FLORAS: Differentially Private Wireless Federated Learning Using Orthogonal Sequences

Xizixiang Wei*, Tianhao Wang§, Ruiquan Huang†, Cong Shen*, Jing Yang†, H. Vincent Poor‡
*Department of Electrical and Computer Engineering, University of Virginia, USA
§ Department of Computer Science, University of Virginia, USA
† Department of Electrical Engineering, The Pennsylvania State University, USA
‡ Department of Electrical and Computer Engineering, Princeton University, USA

***Abstract*—We propose a novel private-preserving uplink over-the-air computation (AirComp) method, termed `FLORAS`, for wireless federated learning (FL) systems. From the communication design perspective, `FLORAS` eliminates the requirement of channel state information at the transmitters (CSIT) by leveraging the properties of orthogonal sequences. From the privacy perspective, we prove that `FLORAS` can offer *pure* differential privacy (DP) guarantee, and explicitly characterize the achievable ϵ-DP level as a function of the `FLORAS` parameter configuration. A novel FL convergence bound is derived which, combined with the pure DP guarantee, allows for a smooth tradeoff between convergence rate and DP guarantee levels. Experiments based on real-world datasets not only corroborate the theoretical findings but also empirically demonstrate the communication and privacy advantages of `FLORAS` over state-of-the-art AirComp methods.**

***Index Terms*—Federated learning; Differential privacy (DP); Orthogonal sequence; Code-division multiple access (CDMA).**

I. Introduction

Uplink communication is known as one of the main bottlenecks of wireless federated learning (FL) [1]. To tackle the scalability problem, over-the-air computation (also known as *AirComp*) mechanisms have been proposed (see [2] and the references therein). Instead of decoding individual local models of clients and then aggregating, AirComp allows multiple clients to transmit uplink signals in a superpositioned fashion, and decodes the average global model directly. The most popular and heuristic AirComp method is based on channel inversion power control [3], which "inverts" the fading channel at each transmitter, so that the aggregated model can be directly obtained at the server. Further enhancements have been proposed along this direction; yet a fundamental limitation of the existing methods is that they mostly require channel state information at the transmitters (CSIT). Enabling CSIT in communication systems is complicated and the precision of CSIT is often worse than the channel state information at the receiver (CSIR). Moreover, channel inversion is well known to "blow up" when one of the users' channels is in deep fade. Hence, exploring CSIT-free AirComp methods becomes attractive [4].

WX and CS were partially supported by the National Science Foundation (NSF) under grants ECCS-2033671 and ECCS-2143559. TW was partially supported by NSF under grant CNS-2220433. RH and JY were partially supported by NSF under grants CNS-1956276, ECCS-2030026, and CNS-2114542. HVP was partially supported by NSF under grants CCF-1908308 and CNS-2128448.

Meanwhile, with the ever-growing importance on data security, privacy preservation of personal information has been increasingly valued by companies and governments. Although FL can intuitively preserve client privacy by keeping training data locally, private information can still be leaked to some extent by analyzing the differences of models trained and uploaded by the clients [5], [6]. To address the privacy concern, a natural approach is to add (artificial) noise to the model parameters in the upload phase of FL, which can be mathematically characterized through the lenses of *differential privacy (DP)* [7]. AirComp has the potential to achieve DP at no extra cost, due to the natural noise in the wireless channel. Different DP levels can be guaranteed "for free" by controlling the effective channel noise. However, AirComp literature rarely characterizes the achievable privacy protection in a mathematically rigorous fashion. Wei et al. [8] proposes an AirComp design to achieve DP by adjusting the effective noise. Seif et al. [9] maximizes the convergence rate while satisfying a desired privacy level by optimizing the power allocation between local gradients and the artificial noise. We note that most research on the DP of AirComp requires power adjustment to achieve the desired DP level, which is often complex and power inefficient. Moreover, as wireless channel noise is usually modeled as a Gaussian distribution, most existing research (e.g. [8], [9]) can only guarantee (ϵ, δ)-DP, which is weaker than ϵ-DP ("pure DP").

To simultaneously remove the CSIT requirement of AirComp and address the privacy challenge, we propose `FLORAS` – Federated Learning using ORthogonAl Sequences, a novel uplink wireless physical layer design for FL by leveraging the properties of *orthogonal sequences*. On the communication design, `FLORAS` preserves all the advantages of AirComp while removing the CSIT requirement. In particular, orthogonal sequences enable the BS to obtain the CSIR via a single pilot, by which global parameters can be estimated through simple linear projections. Therefore, `FLORAS` significantly reduces the channel estimation overhead while allowing the transmit power to be independent of the channel realizations, which avoids increasing the dynamic range of the transmit signal and improves the power efficiency. From the perspective of DP, `FLORAS` achieves a desired ϵ-DP guarantee by adjusting the number of used orthogonal sequences, making it much simpler and independent of the transmit power. Moreover,

978-1-5386-7462-8/23/$31.00 ©2023 IEEE

FLORAS produces an effective noise that follows the *Cauchy distribution*, as opposed to the commonly studied Gaussian distribution, on the global model, which enables *pure DP*. A new FL convergence bound based on the Cauchy noise is derived. Putting the two results together allows us to characterize the trade-off between the model convergence rate and the achievable DP levels. Experiments on real-word datasets validate our theoretical analysis.

II. System Model

A. FL Model

Consider an FL task with a central server and M total clients. Each client $k \in [M]$ stores a (disjoint) local dataset $\mathcal{D}_k$, with its size denoted by D_k. The size of the total data is $D \triangleq \sum_{k\in[M]} D_k$. We use $f_k(\mathbf{w})$ to denote the local loss function at client k, which measures how well a machine learning (ML) model with parameter $\mathbf{w} \in \mathbb{R}^d$ fits its local dataset. Therefore, the global objective function over all M clients can be denoted as $f(\mathbf{w}) = \sum_{k\in[M]} p_k f_k(\mathbf{w})$, where $p_k = \frac{D_k}{D}$ is the weight of each local loss function, and the purpose of FL is to distributively find the optimal model parameter $\mathbf{w}^*$ that minimizes the global loss function: $\mathbf{w}^* \triangleq \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$. We define $\Gamma \triangleq f^* - \sum_{k\in[M]} p_k f_k^*$ to capture the non-independent and identically distributed (non-i.i.d.) degree of local datasets, where f^* and f_k^* are the minima of global and local loss functions, respectively.

The FedAvg framework [1] keeps clients' data locally, and the global model is averaged at the server by the composition of multiple learning rounds. One of the key characteristics of FL is *partial clients participation*, i.e., only a portion of clients are selected in a single learning round for model upload. Here, we assume that K of total M clients are uniformly randomly selected during each learning round for the FL task. To simplify the notation, we use the subscript $k = 1, \cdots, K$ to indicate the K clients during a learning round, acknowledging that they could correspond to different clients in different rounds. A typical wireless FL pipeline executes the following steps iteratively, $\forall t = 1, \cdots, T$:

1) **Downlink communication.** The BS broadcasts the current global model $\mathbf{w}_t$ to all K selected devices over the downlink wireless channel.
2) **Local computation.** Each client uses its local data to train a local model improved upon the received global model $\mathbf{w}_t$. We assume that mini-batch stochastic gradient descent (SGD) is adopted to minimize the local loss function. The parameter is updated iteratively (for E steps) at client k as: $\mathbf{w}_{t,0}^k = \mathbf{w}_t; \mathbf{w}_{t,\tau}^k = \mathbf{w}_{t,\tau-1}^k - \eta_t \nabla \tilde{f}_k(\mathbf{w}_{t,\tau-1}^k, \xi_{t,\tau-1}^k),\ \forall \tau = 1, \cdots, E; \mathbf{w}_{t+1}^k = \mathbf{w}_{t,E}^k$, where $\nabla \tilde{f}_k(\mathbf{w}, \xi)$ denotes the stochastic gradient at client k on model $\mathbf{w}$, using mini-batch ξ.
3) **Uplink communication.** Each involved client uploads its latest local model to the server synchronously over the uplink wireless channel.
4) **Server aggregation.** The BS aggregates the received noisy local models $\tilde{\mathbf{w}}_{t+1}^k$ to generate a new global model. For simplicity, we assume that each local dataset has equal size. Therefore, we have $\mathbf{w}_{t+1} = \Sigma_{k=1}^K \frac{1}{K} \tilde{\mathbf{w}}_{t+1}^k$.

This work focuses on steps 3 and 4 in the FL pipeline. In particular, we leverage the unique properties of orthogonal sequences, which leads to an efficient FL uplink communication design with DP guarantees.

B. Communication Model

Consider a cell with a single-antenna BS and K single-antenna users involving in the FL task. The communication system leverages orthogonal sequences for uplink transmissions. Note that one of the most popular implementations of an orthogonal sequence-based wireless communication system is code-division multiple access (CDMA). We assume a spreading sequence set $\mathcal{A} = \{\mathbf{a}_k, k = 1, \cdots, N\}$ containing N unique spreading sequences ($N \geq K$), where each spreading sequence is denoted as $\mathbf{a}_k = [a_{1,k}, \cdots, a_{L,k}]^T$ and L is the length of each spreading sequence. Each user is (randomly) assigned with a unique spreading sequence $\mathbf{a}_k$ from $\mathcal{A}$ as its signature. We assume that the BS only has the knowledge of the entire spreading sequence set $\mathcal{A}$, *without knowing the specific signature of each user*. We emphasize that this restriction is consistent with our goal of guaranteeing user privacy – BS cannot identify users based on their spreading sequences. More details on this spreading sequence assignment mechanism can be found in the journal version of this paper.

At the uplink step of the t-th round, each client transmits the differential between the received global model and the computed new local model: $\mathbf{x}_t^k = \mathbf{w}_t - \mathbf{w}_{t+1}^k \in \mathbb{R}^d,\ \ \forall k = 1, \cdots, K$, to the BS, where $\mathbf{x}_t^k \triangleq [x_{1,t}^k, \cdots, x_{d,t}^k]^T$. Using a standard normalization technique (see e.g. Appendix of [3]), we can ensure $\mathbb{E}[|x_{i,t}^k|^2] = 1$. Moreover, we can guarantee $|x_{i,t}^k| \leq C$ ($C \gg 1$) by adopting a proper *clipping* technique [10]. To simplify the notation, we omit index t and use x_k^i instead of $x_{i,t}^k$ barring any confusion. We assume that each client transmits every element of the differential model $\{x_k^i\}_{i=1}^d$ via d shared time slots. In addition, block fading is assumed, i.e., the fading channel between each client and the BS h_k remains unchanged within d time slots. We emphasize that we do not make any specific assumption on the fading distribution throughout this paper. In the i-th slot, each client transmits symbol x_k^i spread by its uniquely assigned orthogonal sequence. The received signal at the BS can be written as

$$\mathbf{y}_i = \sum_{k=1}^{K} \mathbf{a}_k h_k x_k^i + \mathbf{n}_i \ \ \forall i = 1, \cdots, d,$$

where $\mathbf{n}_i$ is the additive white Gaussian noise (AWGN) with mean zero and variance σ^2/L per dimension. Note that since the model differential parameters are real signals, we only need to consider the real part of channel coefficients and noise here. Although one-dimensional (real) modulation cannot fully leverage the channel degrees of freedom, it is consistent with the fact that binary phase-shift keying (BPSK) is the most common modulation scheme in CDMA systems [11].

We note that each pair of different spreading sequences are orthogonal, i.e.,

$$\mathbf{a}_i^T \mathbf{a}_i = 1, \ \ \forall i \in [N]; \quad \text{and} \quad \mathbf{a}_i^T \mathbf{a}_j = 0, \ \ \forall i \neq j. \quad (1)$$

At the BS, the receiver will decode the estimated aggregation parameter $\tilde{x}_i \triangleq \sum_{k\in[K]} \hat{x}_k^i$ where $\hat{x}_k^i$ represents a noisy version of x_k^i, and recover $\tilde{\mathbf{x}}_t \triangleq [\tilde{x}_1, \cdots, \tilde{x}_d]^T$ in d slots. After that, the BS can compute the new global model as

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{K}\tilde{\mathbf{x}}_t. \quad (2)$$

Throughout the paper, we assume that all users are synchronized in frames, which can be achieved by the BS sending a beacon signal to initialize uplink transmissions.

III. FLORAS

FLORAS is a novel design for the uplink communication phase in FL. By incorporating the unique characteristics of orthogonal sequences, FLORAS enables the BS to directly obtain the estimates of aggregated parameters, which not only reduces receiver complexity but also achieves better privacy protection.

A. Algorithm design

FLORAS is a four-step framework and each step is detailed as follows.

Step 1: Uplink channel estimation. Before the model differential transmission, the BS schedules all users to transmit a common pilot s simultaneously. The received signal is:

$$\mathbf{y}_s = \sum_{k=1}^{K} \mathbf{a}_k h_k s + \mathbf{n}_s.$$

The BS can utilize $\mathbf{y}_s$ to estimate the channel gain coefficients. For the K spreading sequences adopted by the users[1], we have

$$\hat{h}_k = \frac{\mathbf{a}_k^T \mathbf{y}_s}{s} = \frac{\mathbf{a}_k^T \left[\sum_{k=1}^{K} \mathbf{a}_k h_k s + \mathbf{n}_s\right]}{s} = h_k + \frac{\mathbf{a}_k^T \mathbf{n}_s}{s}, \ \ \forall k = 1, \cdots, K.$$

For the last $N - K$ spreading sequences, the BS obtains

$$\hat{h}_k = \frac{\mathbf{a}_k^T \mathbf{y}_s}{s} = \frac{\mathbf{a}_k^T \left[\sum_{k=1}^{K} \mathbf{a}_k h_k s + \mathbf{n}_s\right]}{s} = \frac{\mathbf{a}_k^T \mathbf{n}_s}{s}, \ \ \forall k = K+1, \cdots, N.$$

Step 2: Projection vector construction. For simplicity, we assume $s = 1$ in the following. After the channel estimation, the BS constructs the following vector based on *all of* the estimated channel coefficients:

$$\mathbf{v} = \sum_{k=1}^{N} \frac{1}{\hat{h}_k} \mathbf{a}_k = \sum_{k=1}^{K} \frac{\mathbf{a}_k}{h_k + \mathbf{a}_k^T \mathbf{n}_s} + \sum_{k=K+1}^{N} \frac{\mathbf{a}_k}{\mathbf{a}_k^T \mathbf{n}_s}.$$

[1]Without loss of generality (w.l.o.g.), we assume the first K spreading sequences from the total set are selected.

We note that since the BS does not have the knowledge on which K of the total N spreading sequences are adopted by the users, it has to use all $\{\hat{h}_k, k = 1, \cdots, N\}$ to construct the projection vector $\mathbf{h}_s$. This seemingly redundant design actually enables better privacy protection, which will be discussed in Section IV.

Step 3: UL model transmission. All users upload each element of the model differentials via d shared time slots:

$$\mathbf{y}_i = \sum_{k=1}^{K} \mathbf{a}_k h_k x_k^i + \mathbf{n}_i \ \ \forall i = 1, \cdots, d.$$

The BS can then use $\mathbf{y}_i$ to estimate the aggregated parameters with the following step.

Step 4: Model decoding. The BS applies the following linear projection to estimate the aggregated model differentials:

$$\begin{aligned}\tilde{x}_i = \mathbf{v}^T \mathbf{y}_i &= \sum_{k=1}^{N} \frac{1}{\hat{h}_k} \mathbf{a}_k \left[\sum_{k=1}^{K} \mathbf{a}_k h_k x_k^i + \mathbf{n}_i\right] \\ &= \left[\sum_{k=1}^{K} \frac{\mathbf{a}_k^T}{h_k + \mathbf{a}_k^T \mathbf{n}_s} + \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T}{\mathbf{a}_k^T \mathbf{n}_s}\right] \left[\sum_{k=1}^{K} \mathbf{a}_k h_k x_k^i + \mathbf{n}_i\right] \\ &= \sum_{k=1}^{K} \frac{h_k}{h_k + \mathbf{a}_k^T \mathbf{n}_s} x_k^i + \sum_{k=1}^{K} \frac{\mathbf{a}_k^T \mathbf{n}_i}{h_k + \mathbf{a}_k^T \mathbf{n}_s} + \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s}, \\ &\forall i = 1, \cdots, d.\end{aligned}$$

After decoding $\{\tilde{x}_i, i = 1, \cdots, d\}$, the BS can compute the new global model following (2) and start the next learning round.

B. Preliminary analysis

In the high signal-to-noise ratio (SNR) regime, where the channel fading effect dominates the noise, we have $\mathbb{E}[\|\mathbf{a}_k^T \mathbf{n}_s\|^2] \ll \mathbb{E}[\|h_k\|^2]$. Therefore, we can establish the following approximation for the estimated model in Step 4:

$$\tilde{x}_i \approx \sum_{k=1}^{K} x_k^i + \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s}, \ \ \forall i = 1, \cdots, d, \quad (3)$$

where $\sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s}$ denotes the dominant noise of the received global model parameters[2]. The distribution of this post-signal-processing noise is not straightforward, and we next present Lemma 1 to establish that the noise term is a *Cauchy random variable*.

Lemma 1. *For i.i.d. Gaussian random vectors* $\mathbf{n}_i, \mathbf{n}_s \sim \mathcal{N}(0, \frac{\sigma^2}{L}\mathbf{I})$*, the derived random variable*

$$X \triangleq \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T \mathbf{n}_i}{\mathbf{a}_k^T \mathbf{n}_s} \sim \textit{Cauchy}(0, N-K), \ \ \forall \mathbf{a}_k \in \mathcal{A},$$

[2]Note that this approximation drops the minor noise term, which results in that the DP guarantee in the later discussion is the lower bound of the actual DP level of FLORAS.

with the Cauchy probability density function (p.d.f.)

$$f_X(x) = \frac{1}{\pi} \frac{N-K}{x^2 + (N-K)^2}, \quad x \in \mathbb{R}.$$

Due to the space limitation, the proof is omitted and is available in the journal version. Cauchy distribution is known as a "fat-tail" distribution, as the tail of its p.d.f. decreases proportionally with $1/x^2$. Lemma 1 suggests that for a fixed K, a larger spreading sequence set will result in a heavier tail in the Cauchy noise. Therefore, we can adjust the size of the spreading sequence set to induce different additive Cauchy noise in (3). We will discuss the effect of different Cauchy noise on DP and convergence in Section IV.

Remark 1. *Compared with the widely investigated channel inversion-based AirComp design, the proposed method does not require CSIT for the uplink communication, which greatly reduces the communication overhead. This is especially attractive for Internet-of-Things (IoT) applications with massive communication links. Moreover, thanks to the orthogonality of spreading sequences,* `FLORAS` *allows the average transmit power to be independent of channel realizations, and thus avoids increasing the dynamic range of the transmit signal. This property can significantly improve the efficiency of power amplifiers (PAs). Note that to further improve the spectrum efficiency of the proposed framework, the system can adopt random orthogonal spreading sequences. More details on the extension to non-orthogonal multiple access (NOMA) systems can be found in the journal version.*

IV. Differential Privacy and Convergence Analysis

A. Preliminaries

We investigate the DP level achieved by `FLORAS`. We briefly introduce the basic concepts of DP in FL and subsequently demonstrate that `FLORAS` achieves different levels of DP via the adjustment of the size of spreading sequence set N and the number of involved clients K. For ease of exposition, we define the noise-free global model as $x_i \triangleq \sum_{k=1}^{K} x_k^i$.

We first introduce the concept of "neighboring datasets". We say that two datasets $\mathcal{D}$ and $\mathcal{D}'$ are neighboring, denoted as $\mathcal{D} \sim \mathcal{D}'$, if they differ in at most one data sample. Based on this concept, we state the definition of ϵ-DP as follows.

Definition 1 (ϵ-DP [7])**.** *A randomized algorithm $\mathcal{M} : X^n \to \mathcal{R}$ provides ϵ-DP with $\epsilon > 0$ if, for all pairs of neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ and all (measurable) sets of outcomes $S \subseteq \mathcal{R}$, we have*

$$Pr[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(\mathcal{D}') \in S].$$

We note that ϵ-DP is also known as *pure DP*, and it provides a more strict privacy guarantee than (ϵ, δ)-DP. In the AirComp FL design, decoding the aggregated global model from the received signal can be regarded as a randomized algorithm for all local datasets $\mathcal{D} = \bigcup_{k=1}^{M} \mathcal{D}_k$:

$$\mathcal{M}(\mathcal{D}) \coloneqq \tilde{x}_i = g(\mathcal{D}) + n, \tag{4}$$

where $g(\mathcal{D}) = x_i$ includes all operations to obtain the noise-free x_i in learning round i (including local SGDs, model differential and global aggregation), and n is a random noise following a certain distribution. We note that $\mathcal{M}(\mathcal{D})$ is a randomized algorithm, whose randomness comes from SGD (stohastic mini-batch), random client participation, and additive random noise n. Impact of the "built-in" SGD and random client participation of FL on DP has been investigated [12], and in this conference paper we focus on the randomness of the single-round post-signal-processing noise for `FLORAS`. More comprehensive analyses that consider all sources of randomness and composition of multiple learning rounds will be provided in the journal version.

To determine the DP level provided by the added noise, we need to define the global sensitivity for operation $g(\cdot)$:

$$GS_g = \max_{\mathcal{D},\mathcal{D}'} |g(\mathcal{D}) - g(\mathcal{D}')|, \tag{5}$$

where $\mathcal{D}'$ is a neighboring dataset of $\mathcal{D}$. Note that the difference between $g(\mathcal{D})$ and $g(\mathcal{D}')$ is only contributed from one single client's model differential. As mentioned in Section II-B, for uplink communications, we can adopt the clipping technique to ensure that $|x_k^i| \leq C$, and it is straightforward to show that $GS_g \leq 2C$.

B. Differential Privacy

We now formally establish the DP level of `FLORAS`.

Theorem 1. *Given a spreading sequence set $\mathcal{A}$ containing N unique sequences and K clients ($K < N$) involved in an FL task,* `FLORAS` *provides ϵ-DP within one single uplink communication with respective to dataset $\mathcal{D}$, where*

$$\epsilon \triangleq \frac{4C}{N-K}.$$

Proof. As indicated by Lemma 1, the distribution of $\mathcal{M}(\mathcal{D})$ in (4) follows $\mathsf{Cauchy}(g(\mathcal{D}), N-K)$. Let neighboring datasets $\mathcal{D} \sim \mathcal{D}'$. Then for any $r \in \mathbb{R}$, we have

$$\frac{\Pr[\mathcal{M}(\mathcal{D}) = r]}{\Pr[\mathcal{M}(\mathcal{D}') = r]} = \frac{(N-K)^2 + (r - g(\mathcal{D}'))^2}{(N-K)^2 + (r - g(\mathcal{D}))^2} \triangleq h(r).$$

To prove Theorem 1, we need to show that $h(r) \leq e^{\epsilon}$.

If $g(\mathcal{D}) = g(\mathcal{D}')$, it is easy to find that $h(r) = 1 \leq e^{\epsilon} (\epsilon > 0)$. If $g(\mathcal{D}) \neq g(\mathcal{D}')$, w.l.o.g., we assume $g(\mathcal{D}) > g(\mathcal{D}')$. To simplify notation, we introduce the auxiliary variables $P \triangleq g(\mathcal{D}) + g(\mathcal{D}')$ and $Q \triangleq g(\mathcal{D}) - g(\mathcal{D}')$. By taking the first-order derivative of $h(r)$ and setting $h'(r) = 0$, we can show that $h(r)$ reaches its maximum at $r_{\max} = \frac{P + \sqrt{Q^2 + (2N-2K)^2}}{2}$, with $h(r_{\max}) = 1 + \frac{2Q(\sqrt{Q^2 + 4(N-K)^2} + Q)}{4(N-K)^2}$. Since $Q > 0$ and $N - K > 0$, we have $\sqrt{Q^2 + 4(N-K)^2} \leq Q + 2(N-K)$. Based on the definition of global sensitivity function in (5), we further have $Q \leq |Q| \leq GS_g \leq 2C$. Therefore, we can establish that

$$h(r) \leq h(r_{\max}) = 1 + \frac{2Q(\sqrt{Q^2 + 4(N-K)^2} + Q)}{4(N-K)^2}$$

$$\leq 1+\frac{2Q(2Q+2(N-K))}{4(N-K)^2} \leq 1+\frac{2C(2C+(N-K))}{(N-K)^2}$$
$$=\frac{(N-K)^2+2C(N-K)+4C^2}{(N-K)^2}$$
$$\leq \frac{(N-K)^2+4C(N-K)+4C^2}{(N-K)^2}=\left(1+\frac{2C}{N-K}\right)^2$$
$$=(1+\frac{1}{2}\epsilon)^2 \leq 1+\epsilon+\frac{1}{2}\epsilon^2+O(\epsilon^3)=e^{\epsilon}.$$

□

Theorem 1 reveals that, for a fixed number of clients K, the expansion of spreading sequence set would achieve higher level of ϵ-DP. In particular, $\epsilon \propto \frac{1}{N-K}$. Since the BS (adversary) has no knowledge on the specific K of the total N spreading sequences the clients have chosen, increasing the size of $\mathcal{A}$ results in a heavier tail of the post-decoding Cauchy noise, which achieves better privacy protection. By choosing different sizes of the spreading sequence set and/or selecting different numbers of active clients, we can provably achieve different DP levels (i.e., ϵ) as shown in Theorem 1. Note that larger noise (better privacy protection) will affect the convergence rate of FL. We will discuss this impact next.

C. Convergence

We analyze the convergence performance theoretically of FLORAS. We make the following standard assumptions on loss functions that are commonly adopted in the convergence analyses of FEDAVG and its variants [13].

Assumption 1. ***L-smooth:*** $\forall$ $\mathbf{v}$ *and* $\mathbf{w}$, $\|f_k(\mathbf{v})-f_k(\mathbf{w})\| \leq L\|\mathbf{v}-\mathbf{w}\|$; ***$\mu$-strongly convex:*** $\forall$ $\mathbf{v}$ *and* $\mathbf{w}$, $\langle f_k(\mathbf{v})-f_k(\mathbf{w}), \mathbf{v}-\mathbf{w}\rangle \geq \mu\|\mathbf{v}-\mathbf{w}\|^2$; ***Unbiased SGD:*** $\forall k \in [M]$, $\mathbb{E}[\nabla \tilde{f}_k(\mathbf{w})] = \nabla f_k(\mathbf{w})$; ***Uniformly bounded gradient:*** $\forall k \in [M]$, $\mathbb{E}\left\|\nabla \tilde{f}_k(\mathbf{w})\right\|^2 \leq H^2$.

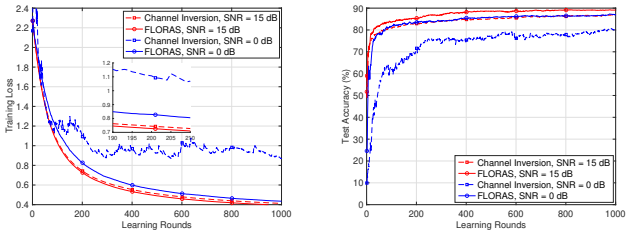
We note that a Cauchy distribution has uncertain (infinity) variance, which brings a significant challenge to the convergence analysis, as the noise variance cannot be bounded. To address this issue, we assume the BS applies a *truncation* operation in the interval $[-B, B]$ on the decoded global parameters in (3):

$$\tilde{x}_i \approx \max\left(\min\left(\sum_{k=1}^{K} x_k^i + \sum_{k=K+1}^{N} \frac{\mathbf{a}_k^T\mathbf{n}_i}{\mathbf{a}_k^T\mathbf{n}_s}, B\right), -B\right) \quad (6)$$

where $B \gg C$. Note that the truncation operation is universal (albeit sometimes implicit) in almost all practical systems, since the signal values in the processing units are always finite. As long as we ensure $B \gg C$, the truncation operation has very little impact on the received signal. For ease of exposition, we make the following assumption on the noise term in (6).

Assumption 2. *The noise term in* (6) *follows a truncated Cauchy distribution within the interval* $[-B, B]$ *with* $B \gg C$, *whose p.d.f. can be expressed as*

$$f_X(x) = \frac{1}{2\arctan\left(\frac{B}{N-K}\right)} \frac{N-K}{x^2+(N-K)^2}, x \in [-B, B].$$



(a) Training loss (b) Test accuracy

Fig. 1. Comparison of FLORAS and the channel inversion method with SNR = 0 and 15 dB.

Note that the DP guarantee of FLORAS in Theorem 1 still holds under Assumption 2. We next establish Theorem 2 as the convergence guarantee of FLORAS.

Theorem 2. *With Assumptions 1 and 2, for some* $\gamma \geq 0$, *if we select the learning rate as* $\eta_t = \frac{2}{\mu(t+\gamma)}$, *a wireless system implementing* FLORAS *for FL uplink communications achieves*

$$\mathbb{E}[f(\mathbf{w}_t)] - f^* \leq \frac{L}{2(t+\gamma)}\left[\frac{4G}{\mu^2} + (1+\gamma)\|\mathbf{w}_0-\mathbf{w}^*\|^2\right],$$

for any $t \geq 1$, *where*

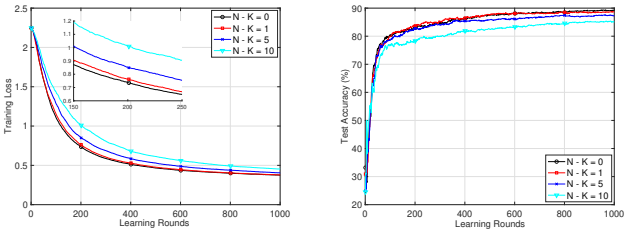
$$G = \sum_{k=1}^{M} \frac{H_k^2}{M^2} + 6L\Gamma + 8(E-1)^2H^2 + \frac{M-K}{M-1}\frac{4}{K}\eta_t^2E^2H^2 + D(\epsilon),$$

and $D(\epsilon) = \frac{64C^2}{K^2\epsilon^2\arctan\left(\frac{B\epsilon}{4C}\right)}\left[\frac{B\epsilon}{4C} - \arctan\left(\frac{B\epsilon}{4C}\right)\right]E^2H^2$.

The proof is omitted due to the space limitation. Theorem 2 demonstrates that FLORAS preserves the $O(1/T)$ convergence rate of SGD. There are multiple components in constant G that affect the convergence speed of FL. In particular, $\sum_{k=1}^{K}\frac{H_k^2}{K^2}$ reveals the *variance reduction* effect of SGD by involving more clients. $6L\Gamma$ and $8(E-1)^2H^2$ highlight the influence of non-i.i.d. datasets and the number of local epochs, respectively. The last term $D(\epsilon)$ in B captures the impact of Cauchy noise, i.e., the level of privacy protection. We note that $D(\epsilon)$ increases as ϵ decreases to 0. It implies that a higher level of privacy protection will decrease the speed of convergence. Therefore, Theorem 2 reveals the trade-off between privacy protection and convergence rate, which provides guidance for the practical system design.

V. EXPERIMENTS

In this section, we evaluate the performance of FLORAS through numerical experiments. We first compare the learning performance of FLORAS with the widely investigated channel inversion AirComp method. Then, we evaluate the effect on the convergence rate of various DP levels. In particular, we validate the theoretical results via real-world FL tasks on the MNIST dataset, under different SNRs and other system configurations. All convergence curves are the average of five individual Monte Carlo trials. Details on the dataset are as follows. The MNIST dataset contains multiple handwritten digit figures of 20×20 pixels. The training



(a) Training loss (b) Test accuracy

Fig. 2. FLORAS with different DP levels with SNR = 20dB.

set contains $4,000$ examples that are evenly distributed over $K = 20$ clients. The data is shuffled and randomly assigned to each client. The test set size is $1,000$. We examine FLORAS and the theoretical results on a multinomial logistic regression problem. Specifically, let $f(\mathbf{w}; x_i)$ denote the predictor with parameter $\mathbf{w} = (\mathbf{W}, \mathbf{b})$ and the form $f(\mathbf{w}; x_i) = \texttt{softmax}(\mathbf{W}x_i + \mathbf{b})$. The loss function is given by $\texttt{loss}(\mathbf{w}) = \frac{1}{D}\sum_{i=1}^{D} \texttt{CrossEntropy}(f(\mathbf{w}; x_i), \mathbf{y}_i) + \lambda \|\mathbf{w}\|^2$. We note that this is a convex optimization problem and we adopt the regularization parameter $\lambda = 0.01$ in the experiments.

We first evaluate the performance of FLORAS compared with the channel inversion method. We assume block Raleigh fading $h_k \sim \mathcal{CN}(0, 1)$. For channel inversion, we adopt a threshold 0.01 for the fading channel gain to avoid deep fading. The following parameters are used for training: local batch size 50, the number of local epochs $E = 1$, and learning rate $\eta = 0.005$. Fig. 1(a) and Fig. 1(b) illustrate the training loss and test accuracy performance versus the learning round of FLORAS and channel inversion with high (red line) and low (blue line) SNR values, respectively. For high SNR, FLORAS and channel inversion have similar performance. However, note that unlike FLORAS, channel inversion requires full CSIT at each client, which incurs significantly larger communication overhead. The advantages of FLORAS become conspicuous in the low SNR case, in which noise becomes the dominant factor of the convergence speed. FLORAS allows all participated clients to make full use of the transmit power and achieves significantly better performance. As shown in Fig. 1(b), FLORAS achieves about 7.5% higher test accuracy compared with channel inversion at SNR = 0 dB.

We next evaluate the convergence performance versus different levels of DP. In this experiment, we keep SNR $= 20$ dB and $K = 20$, while changing the size of available orthogonal sequences in set $\mathcal{A}$ to be $20, 21, 25$ and 30, i.e., $N-K = 0, 1, 5$ and 10. As discussed in Section IV, the larger size of set $\mathcal{A}$, the higher DP level FLORAS achieves. The following parameters are used for training: local batch size 20, the number of local epochs $E = 1$, and learning rate $\eta = 0.005$. The training loss with different privacy levels are shown in Fig. 2(a). It is clear that although higher privacy level decreases the convergence speed, the ML model can still converge with almost the same training loss as that of the no-differential-privacy case ($N - K = 0$). This is consistent with the theoretical analysis in Section IV-C. The test accuracy convergence in Fig. 2(b) further validates the effectiveness of FLORAS. We can see that the test accuracies for moderate DP levels ($N - K = 1$ and 5) are almost the same as the case of $N - K = 0$, i.e., we can achieve certain DP levels *almost for free*. Even when $N - K = 10$, the test accuracy loss is still very small, being about 3.5% compared with the $N - K = 0$ case.

VI. Conclusion

We have proposed FLORAS, a differentially private AirComp FL framework. Compared with the channel inversion method, FLORAS does not require CSIT and performs much more robustly in low SNR cases, which is crucial for IoT applications. The flexibility of adjusting the size of the orthogonal sequence set allows us to easily control the ϵ-DP guarantee of the system. From the analyses of convergence and DP, we have established the trade-off between convergence speed and privacy preservation, which has been further validated by experiments on real-world FL tasks.

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. Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2017, pp. 1273–1282.

[2] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 46–51, 2020.

[3] 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.

[4] G. Zhu, J. Xu, K. Huang, and S. Cui, "Over-the-air computing for wireless data aggregation in massive IoT," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 57–65, 2021.

[5] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1310–1321.

[6] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *Proc. Conference on Computer Communications (INFOCOM)*. IEEE, 2019, pp. 2512–2520.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[8] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3454–3469, 2020.

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

[10] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3454–3469, 2020.

[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[12] Z. Bu, J. Dong, Q. Long, and W. J. Su, "Deep learning with gaussian differential privacy," *Harvard Data Science Review*, vol. 2020, no. 23, 2020.

[13] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.