tau^2\gamma^2\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2$$

$$+ 16\eta^2\tau^2\kappa^2 + 4\tau\eta^2\bar{\zeta}^2.$$

   Proof: The proof is provided in Appendix C in the supplementary. □

Lemma 4 shows the upper bound of local divergence error. Combining Lemmas 3, 4 and choosing a proper learning rate, we have the following convergence bound:

Theorem 4 (Convergence of PFELS). *Under Assumptions 2-5, if the local learning rate satisfies $\eta \leq \min\{1/(24\tau L(\lambda_k+1)\gamma^2), 1/(4\tau L\sqrt{4\gamma^2+2}), 1/12\tau L\}$, then the iterations of PFELS after T rounds satisfy:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 \leq \frac{8(f(\boldsymbol{\theta}^0) - f_{\inf})}{T\eta\tau} + 8\eta\tau L(3\kappa^2 + 2\bar{\zeta}^2)$$

$$+ 8\left(\eta\tau L(2\kappa^2 + \bar{\zeta}^2) + \frac{3\bar{\zeta}^2}{2}\right)\frac{\lambda_k}{r}$$

$$+ \frac{4Lk\sigma_0^2}{\eta\tau r^3 T}\sum_{t=0}^{T-1}\frac{1}{(\beta^t)^2}, \tag{23}$$

where $\lambda_k := 1 - k/d$.

   Proof: When $\pi < 1/2$ and

$$\eta \leq \min\{\frac{1}{4\tau L\sqrt{4\gamma^2+2}}, \frac{1}{12\tau L}, \frac{1-2\pi}{8\tau L(\lambda_k+1)\gamma^2}\},$$

we have:

$$\left(-\frac{\eta\tau(1-\pi)}{2} + 2L\eta^2\tau^2\gamma^2(\lambda_k+1)\right) \leq -\frac{\eta\tau}{4} \tag{24}$$

$$2L^3\eta^2\tau(\lambda_k+1) \leq \frac{\eta L^2}{4\gamma^2} \tag{25}$$

$$4\eta^3\tau^3L^2(2\gamma^2+1) \leq \frac{\eta\tau}{8}, \quad 12\eta\tau L \leq 1. \tag{26}$$

Substituting Lemma 4 into Lemma 3, we have:

$$\mathbb{E}_t[f(\boldsymbol{\theta}^{t+1}) - f(\boldsymbol{\theta}^t)] \leq L\eta^2\tau^2(\lambda_k+1)(2\kappa^2 + \bar{\zeta}^2) + \frac{\eta\tau\lambda_k\bar{\zeta}^2}{2\pi}$$

$$+ \left(\frac{\eta L^2}{2} + 2L^3\eta^2\tau(\lambda_k+1)\right)\sum_{s=1}^{\tau}[16\eta^2\tau^2\gamma^2\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2$$

$$+ 16\eta^2\tau^2\kappa^2 + 4\tau\eta^2\bar{\zeta}^2] + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2}$$

$$+ \left(-\frac{\eta\tau(1-\pi)}{2} + 2L\eta^2\tau^2\gamma^2(\lambda_k+1)\right)\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2$$

$$\leq L\eta^2\tau^2(\lambda_k+1)(2\kappa^2 + \bar{\zeta}^2)$$

$$+ \frac{\eta L^2(2\gamma^2+1)}{4\gamma^2}\sum_{s=1}^{\tau}[16\eta^2\tau^2\gamma^2\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2$$

$$+ 16\eta^2\tau^2\kappa^2 + 4\tau\eta^2\bar{\zeta}^2]$$

$$+ \frac{\eta\tau\lambda_k\bar{\zeta}^2}{2\pi} - \frac{\eta\tau}{4}\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2} \tag{27}$$

$$= L\eta^2\tau^2(\lambda_k+1)(2\kappa^2 + \bar{\zeta}^2) + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2}$$

$$+ \frac{\eta^3\tau^2L^2(2\gamma^2+1)(4\tau\kappa^2 + \bar{\zeta}^2)}{\gamma^2} + \frac{\eta\tau\lambda_k\bar{\zeta}^2}{2\pi}$$

$$+ \left(-\frac{\eta\tau}{4} + 4\eta^3\tau^3L^2(2\gamma^2+1)\right)\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2$$

$$\leq \left(L\eta^2\tau^2(2\kappa^2 + \bar{\zeta}^2) + \frac{\eta\tau\bar{\zeta}^2}{2\pi}\right)\lambda_k + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2}$$

$$+ \eta^2\tau^2L(12\eta\tau L + 2)\kappa^2 + \eta^2\tau^2L(3\eta L + 1)\bar{\zeta}^2$$

$$- \frac{\eta\tau}{8}\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 \tag{28}$$

$$\leq -\frac{\eta\tau}{8}\mathbb{E}_t\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 + \eta^2\tau^2L(3\kappa^2 + 2\bar{\zeta}^2)$$

$$+ \left(L\eta^2\tau^2(2\kappa^2 + \bar{\zeta}^2) + \frac{\eta\tau\bar{\zeta}^2}{2\pi}\right)\lambda_k + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2}, \tag{29}$$

where (27) holds due to (24) and (25), (28) follows from (26) and $\gamma^2 \geq 1$, and (29) follows from (26).

Rearranging the above inequality and summing it from $t = 0$ to $T - 1$, we get:

$$\sum_{t=0}^{T-1}\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 \leq \frac{8}{\eta\tau}\sum_{t=0}^{T-1}\mathbb{E}_t[f(\boldsymbol{\theta}^{t+1}) - f(\boldsymbol{\theta}^t)]$$

$$+ 8T\left(\eta\tau L(2\kappa^2 + \bar{\zeta}^2) + \frac{\bar{\zeta}^2}{2\pi}\right)\lambda_k$$

$$+ 8T\eta\tau L(3\kappa^2 + 2\bar{\zeta}^2)$$

$$+ \sum_{t=0}^{T-1}\frac{4TLk\sigma_0^2}{\eta\tau r^3(\beta^t)^2}, \tag{30}$$

where the expectation is taken over all rounds $t \in [0, T-1]$. Dividing both sides of (30) by T, one yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 \leq \frac{8(f(\boldsymbol{\theta}^0) - f_{\inf})}{T\eta\tau} + 8\eta\tau L(3\kappa^2 + 2\bar{\zeta}^2)$$
$$+ 8\left(\eta\tau L(2\kappa^2 + \bar{\zeta}^2) + \frac{\bar{\zeta}^2}{2\pi}\right)\lambda_k$$
$$+ \frac{4Lk\sigma_0^2}{\eta\tau r^3}\sum_{t=0}^{T-1}\frac{1}{(\beta^t)^2}.$$

By selecting a reasonable constant $\pi = 1/3$ (which satisfies $\pi < 1/2$), we arrive at the conclusion. $\qquad\square$

We can see that the convergence bound (23) contains three parts. The first part $8(f(\boldsymbol{\theta}^0) - f^*)/(T\eta\tau) + 8\eta\tau L(3\kappa^2 + 2\bar{\zeta}^2)$ is the *optimization error* bound in FedAvg [33]. The second part $8(\eta\tau L(2\kappa^2 + \bar{\zeta}^2) + 3\bar{\zeta}^2/2)\lambda_k/r$ is the *compression error* resulting from applying rand_k to local model updates. When the rand_k is not applied, i.e., $k = d$ and $\lambda_k = 0$, the *compression error* equals zero. The last part $4Lk/(\eta\tau r^3 T)\sum_{t=0}^{T-1}\sigma_0^2/(\beta^t)^2$ is the *privacy error*. When there is no privacy noise, i.e., $\sigma_0 = 0$, the *privacy error* is equal to zero. Both privacy error and compression error raise the error floor at convergence. The last two terms explicitly show the trade-off between *compression error* and *privacy error* in terms of compression parameter k in PFELS. As k increases, λ_k gets smaller, leading to a smaller compression error, but the privacy error increases. Therefore, in order to achieve the minimal convergence bound, we need to carefully choose an optimal parameter k to balance these two errors.

7 CONVERGENCE-OPTIMIZED POWER CONTROL UNDER DP GUARANTEE

In this section, we develop a solution method to approximately solve **P1**. First, we consider the power limit constraint (8) and power alignment constraint (12). From (10) and (12), the local transmission signal can be expressed as:

$$\mathbf{x}_i^t = \frac{\beta^t}{|h_i^t|}\mathbf{A}^t\Delta_i^t, \quad \forall i \in \mathcal{S}^t, t. \tag{31}$$

$$\mathbf{x}_i^t = \alpha_i^t\mathbf{A}^t\Delta_i^t = \frac{\beta^t}{|h_i^t|}\mathbf{A}^t\Delta_i^t, \quad \forall i \in \mathcal{S}^t, t. \tag{32}$$

According to the power limit constraint $\mathbb{E}\|\mathbf{x}_i^t\|_2^2 \leq P_i, \forall i, t$, the power alignment coefficients $\{\beta^t\}_{t\in[T-1]}$ are constrained as follows:

$$(\beta^t)^2 \leq \min_{i\in\mathcal{S}^t}\frac{|h_i^t|^2 P_i}{\mathbb{E}\|\mathbf{A}^t\Delta_i^t\|_2^2}, \quad \forall t. \tag{33}$$

Here, to satisfy the power limit constraints, we can use the following lemma to approximate the power consumption $\mathbb{E}\|\mathbf{A}^t\Delta_i^t\|_2^2$ under rand_k sparsification.

Lemma 5 (Bounded Local Updates for rand_k). *Under Assumption 1, given a local update $\Delta_i^t \in \mathbb{R}^d$ and rand_k sparsification with random projection matrix $\mathbf{A}^t$, we have*

$$\mathbb{E}\|\mathbf{A}^t\Delta_i^t\|_2^2 \leq \frac{k}{d}\eta^2\tau^2 C_1^2, \quad \forall i \in \mathcal{S}^t, t. \tag{34}$$

Proof: The proof is provided in Appendix F in the supplementary. $\qquad\square$

Then, we substitute the objective function and the DP constraint in **P1** with the convergence upper bound (23) w.r.t. $\{\beta^t\}_{t\in[T-1]}$ in Theorem 4 and the client-level DP result (20) in Theorem 3, respectively. Therefore, we can approximate Problem **P1** as follows:

$$\textbf{P2} \quad \min_{\{\beta^t\}_{t\in[T]}} \quad \sum_{t=0}^{T-1}\frac{1}{(\beta^t)^2} \tag{35a}$$

$$\text{s.t.} \quad C_2\beta^t \leq \epsilon, \quad \forall t \tag{35b}$$

$$0 < \beta^t \leq \min_{i\in\mathcal{S}^t}\frac{|h_i^t|\sqrt{dP_i}}{C_1\eta\tau\sqrt{k}}, \quad \forall t. \tag{35c}$$

Note that **P2** can be readily solved as shown in the following result:

Theorem 5. *The optimal solution to Problem **P2** is given by:*

$$(\beta^t)^* = \min_{i\in\mathcal{S}^t}\{\frac{|h_i^t|\sqrt{dP_i}}{C_1\eta\tau\sqrt{k}}, \frac{\epsilon}{C_2}\}, \quad \forall t. \tag{36}$$

Proof: The objective (35a) is monotonically decreasing with respect to $\{\beta^t\}_{t\in[T-1]}$, and the constraints are decoupled across time. From the upper bounds of (35b) and (35c), we arrive at (36). $\qquad\square$

8 NUMERICAL EVALUATION

In this section, we conduct extensive experiments on common FL benchmark datasets to verify the performance of the proposed scheme.

8.1 Experimental Setup

We consider a wireless FL system with 1000 devices and one central server. In each round, the server uniformly samples 32 devices to participate in the training process. To get a fair comparison with the baselines in the experimentation, we select the same MNIST-based dataset (FEMNIST) and CIFAR-10 dataset and corresponding model architectures in this paper as prior work in literature such as [31], [36], [37], [38], [39]. Note that while FEMNIST and CIFAR-10 datasets are considered as "solved" in the computer vision community, achieving high accuracy with a strong DP guarantee remains difficult on these datasets [40]. CIFAR-10 is an image dataset that consists of 50,000 training images and 10,000 testing images. Each device has 50 training examples and 10 test examples by partitioning 50,000 samples over 1000 devices in an IID manner. Each image has a size of 32×32 pixels and an associated class label from 10 classes. The trained model on CIFAR-10 is a modified VGG-11 with 9,750,922 parameters in total. FEMNIST is the federated version of EMNIST dataset which has 3,550 users. Each image has a size of 28×28 pixels and an associated class label from 62 classes. We remove the users with less than 100 samples and randomly choose 1,000 users from the remaining set as all the participants. Each device has 90% of images in the training set and 10% in the test set. We train a modified ResNet-18 [41] with 11,192,746 parameters in total on FEMNIST.
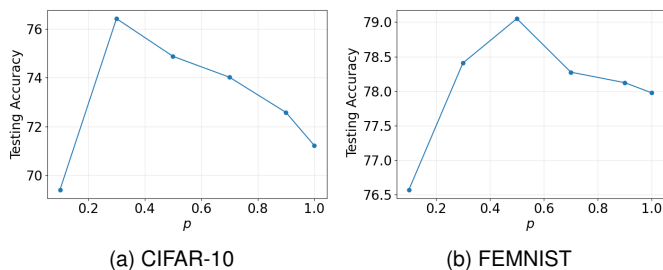
(a) CIFAR-10   (b) FEMNIST

Fig. 3. Test accuracy versus compression ratio p for PFELS under $\epsilon = 1.5$ for CIFAR-10 and $\epsilon = 1.0$ for FEMNIST.

(a) CIFAR-10   (b) FEMNIST

Fig. 4. Test accuracy vs. DP privacy budget ϵ for PFELS and baselines.

TABLE 2
Summary of results on CIFAR-10 dataset under $\epsilon = 1.5$.

Algorithm	Accuracy (%)	Subcarriers (d)	Energy cost
PFELS	76.42%	600	2.55
WFL-P	72.72%	2000	5.19
WFL-PDP	72.23%	2000	3.70

For all experiments, we use mini-batch SGD with a momentum of 0.9 to train the local model with a batch size of 50. The learning rate of each algorithm is tuned from $\{0.01, 0.05, 0.1\}$ for CIAFR-10 and $\{0.001, 0.02, 0.05\}$ for FEMNIST using grid search. Following the implementation in [42], instead of doing τ local training steps per device, we perform τ epochs of training over each device's dataset. For CIFAR-10, the local epoch is set to $\tau = 5$, and for FEMNIST, the local epoch is set to $\tau = 5$. We tune the clipping threshold over the grid $C_1 = \{1.0, 4.0, 8.0, 10.0\}$ and obtain the optimal clipping threshold, i.e., $C_1 = 1.0$ and $C_1 = 10.0$ for CIFAR-10 and FEMNIST, respectively. For all experiments, we set the privacy parameter $\delta = 1/N$.

We assume the channel gain $|h_i^t|$ follows an exponential distribution with a mean of 0.02 and limit its value within the interval $[0.0001, 0.1]$. The variance of the channel noise is set to $\sigma_0 = 1$. The maximum signal-to-noise ratio of each device i is defined as $\mathrm{SNR}_i = P_i/d\sigma_0^2$. We set the maximum SNRs of all users to uniformly sampled from 2dB to 15dB. We set the number of communication rounds $T = 2000$ and $T = 1000$ for CIFAR-10 and FEMNIST, respectively. Instead of using k, we define the *compression ratio* $p := k/d$. We run each experiment with 5 random seeds and report the average. All algorithms are implemented using PyTorch on an Ubuntu server with 4 NVIDIA RTX 8000 GPUs.

To show the effectiveness of PFELS, we simulate it under different constraints and compare it with several baselines. All baselines adopt the same uniform sampling strategy in each FL round.

- WFL-P: *Wireless FL without sparsification and DP constraint.* This baseline optimizes (35a) while transmitting the full model updates and ignoring the DP constraint (35b). According to Theorem 5, when there are no requirements for compression and DP, the optimal power control decision is given by:

$$\beta^t = \min_{i \in \mathcal{S}^t}\{\frac{|h_i^t|\sqrt{P_i}}{C_1 \eta \tau}\}, \quad \forall t. \tag{37}$$

  This baseline belongs to the standard AirComp-based FL algorithm and mimics the existing wireless FL method without privacy consideration [43].
- WFL-PDP: *Wireless FL without sparsification and with DP constraint.* This baseline aims to optimize (35a) while transmitting the full model update in each FL round and satisfying the DP constraint (35b). Similar
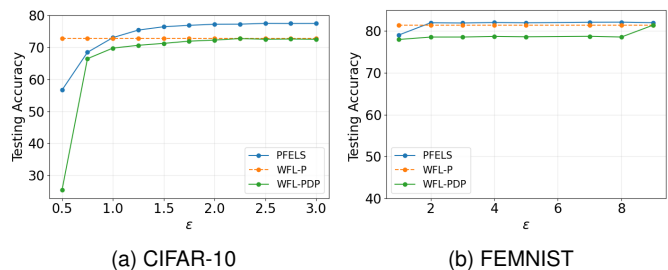
to Theorem 5, the optimal power control decision without sparsification is given by:

$$\beta^t = \min_{i \in \mathcal{S}^t}\{\frac{|h_i^t|\sqrt{P_i}}{C_1 \eta \tau}, \frac{\epsilon}{C_2}\}, \quad \forall t. \tag{38}$$

This baseline mimics the state-of-the-art energy-efficient wireless FL method with privacy consideration [17].

8.2 Experimental Results

8.2.1 Impact of Compression Ratio in PFELS

We first evaluate the impact of compression ratio on the performance of PFELS and show the results in Fig. 3 over both CIFAR-10 and FEMNIST datasets. From the figure, we can observe that the test accuracy of PFELS first increases and then decreases as the compression ratio p increases from 0.1 to 1.0. This is consistent with the analysis in Theorem 4. Note that as p increases, the *compression error* decreases, but the *privacy error* increases. Specifically, when the compression ratio is relatively small, i.e., $p = 0.1$, the magnitude of *compression error* is large and dominates the total convergence error, resulting in a higher training loss and lower testing accuracy. As p increases, the *compression error* reduces and leads to a lower loss and higher accuracy. However, when p exceeds a threshold, i.e., $p = 0.3$ for CIFAR-10 and $p = 0.5$ for FEMNIST, the *privacy error* becomes the dominant term in the total convergence error and keeps increasing as p increases, resulting in a higher training loss and lower test accuracy. Therefore, one needs to carefully choose the optimal p in practice to balance the privacy error and compression error. In the rest of the experiments, we always choose $p = 0.3$ for CIFAR-10 and $p = 0.5$ for FEMNIST in PFELS.

8.2.2 Privacy-Accuracy Tradeoff in PFELS and Baselines

In this section, we compare the testing accuracies of PFELS and baselines by varying the privacy budget ϵ as shown in Fig. 4. From the figure, we have the following observations. First, both the PFELS and WFL-PDP have a higher model accuracy as ϵ increases. This is due to the fact that a higher

TABLE 3
Summary of results on FEMNIST dataset under $\epsilon = 2.0$.

Algorithm	Accuracy (%)	Subcarriers (d)	Energy cost (1e11)
PFELS	81.97%	500	1.90
WFL-P	81.43%	1000	2.89
WFL-PDP	78.55%	1000	2.22
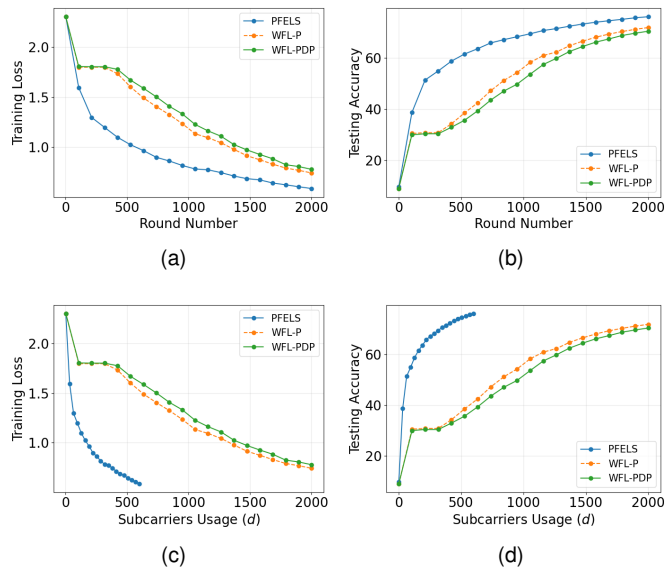
Fig. 5. Training performance versus FL round and communication cost for CIFAR-10 under $\epsilon = 1.5$.

Fig. 6. Training performance versus FL round and communication cost for FEMNIST under $\epsilon = 2.0$.

privacy budget ϵ indicates a lower noise requirement, and hence the useful signal has a relatively higher magnitude than the channel noise, leading to a more accurate model estimation in each round. Second, the accuracy of WFL-P is the upper bound of WFL-PDP because WFL-PDP needs to consider the additional DP constraint besides the same power constraint as in WFL-P. From (37) and (38), we can see that the power control decisions of WFL-PDP and WFL-P are the same when ϵ is large. This is verified in Fig. 4, which shows that the model accuracies of WFL-PDP and WFL-P are the same when $\epsilon \geq 2.5$ for CIFAR-10 and $\epsilon \geq 9$ for FEMNIST. Third, PFELS achieves is always better than WFL-PDP. This is due to PFELS sparsifies the local model updates before transmitting them to reduce the total convergence error while respecting the power and privacy constraints. Fourth, when ϵ is relatively large (e.g., $\epsilon \geq 1.0$ for CIFAR-10 and $\epsilon \geq 2.0$ for FEMNIST), leading to a non-binding DP constraint, PFELS can outperform WFL-P via the use of sparsification. In contrast, when ϵ is small, PFELS needs to adjust the power scaling coefficient to satisfy the strict DP constraint, while WFL-P lacks a DP constraint. Consequently, WFL-P can achieve higher model accuracy compared to PFELS in this case. Note PFELS sparsifies the local model updates before transmitting them to reduce the total convergence error while respecting the power and privacy constraints. This clearly shows the benefit of sparsification in improving the privacy-accuracy trade-off in wireless FL.
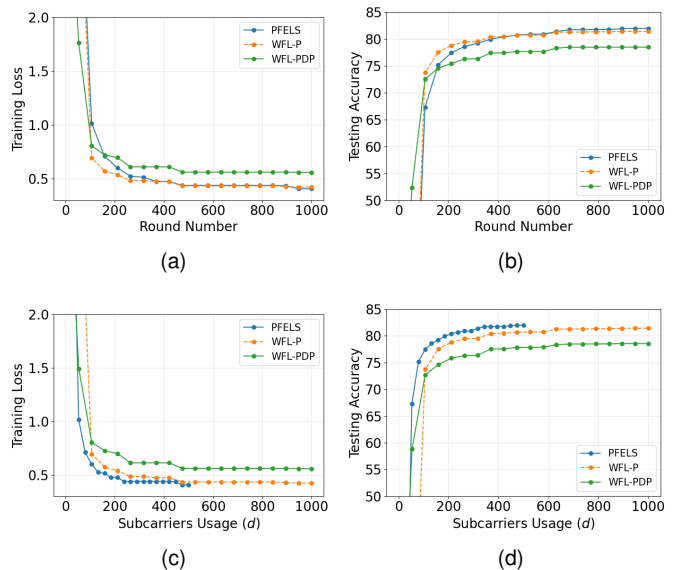
8.2.3 Communication Efficiency Benefits of PFELS

Next, we compare the convergence rate and communication cost in terms of subcarrier usage for PFELS and baselines under a fixed privacy budget. Table 2 and Table 3 summarize the performance of PFELS and baselines after T rounds on CIFAR-10 and FEMNIST, respectively. Fig. 5 and Fig. 6 show the performance during the training on CIFAR-10 and FEMNIST, respectively. First, for CIFAR-10, as shown in Fig. 5a on training loss and Fig. 5b on test accuracy, PFELS has the fastest convergence speed in terms of communication round. This is due to the fact that PFELS uses compression to balance the trade-off between the privacy error and compression error and achieves a better convergence speed. Second, WFL-P has a faster convergence speed in terms of communication round than WFL-PDP because WFL-PDP needs to consider the additional DP constraint besides the same power constraint as in WFL-P. Third, Fig. 5c and Fig. 5d show the training loss and test accuracy w.r.t. the communication cost for PFELS and baselines. The results in Table 2 show that PFELS achieves higher communication efficiency than the baselines by utilizing sparsification with only 600 subcarriers. In contrast, both WFL-P and WFL-PDP incur the same communication cost, using 2000 subcarriers, as they transmit complete model updates. Similar results can be observed for FEMNIST dataset from Fig. 6 and Table 3.

8.2.4 Energy Efficiency Benefits of PFELS

Finally, we evaluate the benefit of PFELS in improving communication energy efficiency compared with the baselines under a fixed privacy budget. Here, we compute the total of accumulated transmission energy of all participated devices as the transmission energy cost. Table 2, 3, and Fig. 7 show the results on CIFAR-10 and FEMNIST. For CIFAR-10, as shown in Fig. 7a, 7b, and Table 2, we observe that PFELS achieves the final test accuracy (76.42%) while utilizing 2.55×10^{11} J in energy consumption, whereas WFL-
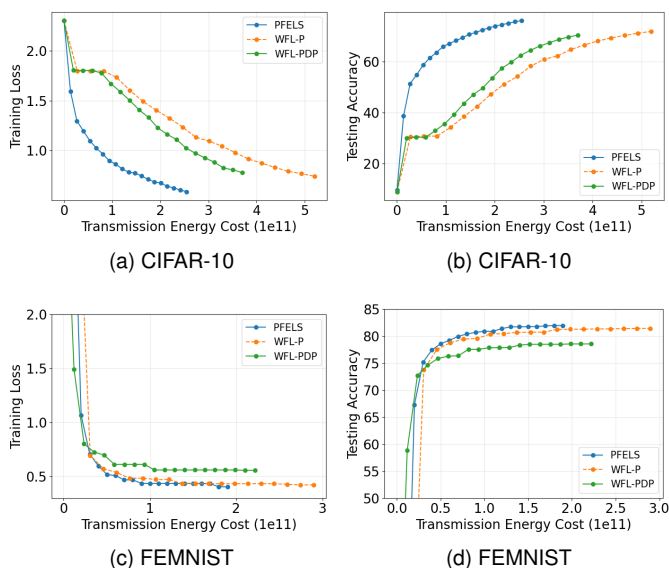
Fig. 7. Transmission energy cost comparisons for CIFAR-10 under $\epsilon = 1.5$ and FEMNIST datasets under $\epsilon = 2.0$.

P consumes 5.19×10^{11} J to achieve the final test accuracy (72.72%) and WFL-P consumes 3.70×10^{11} J to achieve the final test accuracy (72.23%) in energy consumption. For FEMNIST dataset, as depicted in Fig. 7c, 7d, and Table 3, PFELS achieves the final test accuracy (81.97%) while utilizing 1.90×10^{11} J in energy consumption, whereas WFL-P consumes 2.89×10^{11} J to achieve the final test accuracy (81.43%) and WFL-P consumes 2.22×10^{11} J to achieve the final test accuracy (78.55%) in energy consumption. In summary, PFELS outperforms all baselines by saving transmission energy cost.

9 CONCLUSION

In this paper, we have proposed PFELS, a novel wireless FL scheme aimed at achieving client-level DP while maintaining high model accuracy and improving communication and energy efficiency by harnessing the intrinsic channel noise, signal-superposition nature of the wireless channel, and update compression in AirComp. Furthermore, we have analyzed the convergence and privacy properties of PFELS under the general non-convex and non-iid setting. Experimental results have demonstrated that compared with prior wireless FL frameworks, PFELS can greatly improve the model accuracy, communication and energy efficiency simultaneously under the same DP guarantee. In the future, we will investigate other compression methods and extend our algorithm to the setting with imperfect channel status.

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. Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[2] D. Shi, L. Li, R. Chen, P. Prakash, M. Pan, and Y. Fang, "Toward energy-efficient federated learning over 5g+ mobile devices," *IEEE Wireless Commun.*, vol. 29, no. 5, pp. 44–51, 2022.

[3] Y. Guo, Y. Sun, R. Hu, and Y. Gong, "Hybrid local SGD for federated learning with heterogeneous communications," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[4] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7947–7962, 2021.

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

[6] 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.

[7] 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.

[8] B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *Proc. Int. Conf. Learn. Represent.*, 2018.

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

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

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

[12] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[13] ——, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, 2020.

[14] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, 2021.

[15] 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, 2021.

[16] A. Sonee, S. Rini, and Y. Huang, "Wireless federated learning with limited communication and differential privacy," in *Proc. IEEE Glob. Commun. Conf.*, 2021, pp. 01–06.

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

[18] 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, 2021.

[19] Y. Zhu, X. Yu, Y.-H. Tsai, F. Pittaluga, M. Faraki, M. Chandraker, and Y.-X. Wang, "Voting-based approaches for differentially private federated learning," in *Proc. NIPS Workshop Federated Learn.: Recent Adv. New Challenges*, 2022.

[20] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[21] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 308–318.

[22] R. Liu, Y. Cao, M. Yoshikawa, and H. Chen, "FedSel: Federated SGD under local differential privacy with top-k dimension selection," in *Proc. Int. Database Syst. Adv. Appl.* Springer, 2020, pp. 485–501.

[23] R. Hu, Y. Gong, and Y. Guo, "Federated learning with sparsification-amplified privacy and adaptive optimization," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, p. 1463–1469.

[24] R. Hu, Y. Guo, and Y. Gong, "Concentrated differentially private federated learning with performance analysis," *IEEE Open J. Comput. Soc.*, vol. 2, pp. 276–289, 2021.

[25] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong, "Personalized federated learning with differential privacy," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9530–9539, 2020.

[26] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5194–5208, 2021.

[27] Y. Shao, D. Gündüz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3951–3964, 2022.

[28] S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi, "Error feedback fixes SignSGD and other gradient compression schemes," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3252–3261.

[29] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[30] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2019, pp. 6155–6165.

[31] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," in *Proc. NIPS Workshop Mach. Learn. Phone Consumer Devices*, 2017.

[32] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

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

[34] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5381–5393.

[35] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7611–7623.

[36] R. Hu, Y. Gong, and Y. Guo, "Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy," *arXiv preprint arXiv:2202.07178*, 2022.

[37] A. Cheng, P. Wang, X. S. Zhang, and J. Cheng, "Differentially private federated learning with local regularization and sparsification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 122–10 131.

[38] P. Kairouz, Z. Liu, and T. Steinke, "The distributed discrete Gaussian mechanism for federated learning with secure aggregation," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 5201–5212.

[39] K. Wei, J. Li, M. Ding, C. Ma, H. Su, B. Zhang, and H. V. Poor, "User-level privacy-preserving federated learning: Analysis and performance optimization," *IEEE Trans. Mobile Comput.*, vol. 21, no. 9, pp. 3388–3401, 2021.

[40] N. Papernot, A. Thakurta, S. Song, S. Chien, and Ú. Erlingsson, "Tempered sigmoid activations for deep learning with differential privacy," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 10, 2021, pp. 9312–9321.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[43] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, 2020.

APPENDIX A
PROOF OF LEMMA 1

Proof: For the rand_k, taking the expectation on the active set ω, we have:

$$\mathbb{E}_\omega[\hat{\Delta}^t] = \mathbb{E}_\omega[\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t + \frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}] \tag{39a}$$

$$= \frac{1}{r}\sum_{i\in\mathcal{S}^t}\mathbb{E}_\omega[(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t] \tag{39b}$$

$$= \frac{1}{r}\sum_{i\in\mathcal{S}^t}[\frac{k}{d}[\Delta_i^t]_1,\ldots,\frac{k}{d}[\Delta_i^t]_d] \tag{39c}$$

$$= \frac{k}{d}\sum_{i\in\mathcal{S}^t}\frac{\Delta_i^t}{r}, \tag{39d}$$

where (39b) holds due to the zero mean of Gaussian noise. $\qquad\square$

APPENDIX B
PROOF OF LEMMA 3

Proof: According to the Assumption 2, we have:

$$\mathbb{E}_t[f(\boldsymbol{\theta}^{t+1} - f(\boldsymbol{\theta}^t)] \leq \mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\rangle + \frac{L}{2}\mathbb{E}_t\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2$$

$$= \mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \mathbb{E}_{\mathcal{S}^t}\left[\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t + \frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}\right]\rangle + \frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t + \frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}\right\|^2$$

$$\overset{(a)}{=} \mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \mathbb{E}_{\mathcal{S}^t}\left[\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t\right]\rangle + \frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t + \frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}\right\|^2$$

$$\overset{(b)}{=} \mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \frac{1}{r}r\frac{1}{N}\sum_{i=1}^N(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t\rangle + \frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t + \frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}\right\|^2$$

$$= \underbrace{\mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \frac{1}{N}\sum_{i=1}^N(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t\rangle}_{T_1} + \underbrace{\frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t + \frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}\right\|^2}_{T_2} \tag{40}$$

where (a) holds due to the zero-mean Gaussian noise, (b) follows from the unbiased sampling. For T_1, we have:

$$T_1 = \mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \frac{1}{N}\sum_{i=1}^N(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t - \frac{1}{N}\sum_{i=1}^N\Delta_i^t + \frac{1}{N}\sum_{i=1}^N\Delta_i^t\rangle$$

$$= \underbrace{\mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \frac{1}{N}\sum_{i=1}^N\Delta_i^t\rangle}_{A_1} + \underbrace{\mathbb{E}_t\langle\nabla f(\boldsymbol{\theta}^t), \frac{1}{N}\sum_{i=1}^N(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t - \frac{1}{N}\sum_{i=1}^N\Delta_i^t\rangle}_{A_2} \tag{41}$$

For A_1, we have:

$$A_1 = \mathbb{E}_t \langle \nabla f(\boldsymbol{\theta}^t), \frac{1}{N} \sum_{i=1}^N \Delta_i^t \rangle$$

$$\overset{(a)}{=} -\eta\tau \mathbb{E}_t \langle \nabla f(\boldsymbol{\theta}^t), \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \rangle$$

$$\overset{(b)}{=} -\frac{\eta\tau}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 - \frac{\eta\tau}{2} \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2 + \frac{\eta\tau}{2} \left\| \nabla f(\boldsymbol{\theta}^t) - \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2$$

$$= -\frac{\eta\tau}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta\tau}{2} \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau (\nabla f_i(\boldsymbol{\theta}^t) - \nabla f_i(\boldsymbol{\theta}_i^{t,s-1})) \right\|^2 - \frac{\eta\tau}{2} \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2$$

$$\overset{(c)}{\leq} -\frac{\eta\tau}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta}{2N} \sum_{i=1}^N \sum_{s=1}^\tau \left\| \nabla f_i(\boldsymbol{\theta}^t) - \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2 - \frac{\eta\tau}{2} \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2$$

$$\overset{(d)}{\leq} -\frac{\eta\tau}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta L^2}{2N} \sum_{i=1}^N \sum_{s=1}^\tau \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1} \right\|^2 - \frac{\eta\tau}{2} \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2, \tag{42}$$

where (a) holds due to $\boldsymbol{\theta}_i^{t,\tau} = \boldsymbol{\theta}^t - \eta \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1}$ and the unbiased stochastic gradient Assumption 3, (b) follows from $2\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a} - \boldsymbol{b}\|^2$, (c) holds due to Jensen's inequality, and (d) comes from the L-smoothness Assumption 2.

For A_2, let $\lambda_k = 1 - \frac{k}{d}$, we have:

$$A_2 \overset{(a)}{=} \eta\tau \mathbb{E}_t \langle \nabla f(\boldsymbol{\theta}^t), \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1} - \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau (\mathbf{A}^t)^\intercal \mathbf{A}^t \mathbf{g}_i^{t,s-1} \rangle$$

$$\overset{(b)}{\leq} \frac{\eta\tau}{2} \left(\pi \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \pi^{-1} \mathbb{E}_t \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1} - \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau (\mathbf{A}^t)^\intercal \mathbf{A}^t \mathbf{g}_i^{t,s-1} \right\|^2 \right)$$

$$\overset{(c)}{\leq} \frac{\eta\tau\pi}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta\tau\lambda_k}{2\pi} \mathbb{E}_t \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1} \right\|^2$$

$$= \frac{\eta\tau\pi}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta\tau\lambda_k}{2\pi} \mathbb{E}_t \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1} - \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) + \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2$$

$$\overset{(d)}{=} \frac{\eta\tau\pi}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta\tau\lambda_k}{2\pi} \mathbb{E}_t \left[\left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1} - \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2 + \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2 \right]$$

$$\overset{(e)}{\leq} \frac{\eta\tau\pi}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta\tau\lambda_k}{2N\pi} \sum_{i=1}^N \mathbb{E}_t \left\| \frac{1}{\tau} \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1} - \frac{1}{\tau} \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2 + \frac{\eta\tau\lambda_k}{2\pi} \mathbb{E}_t \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2$$

$$\overset{(f)}{\leq} \frac{\eta\tau\pi}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta\tau\lambda_k\bar\zeta^2}{2\pi} + \frac{\eta\tau\lambda_k}{2\pi} \mathbb{E}_t \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2 \tag{43}$$

where (a) holds due to $\boldsymbol{\theta}_i^{t,\tau} = \boldsymbol{\theta}^t - \eta \sum_{s=1}^\tau \mathbf{g}_i^{t,s-1}$, (b) follows from Lemma 9, (c) follows from Lemma 10, (d) uses unbiased Assumption 3, (e) uses Lemma 6, and (f) uses (55) and $\bar\zeta^2 = (1/N) \sum_{i=1}^N \zeta_i^2$.

Combining (41), (42) and (43), if $\lambda_k \leq \pi$, we get:

$$T_1 \leq -\frac{\eta\tau(1-\pi)}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta L^2}{2N} \sum_{i=1}^N \sum_{s=1}^\tau \mathbb{E}_t \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1} \right\|^2 + \frac{\eta\tau\lambda_k\bar\zeta^2}{2\pi} - \frac{\eta\tau(1-\lambda_k/\pi)}{2} \mathbb{E}_t \left\| \frac{1}{N\tau} \sum_{i=1}^N \sum_{s=1}^\tau \nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) \right\|^2$$

$$\leq -\frac{\eta\tau(1-\pi)}{2} \|\nabla f(\boldsymbol{\theta}^t)\|^2 + \frac{\eta L^2}{2N} \sum_{i=1}^N \sum_{s=1}^\tau \mathbb{E}_t \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1} \right\|^2 + \frac{\eta\tau\lambda_k\bar\zeta^2}{2\pi}. \tag{44}$$

For T_2, using lemma because the expectation of Gaussian noise is zero, we have:

$$
\begin{aligned}
T_2 &= \frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t + \frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}\right\|^2 \\
&\overset{(a)}{=} \frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}_t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t\right\|^2 + \frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}^t}\frac{(\mathbf{A}^t)^\intercal\mathbf{z}^t}{r\beta^t}\right\|^2 \\
&= \frac{L}{2}\mathbb{E}_t\left\|\frac{1}{r}\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t - \frac{1}{r}\sum_{i\in\mathcal{S}^t}\Delta_i^t + \frac{1}{r}\sum_{i\in\mathcal{S}^t}\Delta_i^t\right\|^2 + \frac{L}{2r^2}r\frac{\sigma_0^2 k}{(r\beta^t)^2} \\
&\overset{(b)}{\leq} \frac{L}{r^2}\mathbb{E}_t\left\|\sum_{i\in\mathcal{S}^t}(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t - \Delta_i^t\right\|^2 + \frac{L}{r^2}\mathbb{E}_t\left\|\sum_{i\in\mathcal{S}^t}\Delta_i^t\right\|^2 + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2} \\
&\overset{(c)}{\leq} \frac{L}{r^2}r\mathbb{E}_t\left[\sum_{i\in\mathcal{S}^t}\left\|(\mathbf{A}^t)^\intercal\mathbf{A}^t\Delta_i^t - \Delta_i^t\right\|^2\right] + \frac{L}{r^2}r\mathbb{E}_t\left[\sum_{i\in\mathcal{S}^t}\left\|\Delta_i^t\right\|^2\right] + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2} \\
&\overset{(d)}{=} \frac{L}{r}\lambda_k\mathbb{E}_t\left[\sum_{i\in\mathcal{S}^t}\left\|\Delta_i^t\right\|^2\right] + \frac{L}{r}\mathbb{E}_t\left[\sum_{i\in\mathcal{S}^t}\left\|\Delta_i^t\right\|^2\right] + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2} \\
&= L(\lambda_k+1)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t\left\|\Delta_i^t\right\|^2 + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2} \\
&\overset{(e)}{=} \frac{L\eta^2\tau^2}{N}(\lambda_k+1)\sum_{i=1}^{N}\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1}\right\|^2 + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2},
\end{aligned}
$$

where (a) holds due to the fact that $\mathbf{z}^t$ is zero-mean, (b) and (c) follow from Lemma 7, (d) follows from Lemma 10, and (e) holds due to $\boldsymbol{\theta}_i^{t,\tau} = \boldsymbol{\theta}^t - \eta\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1}$.

According to Lemma 11, we have:

$$
\begin{aligned}
T_2 \leq{} &2L\eta^2\tau^2\gamma^2(\lambda_k+1)\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 + \frac{2L^3\eta^2\tau}{N}(\lambda_k+1)\sum_{i=1}^{N}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1}\right\|^2 \\
&+ L\eta^2\tau^2(\lambda_k+1)(\bar{\zeta}^2 + 2\kappa^2) + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2}
\end{aligned}
\tag{45}
$$

Combining (40), (44) and (45), we get:

$$
\begin{aligned}
\mathbb{E}_t[f(\boldsymbol{\theta}^{t+1}) - f(\boldsymbol{\theta}^t)] \leq{} &\left(-\frac{\eta\tau(1-\pi)}{2} + 2L\eta^2\tau^2\gamma^2(\lambda_k+1)\right)\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 \\
&+ \left(\frac{\eta L^2}{2N} + \frac{2L^3\eta^2\tau^2(\lambda_k+1)}{n\tau}\right)\sum_{i=1}^{N}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1}\right\|^2 + \frac{\eta\tau\lambda_k\bar{\zeta}^2}{2\pi} \\
&+ L\eta^2\tau^2(\lambda_k+1)(\bar{\zeta}^2 + 2\kappa^2) + \frac{Lk\sigma_0^2}{2r^3(\beta^t)^2}.
\end{aligned}
$$

$\square$

APPENDIX C
PROOF OF LEMMA 4

Proof: According to the local update rule, we have

$$
\mathbb{E}_t \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1} \right\|^2 = \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t - \eta \mathbf{g}_i^{t,s-2} \right\|^2
$$

$$
= \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t - \eta \mathbf{g}_i^{t,s-2} + \eta \nabla f_i(\boldsymbol{\theta}_i^{t,s-2}) - \eta \nabla f_i(\boldsymbol{\theta}_i^{t,s-2}) + \eta \nabla f_i(\boldsymbol{\theta}^t) - \eta \nabla f_i(\boldsymbol{\theta}^t) \right\|^2
$$

$$
\overset{(a)}{=} \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t - \eta \nabla f_i(\boldsymbol{\theta}_i^{t,s-2}) + \eta \nabla f_i(\boldsymbol{\theta}^t) - \eta \nabla f_i(\boldsymbol{\theta}^t) \right\|^2 + \eta^2 \mathbb{E}_t \left\| \mathbf{g}_i^{t,s-2} - \nabla f_i(\boldsymbol{\theta}_i^{t,s-2}) \right\|^2
$$

$$
\overset{(b)}{\leq} (1 + \frac{1}{2\tau - 1}) \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t \right\|^2 + 2\eta^2 \tau \mathbb{E}_t \left\| \nabla f_i(\boldsymbol{\theta}_i^{t,s-2}) - \nabla f_i(\boldsymbol{\theta}^t) + \nabla f_i(\boldsymbol{\theta}^t) \right\|^2 + \eta^2 \zeta_i^2
$$

$$
\overset{(c)}{\leq} (1 + \frac{1}{2\tau - 1}) \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t \right\|^2 + 4\eta^2 \tau \mathbb{E}_t \left\| \nabla f_i(\boldsymbol{\theta}_i^{t,s-2}) - \nabla f_i(\boldsymbol{\theta}^t) \right\|^2 + 4\eta^2 \tau \mathbb{E}_t \left\| \nabla f_i(\boldsymbol{\theta}^t) \right\|^2 + \eta^2 \zeta_i^2
$$

$$
\overset{(d)}{\leq} (1 + \frac{1}{2\tau - 1}) \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t \right\|^2 + 4\eta^2 L^2 \tau \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t \right\|^2 + 4\eta^2 \tau \mathbb{E}_t \left\| \nabla f_i(\boldsymbol{\theta}^t) \right\|^2 + \eta^2 \zeta_i^2,
$$

where (a) holds due to the unbiased gradient estimation in Assumption 3, (b) follows from Assumption 3 and Lemma 8 with $\alpha = \frac{1}{2\tau - 1}$, (c) follows from Lemma 7, and (d) follows from the smoothness assumption 2. Next, taking the average of sum of all clients, we get:

$$
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_t \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1} \right\|^2 \leq (1 + \frac{1}{2\tau - 1} + 4\eta^2 L^2 \tau) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t \right\|^2 + \frac{4\eta^2 \tau}{N} \sum_{i=1}^{N} \mathbb{E}_t \left\| \nabla f_i(\boldsymbol{\theta}^t) \right\|^2 + \frac{\eta^2}{N} \sum_{i=1}^{N} \zeta_i^2
$$

$$
\overset{(a)}{\leq} (1 + \frac{1}{2\tau - 1} + 4\eta^2 L^2 \tau) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t \right\|^2 + 4\eta^2 \tau \gamma^2 \mathbb{E}_t \left\| \nabla f(\boldsymbol{\theta}^t) \right\|^2 + 4\eta^2 \tau \kappa^2 + \eta^2 \bar{\zeta}^2,
$$

where (a) holds due to the dissimilarity Assumption 5 and the notation $\bar{\zeta}^2 = (1/N) \sum_{i=1}^{N} \zeta_i^2$. When $\eta \leq \frac{1}{2\sqrt{2}\tau L}$, then $4\eta^2 L^2 \tau \leq \frac{1}{2\tau}$. We have:

$$
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_t \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1} \right\|^2 \leq (1 + \frac{1}{\tau - 1}) \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_t \left\| \boldsymbol{\theta}_i^{t,s-2} - \boldsymbol{\theta}^t \right\|^2 + 4\eta^2 \tau \gamma^2 \mathbb{E}_t \left\| \nabla f(\boldsymbol{\theta}^t) \right\|^2 + 4\eta^2 \tau \kappa^2 + \eta^2 \bar{\zeta}^2
$$

Unrolling the recursion, we get:

$$
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_t \left\| \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1} \right\|^2 \leq \sum_{h=0}^{s-1} (1 + \frac{1}{\tau - 1})^h \left[4\eta^2 \tau \gamma^2 \mathbb{E}_t \left\| \nabla f(\boldsymbol{\theta}^t) \right\|^2 + 4\eta^2 \tau \kappa^2 + \eta^2 \bar{\zeta}^2 \right]
$$

$$
\leq (\tau - 1) \left[(1 + \frac{1}{\tau - 1})^\tau - 1 \right] \times \left[4\eta^2 \tau \gamma^2 \mathbb{E}_t \left\| \nabla f(\boldsymbol{\theta}^t) \right\|^2 + 4\eta^2 \tau \kappa^2 + \eta^2 \bar{\zeta}^2 \right]
$$

$$
\overset{(a)}{\leq} 16 \eta_l^2 \tau^2 \gamma^2 \left\| \nabla f(\boldsymbol{\theta}^t) \right\|^2 + 16 \eta^2 \tau^2 \kappa^2 + 4\tau \eta^2 \bar{\zeta}^2,
$$

where (a) follows from $(1 + \frac{1}{\tau - 1})^\tau \leq 5$ when $\tau > 1$. $\qquad \square$

APPENDIX D
USEFUL INEQUALITIES

For ease of notation, we use $\| \cdot \|$ to denote the ℓ_2 vector norm.

Lemma 6 (Jensen's inequality). *For arbitrary set of N vectors $\{\boldsymbol{a}_i\}_i^N$, $\boldsymbol{a}_i \in \mathbb{R}^d$ and positive weights $\{w_i\}_{i \in [N]}$, $\sum_i^N w_i = 1$,*

$$
\left\| \sum_{i=1}^{N} w_i \boldsymbol{a}_i \right\|^2 \leq \sum_{i=1}^{N} w_i \| \boldsymbol{a}_i \|^2 \tag{46}
$$

Lemma 7 (Cauchy-Schwarz inequality). *For arbitrary set of N vectors $\{\boldsymbol{a}_i\}_{i=1}^N$, $\boldsymbol{a}_i \in \mathbb{R}^d$,*

$$
\left\| \sum_{i=1}^{N} \boldsymbol{a}_i \right\|^2 \leq N \sum_{i=1}^{N} \| \boldsymbol{a}_i \|^2 \tag{47}
$$

Lemma 8. *For given two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$,*

$$
\| \boldsymbol{a} + \boldsymbol{b} \|^2 \leq (1 + \alpha) \| \boldsymbol{a} \|^2 + (1 + \alpha^{-1}) \| \boldsymbol{b} \|^2, \forall \alpha \geq 0. \tag{48}
$$

Lemma 9. *For given two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$,*

$$
2 \langle \boldsymbol{a}, \boldsymbol{b} \rangle \leq \pi \| \boldsymbol{a} \|^2 + \pi^{-1} \| \boldsymbol{b} \|^2, \forall \pi > 0. \tag{49}
$$

APPENDIX E
INTERMEDIATE RESULTS

Lemma 10 (Bounded Sparsification). *Given a vector $\mathbf{a} \in \mathbb{R}^d$ and a parameter $k \in [d]$, The rand_k projection matrix $\mathbf{A} \in \mathbb{R}^{k \times d}$ generated from the activate subset ω holds that*

$$\mathbb{E}_\omega[\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{a}] = \frac{k}{d}\mathbf{a}, \quad \mathbb{E}_\omega\|\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{a} - \mathbf{a}\|^2 = (1 - \frac{k}{d})\|\mathbf{a}\|^2. \tag{50}$$

Proof: For the rand_k, taking the expectation on the active set ω, we have:

$$\mathbb{E}_\omega[\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{a}] = [\frac{k}{d}[\mathbf{a}]_1, \dots, \frac{k}{d}[\mathbf{a}]_d] = \frac{k}{d}\mathbf{a} \tag{51}$$

For the variance of rand_k, we get:

$$\mathbb{E}_\omega\|\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{a} - \mathbf{a}\|^2 = \sum_{n=1}^{d}\left(\frac{k}{d}([\mathbf{a}]_n - [\mathbf{a}]_n)^2 + (1 - \frac{k}{d})[\mathbf{a}]_n^2\right) = (1 - \frac{k}{d})\|\mathbf{a}\|^2 \tag{52}$$

$\square$

Lemma 11 (Bounded Local Model Update). *Under Assumptions 2, 3 and 5, the local model update Δ_i^t at round t is bounded by:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1}\right\|^2 \leq \frac{2L^2}{N\tau}\sum_{i=1}^{N}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1}\right\|^2 + 2(\gamma^2\|\nabla f(\boldsymbol{\theta}^t)\|^2 + \kappa^2) + \bar{\zeta}^2 \tag{53}$$

where $\bar{\zeta}^2 = (1/N)\sum_{i=1}^{N}\zeta_i^2$.

Proof:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1}\right\|^2 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1} - \frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) + \frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right\|^2$$

$$\overset{(a)}{=} \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1} - \frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right\|^2 + \mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right\|^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1} - \frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right\|^2 + \mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) - \nabla f_i(\boldsymbol{\theta}^t) + \nabla f_i(\boldsymbol{\theta}^t)\right\|^2\right]$$

$$\overset{(b)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1} - \frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right\|^2 + 2\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) - \nabla f_i(\boldsymbol{\theta}^t)\right\|^2\right]$$

$$+ 2\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t\|\nabla f_i(\boldsymbol{\theta}^t)\|^2$$

$$\overset{(c)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_t\left\|\frac{1}{\tau}\left(\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1} - \sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right)\right\|^2 + 2\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) - \nabla f_i(\boldsymbol{\theta}^t)\right\|^2\right]$$

$$+ 2(\gamma^2\|\nabla f(\boldsymbol{\theta}^t)\|^2 + \kappa^2) \tag{54}$$

where (a) follows from the unbiased stochastic gradient Assumption 3, (b) uses Lemma 7, and (c) holds due to Assumption 5.

$$\mathbb{E}_t\left\|\frac{1}{\tau}\left(\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1} - \sum_{s=1}^{\tau}\nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right)\right\|^2 \leq \frac{1}{\tau}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\mathbf{g}_i^{t,s-1} - \nabla f_i(\boldsymbol{\theta}_i^{t,s-1})\right\|^2 \leq \frac{1}{\tau}\sum_{s=1}^{\tau}\zeta_i^2 = \zeta_i^2, \tag{55}$$

where (55) follows from Lemma 6 and Assumption 3.

$$\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\left(\nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) - \nabla f_i(\boldsymbol{\theta}^t)\right)\right\|^2 \leq \frac{1}{\tau}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\nabla f_i(\boldsymbol{\theta}_i^{t,s-1}) - \nabla f_i(\boldsymbol{\theta}^t)\right\|^2 \leq \frac{L^2}{\tau}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\boldsymbol{\theta}_i^{t,s-1} - \boldsymbol{\theta}^t\right\|^2, \tag{56}$$

where (56) follows from Lemma 6 and Assumption 2. Combining (54), (55) and (56), we have:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t\left\|\frac{1}{\tau}\sum_{s=1}^{\tau}\mathbf{g}_i^{t,s-1}\right\|^2 \leq \frac{2L^2}{N\tau}\sum_{i=1}^{N}\sum_{s=1}^{\tau}\mathbb{E}_t\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,s-1}\right\|^2 + 2(\gamma^2\|\nabla f(\boldsymbol{\theta}^t)\|^2 + \kappa^2) + \bar{\zeta}^2 \tag{57}$$

$\square$

APPENDIX F
PROOF OF LEMMA 5

Proof: Taking the expectation on the rand_k, we have:

$$\mathbb{E}\|\mathbf{A}^t\Delta_i^t\|_2^2 = \sum_{j=1}^{d} \frac{k}{d}[\Delta_i^t]_j^2 \tag{58a}$$

$$= \frac{k}{d}\|\boldsymbol{\theta}_i^{t,\tau} - \boldsymbol{\theta}^t\|_2^2 \tag{58b}$$

$$= \frac{k}{d}\eta^2\left\|\sum_{s=1}^{\tau} \mathbf{g}_i^{t,s-1}\right\|_2^2 \tag{58c}$$

$$\leq \frac{k}{d}\eta^2\tau^2 C_1^2, \tag{58d}$$

where (58c) holds due to (18), (58d) follows from Assumption 1 and Cauchy–Schwarz inequality. $\qquad\square$