Differentially Private Federated Learning via Reconfigurable Intelligent Surface

Yuhan Yang®, Student Member, IEEE, Yong Zhou®, Member, IEEE, Youlong Wu®, Member, IEEE, and Yuanming Shi®, Senior Member, IEEE

Abstract—Federated learning (FL), as a disruptive machine learning (ML) paradigm, enables the collaborative training of a global model over decentralized local data sets without sharing them. It spans a wide scope of applications from the Internet of Things (IoT) to biomedical engineering and drug discovery. To support low-latency and high-privacy FL over wireless networks, in this article, we propose a reconfigurable intelligent surface (RIS)-empowered over-the-air FL system to alleviate the dilemma between learning accuracy and privacy. This is achieved by simultaneously exploiting the channel propagation reconfigurability with RIS for boosting the received signal power, as well as the waveform superposition property with over-the-air computation (AirComp) for fast model aggregation. By considering a practical scenario, where high-dimensional local model updates are transmitted across multiple communication blocks, we characterize the convergence behaviors of the differentially private federated optimization algorithm. We further formulate a system optimization problem to optimize the learning accuracy while satisfying privacy and power constraints via the joint design of transmit power, artificial noise, and phase shifts at RIS, for which a two-step alternating minimization framework is developed. Simulation results validate our systematic, theoretical, and algorithmic achievements and demonstrate that RIS can achieve a better tradeoff between privacy and accuracy for over-the-air FL systems.

Index Terms—Biomedical monitoring, differential privacy (DP), federated learning (FL), over-the-air computation (AirComp), reconfigurable intelligent surface (RIS).

I. INTRODUCTION

WITH the rapid advancement of communication technologies for the Internet of Things (IoT), massive amounts of sensory data generated by various edge devices (e.g., smartphones, wearables) can be leveraged to support various intelligent applications and services [1]. However, the concern on data privacy makes the data sharing among edge devices unappealing and hinders the exploration of the potential values for the decentralized data sets. Recently, federated learning (FL) [2] has been recognized as a disruptive machine learning (ML) paradigm that is capable of preserving data privacy without sharing private raw data. FL spans a wide scope of applications from 6G [1], [3], [4], IoT [5], [6], and keyboard prediction [7], to the recent applications in healthcare informatics [8]–[10], medical imaging [11], drug discovery [12], and diabetes mellitus [13]. In particular, for cross-device FL, many edge devices collaboratively train a common model under the orchestration of a central edge server by periodically exchanging their local model updates. Despite the promising benefits, FL still encounters many challenges, including statistical heterogeneity, communication bottleneck, as well as privacy and security concerns [4], [14], [15].

The communication overhead incurred by the periodic model update exchange is a key performance-limiting factor of FL. To enable efficient model aggregation, over-the-air computation (AirComp) has been recently proposed as a promising aggregation scheme by integrating communication and computation [16]. Specifically, multiple edge devices can upload the local models concurrently by exploiting the waveform superposition property of a multiple-access channel [16]–[19], thereby achieving high spectrum efficiency and low transmission latency. In particular, AirComp was leveraged in [17] to reduce latency for FL over broadband channels compared with traditional orthogonal multiple access schemes. The learning performance of over-the-air FL was improved in [16] by jointly optimizing the receive beamforming design and device selection based on AirComp. Amiri and Gündüz [18] also proposed a gradient sparsification scheme for over-the-air stochastic gradient method to reduce the dimensionality of the exchanged updates.

Despite the aforementioned advantages, the heterogeneity of wireless links inevitably degrades the learning performance due to the magnitude attenuation and misalignment of signals received at the edge server. To alleviate the detrimental effect of wireless fading channels, a reconfigurable intelligent surface (RIS) has recently been introduced as a key enabling technology to support fast and reliable model aggregation for over-the-air FL systems by reconfiguring the propagation environment [20]–[22]. Specifically, RIS is a planar metasurface equipped with an array of cost-effective passive reflecting elements, which are orchestrated by a software-enabled controller to adjust the phase shifts of the incident signals [23]–[26]. This helps improve the power of received signals, thereby
enhancing the communication performance between all edge devices and the edge server. The waveform superposition property of a wireless multiple-access channel based on RIS is thus able to adapt to the local model updates, thereby improving the learning performance for over-the-air FL systems via jointly optimizing phase shifts, receive beamforming, and edge device selection [21], [22].

Besides, as illustrated in [27]–[29], transmitting the model updates (e.g., local gradients) can still cause privacy leakage. To quantify the privacy disclosure, a rigorous mathematical framework called differential privacy (DP) [30], [31] has been proposed. Various privacy-preserving mechanisms have been developed by injecting random perturbations that obey specific distributions, e.g., Gaussian [30], Laplacian [32], and Binomial [33]. In particular, Liu and Simeone [34] developed an AirComp-based FL scheme to achieve privacy protection for free. This procedure is accomplished by leveraging the inherent wireless channel noise to protect user privacy. Hasircioğlu and Gündüz [35] proved the inherent anonymity of AirComp which hides each private local update in the crowd to ensure high privacy, thereby reducing the amount of artificial noise added to the local updates. A disturbance scheme with additional perturbations added to a partial set of edge devices to benefit the whole system was proposed in [36]. However, the learning accuracy of FL may significantly degenerate due to the reduction of signal-to-noise ratio (SNR) for privacy guarantee [34], which yields a trade-off between accuracy and privacy.

To balance the privacy–accuracy tradeoff, the existing works [34]–[38] only focus on designing power allocation schemes, without considering the reconfigurability of the wireless environment. In comparison, we propose an RIS-enabled over-the-air FL system to balance the tradeoff between privacy and accuracy, which is achieved by simultaneously exploiting the channel propagation reconfigurability with RIS and waveform superposition property with AirComp. Besides, we consider a practical local model updates transmission scheme by distributing the high-dimensional model updates (e.g., deep neural network model) across multiple communication blocks in one learning round [39]. However, most of the previous works often make the lightweight assumption on the learning models, which requires the dimension of model updates to be small enough to be transmitted within one communication block [22], [34]–[38], [40], [41]. Through convergence analysis and system optimization, we reveal that the RIS-enabled FL system can enhance system SNR and boost the received signal power to establish high learning accuracy while satisfying the privacy requirements.

The major contributions of this article are summarized as follows.

1) From a systematic perspective, we develop an RIS-enabled FL system with privacy guarantees for fast and reliable model aggregation to alleviate the dilemma between learning accuracy and privacy by exploiting channel propagation reconfigurability and waveform superposition property.

2) From a theoretical perspective, for high-dimensional model updates transmission across multiple communication blocks in one learning round, we propose a privacy-preserving transmission scheme based on DP and analyze its convergence behavior.

3) From an algorithmic perspective, we propose a two-step alternating minimization framework to jointly optimize the transmit power, artificial noise, and phase shifts for system optimization, leading to an optimal power allocation scheme.

4) Simulation results validate our systematic, theoretical, and algorithmic achievements and demonstrate the advantages of deploying RIS for enhancing both privacy and accuracy performance in the learning process.

The remainder of this article is organized as follows. The privacy-preserving RIS-enabled FL system is presented in Section II. Section III analyzes the training loss, privacy, and power constraints, yielding a nonconvex system optimization problem. In Section IV, an alternating minimization framework is proposed for solving this nonconvex problem. Simulation results are elaborated in Section V to demonstrate the advantages of the RIS-enabled FL system. Finally, Section VI concludes this work.

Notations: Italic and boldface letters denote scalar and vector (matrix), respectively. \(\mathbb{R}^{m \times n}\) and \(\mathbb{C}^{m \times n}\) denote the real and complex domains with the space of \(m \times n\), respectively. For a positive integer \(i\), we let \([i] = \{1, \ldots, i\}\). The operators \((\cdot)^\dagger\), \((\cdot)^H\), \(\text{Tr} (\cdot)\), rank \((\cdot)\), \(\mathbb{E} (\cdot)\), and \(\text{diag} (\cdot)\) represent the transpose, Hermitian transpose, trace, rank, statistical expectation, and diagonal matrix, respectively. \(\wedge\) denotes logical and operation. The operator \(|\cdot|\) is the cardinality of a set or the absolute value of a scalar number, and \(|\cdot|\) denotes the Euclidean norm.

II. SYSTEM MODEL

In this section, we first introduce an RIS-enabled FL system based on AirComp for fast and accurate model aggregation, followed by presenting a DP-based privacy-preserving pipeline for privacy concerns. The notations are listed in Table I.

A. Distributed Federated Learning

As shown in Fig. 1, we consider an RIS-enabled wireless FL system consisting of one single-antenna edge server, \(K\) single-antenna edge devices indexed by set \(\mathcal{K} = \{1, \ldots, K\}\), and an RIS equipped with \(N\) passive reflecting elements. We assume that each edge device \(k \in \mathcal{K}\) has its own local data set \(\mathcal{D}_k\) with \(D_k = |\mathcal{D}_k|\) data samples. For a given \(d\)-dimensional model parameter \(\theta \in \mathbb{R}^d\), the local loss function of edge device \(k\) is defined as

\[
F_k(\theta) = \frac{1}{\mathcal{D}_k} \sum_{(x, y) \in \mathcal{D}_k} f(x, y; \theta)
\]

where \(f(x, y; \theta)\) represents the samplewise loss function quantifying the prediction error of model \(\theta\) on training sample \(x\) with respect to its true label \(y\) in \(\mathcal{D}_k\). We assume that all local data sets have the same size [17], i.e., \(D_k = D \forall k \in \mathcal{K}\). Then, the global loss function can be defined as

\[
F(\theta) = \frac{1}{\sum_{k=1}^{K} \mathcal{D}_k} \sum_{k=1}^{K} \mathcal{D}_k F_k(\theta) = \frac{1}{K} \sum_{k=1}^{K} F_k(\theta)
\]
TABLE I
SUMMARY OF NOTATIONS IN THIS ARTICLE

| Description | Notation |
|-------------|----------|
| Number of edge devices | $K$ |
| Size of local dataset in the $k$-th edge device | $D_k$ |
| Channel coherence length | $\epsilon$ |
| $d$-dimensional model parameter | $\theta$ |
| Local update of edge device $k$ in the $t$-th learning round | $g_{k,t}$ |
| Transmit scalar of edge device $k$ in the $t$-th block | $\alpha_k(t)$ |
| Pre-processed signal of edge device $k$ in the $t$-th block | $s_{k,t}(i)$ |
| Aggregated signal in the $t$-th block of learning round $t$ | $r_t(i)$ |
| Channel coefficient from edge device $k$ to the edge server | $h_{k,t}^x(i)$ |
| Phase shift matrix of RIS in the $t$-th block | $\Theta_t(i)$ |
| Power of artificial noise | $\sigma_{k,t}^2(i)$ |

| Description | Notation |
|-------------|----------|
| Number of passive reflecting elements of RIS | $N$ |
| Total number of learning rounds | $T$ |
| Number of communication blocks in one learning round | $I$ |
| Privacy level and failure probability | $(\epsilon, \delta)$ |
| Artificial noise of edge device $k$ in the $t$-th block | $n_{k,t}(i)$ |
| Uniform power scaling factor in the $t$-th block | $\eta_t(i)$ |
| Transmit signal of edge device $k$ in the $t$-th block | $s_{k,t}(i)$ |
| Additive white Gaussian noise in the $t$-th block | $w_t(i)$ |
| Channel coefficient from edge device $k$ to RIS | $h_{k,t}^r(i)$ |
| Composite channel response of the $k$-th edge device | $P_0$ |
| Maximum transmit power of all edge devices | $N_0$ |

which refers to the empirical average of the samplewise loss functions on the global data set $D = \bigcup_{k=1}^K D_k$. We also assume that each edge device observes independent and identically distributed (i.i.d.) samples from a common distribution [34]. The non-i.i.d. setting in FL is also of special interest [42], but it is not the focus of this work.

The learning process minimizes (2) by updating $\theta$ based on the local gradients sent by edge devices, i.e.,

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} F(\theta).$$

(3)

This problem can be tackled by using the popular decentralized optimization method, e.g., FedSGD [2]. For analytical ease, in this article, full-batch gradient descent is adopted for local model update and $\theta$ is updated periodically within $T$ training rounds [16], [34], [38]. The $t$th learning round consists of the following procedures.

1) **Broadcast**: The edge server broadcasts the current global model parameter $\theta_t$ to all edge devices.

2) **Local Update**: Each edge device $k$ computes its local update $g_{k,t}$ with respect to its local data set $D_k$ based on $\theta_t$, i.e.,

$$g_{k,t} = \nabla F_k(\theta_t) \in \mathbb{R}^d.$$  

(4)

3) **Model Aggregation**: All edge devices in $\mathcal{K}$ upload $\{g_{k,t}\}$ to the edge server. Then, the edge server updates $\theta_{t+1}$ based on the aggregated local updates via gradient descent, i.e.,

$$\theta_{t+1} = \theta_t - \lambda \hat{g}_t,$$

(5)

where $\lambda$ denotes the learning rate and $\hat{g}_t$ is an estimation of the global gradient $g_t = \nabla F(\theta_t) = (1/K) \sum_{k=1}^K \hat{g}_{k,t}$.

### B. Channel Model for RIS-Enabled FL Systems

The existence of unfavorable channel propagations and the power limitation of each edge device may severely degrade the accuracy for model aggregation, thereby reducing the learning accuracy of the FL system. To address this issue, an RIS equipped with $N$ reflecting elements is deployed to enhance the channel conditions of the channel links between all edge devices and the edge server [22].

We focus on the information exchange process among the edge server, RIS, and edge devices, as shown in Fig. 1. To simplify the theoretical analysis, the downlink channels are assumed to be noise free [18], [41], while we mainly consider the uplink fading channels for aggregating the local models [21]. As the dimension $d$ of the model parameters is often much larger than the channel coherence length $\epsilon$ [39], we propose a practical transmission scheme illustrated in Fig. 2, where each edge device uploads its local update sequentially over several consecutive communication blocks. Specifically, in the $t$th learning round, the whole update message $x_{k,t} = [x_{k,t}^1(1), \ldots, x_{k,t}^I(I)]^T \in \mathbb{C}^d$ (which will be elaborated in Section II-D) that is a function of $g_{k,t}$ is evenly divided into $I = \lceil d/\epsilon \rceil$ $\epsilon$-dimensional vectors denoted by

![Fig. 1. RIS-enabled wireless FL system.](image-url)
Then, \( \{x_{k,t}(i)\} \) are sequentially transmitted across \( I \) communication blocks. We further assume \( I = d/e \) for simplicity.

For the \( t \)th communication block in learning round \( t \), let \( h^d_{k,t}(i) \in \mathbb{C}, h^r_{k,t}(i) \in \mathbb{C}^N \), and \( m(i) \in \mathbb{C}^N \) denote the channel responses from edge device \( k \) to the edge server, from edge device \( k \) to RIS, and from RIS to the edge server, respectively. We assume that the channel coefficients are independent across different communication blocks and remain invariant during one communication block. We also assume that perfect channel state information (CSI) is available [18], [21], [22]. Moreover, the phase-shift matrix of the RIS is denoted as \( \Theta_i(i) = \text{diag}(\beta \epsilon^2 \delta(i), \ldots, \beta \epsilon^2 \delta(i)) \in \mathbb{C}^{N \times N} \), where \( \phi_{n,i} \in [0, 2\pi], n \in [1 \ldots N] \). Without loss of generality, the amplitude reflection coefficient \( \beta \) at each reflecting element is set to be one [43]. We assume that \( \Theta_i(i) \) is designed at the edge server and then transmitted to the RIS via error-free downlink channels. We also assume that signals reflected twice or more times by the RIS can be ignored [24]. The composite channel response of the \( k \)th edge device is thus denoted as \( h_{k,t}(i) = m^T(i) \Theta_i(i) h^r_{k,t}(i) + h^d_{k,t}(i) \).

In the \( t \)th communication block of learning round \( t \), the aggregated signal at the edge server is given by

\[
\tilde{r}_t(i) = \sum_{k=1}^{K} h_{k,t}(i)x_{k,t}(i) + w_t(i) \tag{6}
\]

where \( w_t(i) \sim \mathcal{CN}(0, \sigma^2) \) is the additive white Gaussian noise (AWGN) and the aggregated signal \( \tilde{r}_t \in \mathbb{C}^d \) can be written as \( \tilde{r}_t = [\tilde{r}_t^1(1), \ldots, \tilde{r}_t^I(T)]^T \) for the \( t \)th learning round.

### C. Differential Privacy

In the RIS-enabled FL system, the sensitive raw data never leave the edge devices to protect users’ privacy. Nevertheless, as revealed in [29], the transmission of model updates \( g_{k,t} \) may leak information about the local data set statistically, which requires additional mechanisms to provide stronger privacy guarantees. We assume that the edge server is honest-but-curious [40], i.e., the edge server aims to infer the local information of each edge device based on the received signals \( r = [r_t]_{t=1}^T \). We introduce DP based on the concept of neighboring data sets. Let \( D_k = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}^n \) denote a data set comprising \( n \) data points from \( \mathcal{X} \). Two data sets \( D_k = \{x_1, \ldots, x_n\} \) and \( D'_k = \{x'_1, \ldots, x'_n\} \) with the same cardinality are neighboring if they differ only by one element, i.e., there exists an index \( i \in [n] \) such that \( x_i \neq x'_i \) and \( x_j = x'_j \) for all \( i \neq j \). We then present the following definition [30], [31] for DP.

**Definition 1:** Given \( \epsilon > 0, 0 \leq \delta \leq 1 \), and an arbitrary protocol \( \mathcal{M} : \mathcal{X}^n \rightarrow \mathcal{Y} \). For any two possible neighboring data sets \( D_k, D'_k \in \mathcal{X}^n \) and any subset \( \Lambda \subseteq \mathcal{Y} \), protocol \( \mathcal{M} \) is \( (\epsilon, \delta) \)-differentially private (in short, \( (\epsilon, \delta) \)-DP), if the following inequality holds:

\[
\Pr[\mathcal{M}(D_k) \in \Lambda] \leq e^\epsilon \Pr[\mathcal{M}(D'_k) \in \Lambda] + \delta. \tag{7}
\]

Note that the case of \( \delta = 0 \) is called pure \( \epsilon \)-DP.

Specifically, for edge device \( k \), we set the received signal \( r \) as test variable to be the output of protocol \( \mathcal{M} \). Based on this, \( (\epsilon, \delta) \)-DP guarantees that for any neighboring data sets \( D_k \) and \( D'_k \), the DP loss

\[
\ell_{D_k, D'_k}(r) = \ln \frac{\Pr(r|D_k)}{\Pr(r|D'_k)}. \tag{8}
\]

That is, the log-likelihood ratio of the neighboring data sets \( D_k \) and \( D'_k \) satisfies

\[
\Pr(\ell_{D_k, D'_k}(r) \leq \epsilon) \geq 1 - \delta. \tag{9}
\]

From (9), we can see that a lower \( \epsilon \) yields a higher privacy guarantee for the FL system.

To achieve the privacy level \( \epsilon \), a common method is to add random perturbations to the signals. We thus uniformly assign the artificial noise \( n_{k,t} = [n^1_{k,t}(1), \ldots, n^I_{k,t}(T)]^T \in \mathbb{C}^d \) with \( n_{k,t}(i) \sim \mathcal{CN}(0, \sigma^2) \). Each communication block, yields a noisy version of the local update \( g_{k,t} \). Therefore, the preprocessed transmit signal of edge device \( k \) is given as

\[
s_{k,t}(i) = D_t g_{k,t}(i) + n_{k,t}(i). \tag{10}
\]

Note that, for simplicity, we omit some procedures such as gradient clipping [44] which may preclude a large value of \( D_t g_{k,t}(i) \).

### D. Model Aggregation via Over-the-Air Computation

We leverage AirComp for fast model aggregation by exploiting the signal superposition of a wireless multiple-access channel [20]. In AirComp, all edge devices transmit their local updates \( [g_{k,t}(i)]_{k=1}^K \) synchronously using the same time-frequency resources in the communication block, thereby achieving high communication efficiency [16].

Motivated by [45], based on the perfect CSI of each edge device, we design the block-based uniform-forcing transmit signal \( x_{k,t}(i) \) in the \( t \)th communication block of learning round \( t \) as follows:

\[
x_{k,t}(i) = \alpha_{k,t}(i) s_{k,t}(i) = \frac{\sqrt{\eta_t(i)}}{|h_{k,t}(i)|} s_{k,t}(i) \tag{11}
\]

where \( \alpha_{k,t}(i) \in \mathbb{C} \) denotes the transmit scalar which is given by \( \alpha_{k,t}(i) = \sqrt{\eta_t(i)/h_{k,t}(i)} \) and \( \sqrt{\eta_t(i)} > 0 \) is the uniform power scaling factor. Given a maximum transmit power \( P_0 > 0 \) for all edge devices, we have the following power constraints in each communication block:

\[
E\left(\left|\left|x_{k,t}(i)\right|^2\right|\right) = E\left(\left|\alpha_{k,t}(i)s_{k,t}(i)\right|^2\right) \leq P_0 \quad \forall k, i, t \tag{12}
\]

by taking the expectation over the additive Gaussian noise in (10). Note that for each communication block, \( \eta_t(i) \) and \( \Theta_i(i) \) are elaborately designed by the edge server and broadcasted to the edge devices and RIS, which result in additional communication overheads.

Therefore, based on (10) and (11), the aggregated signal \( r_t(i) \) defined in (6) can be rewritten as

\[
r_t(i) = \sqrt{\eta_t(i)} \sum_{k=1}^{K} D_t g_{k,t}(i) + \sqrt{\eta_t(i)} \sum_{k=1}^{K} n_{k,t}(i) + w_t(i). \tag{13}
\]
Then, the edge server computes the estimated global update \( \hat{g}_t(i) \) by
\[
\hat{g}_t(i) = \frac{1}{KD \sqrt{\eta_t(i)}} \text{Re}[r_t(i)] \\
= g_t(i) + \frac{1}{KD} \sum_{k=1}^{K} \text{Re}[n_{k,t}(i)] + \text{Re}[w_t(i)]
\]
where \( g_t(i) = \sum_{k=1}^{K} g_{k,t}(i)/K \). After \( I \) rounds of model aggregation, the resulting \( \hat{g}_t = [\hat{g}_1^T(i), ..., \hat{g}_I^T(i)]^T \) is an unbiased estimation of the true global gradient \( g_t = [g_1^T(i), ..., g_I^T(i)]^T \). But the existence of the artificial noise and wireless channel noise leads to inevitable inaccuracy in \( \hat{g}_t \), which yields a trade-off between accuracy and privacy. In particular, the learning accuracy also depends on SNR, which is defined as the ratio of the maximum transmit power and the noise power in one communication block, i.e.,
\[
\text{SNR} = \frac{P_0}{e \times N_0}
\]
where \( e \times N_0 \) represents the channel noise power within one communication block.

**Remark 1 (Symbol-Level Synchronization):** Note that the AirComp-based aggregation rule relies on symbol-level synchronization among edge devices. To achieve this, a feasible technique is timing advance mechanism [46], which is widely used in 4G long-term evolution (LTE) and 5G new radio (NR) systems. Specifically, consider the case when deploying AirComp in traditional orthogonal frequency-division multiplexing (OFDM) systems, we note that 1-MHz synchronization bandwidth can limit the timing offset within 0.1 \( \mu s \) [47], which is shorter than the typical length of cyclic prefix (5 \( \mu s \)) in LTE systems. Thus, the symbol-level synchronization for AirComp is feasible.

**Remark 2 (Limitation of Uniform-Forcing Scalar):** According to (11), channel-inversion power control is adopted to achieve the amplitude alignment at the edge server. However, due to the power constraint in (12), the uniform-forcing approach may be inefficient when some edge devices encounter deep fadings, which results in a small value of \( \eta_t(i) \). To overcome this issue, we can adopt the truncated-channel-inversion scheme [17], which allocates power to an edge device only if its channel gain exceeds a hard threshold. Besides, the deployment of RIS can also alleviate deep fadings by reconfiguring the wireless environment, which is elaborated in Section IV-A.

**E. Assumptions**

We list several widely used assumptions [22], [48] for theoretical analysis of differentially private FL systems.

**Assumption 1 (Strong Convexity):** The global loss function \( F(\theta) \) is strongly convex with parameter \( \mu \), i.e.,
\[
F(\theta) \geq F(\theta') + \nabla F(\theta')^T (\theta - \theta') + \frac{\mu}{2} \| \theta - \theta' \|^2 \quad \forall \theta, \theta'.
\]

**Assumption 2 (Smoothness):** The global loss function \( F(\theta) \) is Lipschitz continuous with parameter \( L \), i.e.,
\[
F(\theta) \leq F(\theta') + \nabla F(\theta')^T (\theta - \theta') + \frac{L}{2} \| \theta - \theta' \|^2 \quad \forall \theta, \theta'.
\]

which implies the following inequality:
\[
\| \nabla F(\theta) - \nabla F(\theta') \| \leq L \| \theta - \theta' \| \quad \forall \theta, \theta' \in \mathbb{R}^d.
\]

**Assumption 3 (Block Gradient Bound):** In any learning round \( t \), for any training data sample \((x, y)\), the samplewise loss function is upper bounded by a given constant \( \gamma_t \), i.e.,
\[
\| \nabla f(x, y; \theta_t) \| \leq \gamma_t \quad \forall t.
\]

Based on the triangular inequality, for edge device \( k \), there always exists a constant \( \zeta_{k,t}(i) \) satisfying \( \| g_{k,t}(i) \| \leq \zeta_{k,t}(i) \).

**III. PERFORMANCE ANALYSIS AND SYSTEM OPTIMIZATION**

In this section, we shall analyze the convergence behavior of the RIS-enabled FL system and show the key component of the privacy-preserving mechanism. We then provide the system optimization formulation to model the dilemma between privacy and accuracy.

**A. Convergence Analysis**

In the \( r \)th learning round, the edge server updates the global model based on the received signal \( r_t \). However, the fading channels \( \{h_{k,t}(i)\} \) as well as the artificial noise \( \{n_{k,t}(i)\} \) and AWGN \( \{w_t(i)\} \) may severely degrade the learning performance. To characterize the convergence behavior of the RIS-enabled FL system, we leverage the expected value of the gap between the optimal value \( F^* \) and the iteration value \( F(\theta_t) \) to measure the distortion over \( T \) learning rounds, i.e.,
\[
\mathbb{E}[F(\theta_{T+1}) - F^*].
\]

**Theorem 1:** Under Assumptions 1 and 2, given learning rate \( \lambda = 1/L \), after \( T \) learning iterations, the averaged optimality gap is upper bounded as
\[
\mathbb{E}[F(\theta_{T+1})] - F^* \leq \left(1 - \frac{\mu}{L}\right)^T \left[F(\theta_1) - F^*\right] + \frac{1}{4L} \\
\times \frac{e}{(KD)^2} \sum_{i=1}^{N_0} \left(1 - \frac{\mu}{L}\right)^{T-t} \left(\sum_{k=1}^{K} \sum_{i=1}^{I} \sigma_{k,i}^2(i) + \sum_{i=1}^{N_0} \eta_t(i)\right).
\]

**Proof:** Refer to Appendix A.

According to Theorem 1, we observe that the average optimality gap \( \mathbb{E}[F(\theta_{T+1}) - F^*] \) is upper bounded by two terms. The first term will converge to 0 as \( T \) goes to infinity, while the second term in terms of \( \{\sigma_{k,t}(i), \eta_t(i)\} \) cannot vanish due to the added artificial noise and wireless channel noise in each communication block. Besides, (18) is also related to \( \Theta_t(i) \) as the selection of \( \eta_t(i) \) depends on \( \Theta_t(i) \), which will be presented in Theorem 3.

**B. Privacy-Preserving Mechanism**

We develop a privacy-preserving scheme based on DP to alleviate privacy concerns for edge devices. We assume that \( \{\eta_t(i), \sigma_{k,t}(i), \Theta_t(i)\} \) are fixed constants, as they do not reveal any information about the local data sets. Therefore, we only focus on the upload process from edge devices to the edge
server. In the $t$th learning round, for edge device $k$, the disclosed signals regarding its local data set $\mathbb{D}_k$ are $\{g_{k,t}(i)\}$ across $I$ blocks. We decouple the received signal $r_t(i)$ as

$$r_t(i) = \sqrt{\eta_t(i)D_k}g_{k,t}(i) + \sqrt{\eta_t(i)} \sum_{j \neq k} D_j g_{j,t}(i) + q_t(i)$$

(19)

where

$$q_t(i) = \sqrt{\eta_t(i)} \sum_{j=1}^{K} n_{j,t}(i) + w_t(i).$$

(20)

In (19), the first term is the disclosed signal concerning $\mathbb{D}_k$, the second term is independent of edge device $k$, and the third term $q_t(i)$ is the Gaussian noise. We observe that the privacy-preserving disturbance in (20) includes the added artificial noise and the inherent channel noise. Besides, all update messages, i.e., the first and second terms in (19), enjoy the same privacy protection provided by $q_t(i)$. For notational ease, let $\mathbf{q}_t = \{q_t^T(1), \ldots, q_t^T(I)\}^T$ denote the effective noise to ensure privacy for the transmission of $\{g_{j,t}(i)\}_{i=1}^I$ in the $t$th learning round.

For edge device $k$, its privacy level $(\epsilon, \delta)$ depends on the sensitivity of the disclosed noise-free function regarding its local data set. We first recall the definition of $\ell_2$-sensitivity [30].

**Definition 2:** Let $\mathcal{M}$ be an arbitrary mechanism on the transmitted signals. The $\ell_2$-sensitivity $\Delta$ is defined as the maximum difference of the outputs from two neighboring data sets $\mathbb{D}_k$ and $\mathbb{D}_k'$, i.e.,

$$\Delta = \max_{\mathbb{D}_k, \mathbb{D}_k'} \|\mathcal{M}(\mathbb{D}_k) - \mathcal{M}(\mathbb{D}_k')\|_2.$$ 

(21)

Recalling that the first term in (19) is the only disclosed function concerning $\mathbb{D}_k$, let $h_{k,t}$ denote the difference between the outputs from two neighboring data sets $\mathbb{D}_k$ and $\mathbb{D}_k'$, which is given by

$$u_{k,t} = \left[h_{k,t}(1)\alpha_{k,t}(1)\Delta g_{k,t}^T(1), \ldots, h_{k,t}(I)\alpha_{k,t}(I)\Delta g_{k,t}^T(I)\right]^T$$

(22)

where

$$\Delta g_{k,t}(i) = \sum_{(x,y) \in \mathbb{D}_k} \nabla f_i(x, y; \theta_t) - \sum_{(x,y) \in \mathbb{D}_k'} \nabla f_i(x, y; \theta_t)$$

with $\nabla f_i(x, y; \theta_t)$ denoting the $(i-1)\epsilon + 1$th to $i\epsilon$-th elements in $\nabla f(x, y; \theta_t)$. Then, the $\ell_2$-sensitivity $\Delta_{k,t}$ of device $k$ is

$$\Delta_{k,t} = \max_{\mathbb{D}_k, \mathbb{D}_k'} \|u_{k,t}\|_2.$$ 

(23)

By using the triangular inequality and Assumption 3, we have

$$\Delta_{k,t} \leq 2\gamma \max_{i \in [I]} \sqrt{\eta_t(i)}.$$ 

(24)

Based on (24), the privacy constraint for all edge devices is given in the following theorem.

**Theorem 2:** For any fixed sequence of $(\eta_t(i), \sigma_{k,t}(i))_{t=1}^T$, the RIS-enabled FL system satisfies $(\epsilon, \delta)$-DP if we have

$$\sum_{t=1}^T \frac{4\gamma^2}{\xi_t} \max_{i \in [I]} \eta_t(i) \leq \mathcal{R}_{dp}(\epsilon, \delta)$$

(25)

denotes a lower bound of the power of the effective noise $q_t(i)$, and $\mathcal{R}_{dp}(\epsilon, \delta) = (\sqrt{\epsilon} + [C^{-1}(1/\delta)]^2 - C^{-1}(1/\delta))^2$ with $C^{-1}(x)$ denoting the inverse function of $C(x) = \sqrt{\pi x}e^x$.

**Proof:** Refer to Appendix B.

From Theorem 2, we emphasize that both the artificial noise and the receiver noise contribute to the privacy guarantee. Besides, based on (26), $\xi_t$ represents a lower bound of the weakest privacy-preserving perturbation (i.e., the lowest effective noise power) across $I$ communication blocks in the $t$th learning round. Furthermore, the privacy guarantee is related to the joint effect of $\xi_t$ and the maximum transmit power scalar $\eta_t(i)$ among $I$ communication blocks. We also notice that the relationship between the privacy constraint and the phase-shift matrix $\Theta_t(i)$ is similar to Theorem 1.

C. System Optimization

To design the RIS-enabled FL system, we shall propose to minimize the optimality gap in (18) while satisfying the privacy and the maximum transmit power constraints given in (25) and (12), respectively, across $T$ learning rounds. Specifically, given the privacy level $(\epsilon, \delta)$, we jointly optimize $\{\sigma_{k,t}(i), \Theta_t(i), \eta_t(i), \xi_t\}$ in each communication block. Based on (10)–(12), we obtain the power constraints

$$\mathbb{E} \left( \|x_t(i)\|^2 \right) = \frac{\eta_t(i)}{\|h_{k,t}(i)\|^2} \left[ D_k^2 \xi_{k,t}(i) + 2 \eta_t(i) \sum_{l=1}^I \eta_l(i) \right] \leq P_0 \quad \forall i, k, t$$

(27a)

subject to

$$\sum_{i=1}^T \frac{4\gamma^2}{\xi_t} \max_{i \in [I]} \eta_t(i) \leq \mathcal{R}_{dp}(\epsilon, \delta)$$

(27b)

$$\xi_t \leq \eta_t(i) \sum_{k=1}^K \sigma_{k,t}(i)^2 + N_0 \quad \forall i, t$$

(27c)

$$\eta_t(i) \leq \frac{m_t^c(i) \Theta_t(i) h_{k,t}(i) + h_{k,t}(i)^2}{\|\Theta_t(i)(n, n)\|_1 = 1} \quad \forall i, n, t$$

(27d)

where (27b) and (27c) denote the privacy constraints for all edge devices as presented in (25) and (26), respectively, (27d) represents the power constraints by substituting $h_{k,t}(i) = m_t^c(i) \Theta_t(i) h_{k,t}(i) + h_{k,t}(i)^2$ into (12), and (27e) indicates the unit modulus constraints of all reflecting elements at RIS with $\Theta_t(i)(n, n)$ denoting as the $(n, n)$th entry of $\Theta_t(i)$. 

Unfortunately, due to the nonconvexity of the privacy and power constraints, and the nonconvex unimodular constraints of the phase shifts at RIS, problem $\mathcal{P}$ is highly intractable and computationally challenging.

IV. TWO-STEP ALTERNATING LOW-RANK OPTIMIZATION FRAMEWORK

In this section, we propose a two-step alternating minimization framework for solving problem $\mathcal{P}$. Specifically, the transmit power scalar $\eta_t(i)$ and artificial noise $\sigma_{k,t}(i)$ at each edge device, as well as the phase-shift matrix $\Theta_t(i)$ at the RIS are optimized in an alternative manner until the algorithm converges.

A. Co-Design of Artificial Noise and Power Scalar

The main idea of this framework is to alternately optimize $[\eta_t(i), \sigma_{k,t}(i), \Theta_t(i), \xi_t]$ in $T$ learning rounds. Note that problem $\mathcal{P}$ is a noncausal setting, where parameters $[h_{k,t}(i), m_t(i), h_{k,t}^d(i), \xi_t(i), \gamma_t]$ need to be known in advance, which will be discussed in Section VI-C.

Given the phase-shift matrix $\Theta_t(i)$, problem $\mathcal{P}$ is reduced to problem $\mathcal{P}_1$ as follows:

\[
\begin{align*}
\min_{\{\eta_t, \sigma_{k,t}, \xi_t\}} & \quad \sum_{t=1}^{T} \left(1 - \frac{\mu}{L}\right)^{t} \left(\sum_{t=1}^{T} \sum_{i=1}^{K} \sigma_{k,t}^2(i) + \sum_{i=1}^{I} N_0 \eta_t(i)\right) \\
\text{subject to} & \quad \sum_{t=1}^{T} \frac{4\eta_t^2}{\xi_t} \max_{i \in [I]} \eta_t(i) \leq \mathcal{R}_dp(\epsilon, \delta) \\
& \quad \xi_t \leq \eta_t(i) \sum_{k=1}^{K} \sigma_{k,t}^2(i) + N_0 \quad \forall i, t \\
& \quad \frac{D^2 h_{k,t}^2(i) + e^2 \sigma_{k,t}^2(i)}{|h_{k,t}^2(i)|^2} \eta_t(i) \leq P_0 \quad \forall i, k, t
\end{align*}
\]

where $h_{k,t}(i) = m_t(i) \Theta_t(i) h_{k,t}^d(i) + h_{k,t}^d(i)$ denotes the composite channel response which remains constant given $\Theta_t(i)$. We note that through a change of variables, problem $\mathcal{P}_1$ can be transformed into a convex problem tackled by KKT conditions, leading to an adaptive power allocation mechanism, i.e., Theorem 3. For analytical ease, we denote $\tau_t$ as the number of communication blocks which are restricted by privacy in the $t$th learning round.

**Theorem 3:** The optimal solution to problem $\mathcal{P}_1$ is given by

\[
\xi_t = N_0 \quad \forall t, \quad \sigma_{k,t}(i) = 0 \quad \forall i, k, t
\]

and the selection of $[\eta_t(i)]$ depends on the joint effect of privacy and power, which yields the following three cases.

1) **Privacy Field:** If $\tau_t = I$ for all learning rounds, then the optimal $[\eta_t(i)]_{t=1}^T$ remains a constant $\eta$ with

\[
\eta_t = \frac{N_0 \sqrt{T}}{2\gamma_t \beta_t} \left(1 - \frac{\mu}{L}\right) \frac{1}{L} \wedge \eta_t < P_0 \min_{i,k} \frac{|h_{k,t}(i)|^2}{D^2 h_{k,t}^2(i)}
\]

2) **Privacy Power Field:** It is a generalized version of 1) with $\tau_t \in [I]$, and the optimal $\eta_t(i)$ is given by

\[
\eta_t(i) = \min \left\{ \frac{N_0 \sqrt{T}}{2\gamma_t \beta_t} \left(1 - \frac{\mu}{L}\right) \frac{1}{L} P_0 \min_{i,k} \frac{|h_{k,t}(i)|^2}{D^2 h_{k,t}^2(i)} \right\} = \mathcal{R}_dp(\epsilon, \delta).
\]

3) **Power Field:** If $\tau_t = 0$ holds for all learning rounds, i.e.,

\[
\sum_{t=1}^{T} \frac{4\eta_t^2}{\xi_t} \max_{i \in [I]} \eta_t(i) = \sum_{t=1}^{T} \frac{4\eta_t^2}{\xi_t} \min_{i \in [I]} \left\{ \frac{N_0 \sqrt{T}}{2\gamma_t \beta_t} \left(1 - \frac{\mu}{L}\right) \frac{1}{L} P_0 \right\} \left\{ \min_{i,k} \frac{|h_{k,t}(i)|^2}{D^2 h_{k,t}^2(i)} \right\} < \mathcal{R}_dp(\epsilon, \delta)
\]

the unique optimal power scalar $\eta_t(i)$ is given by

\[
\eta_t(i) = P_0 \min_{i,k} \frac{|h_{k,t}(i)|^2}{D^2 h_{k,t}^2(i)}.
\]

**Proof:** Refer to Appendix C.

From Theorem 3, the power of artificial noise becomes zero, which indicates that the additive channel noise in model aggregation serves as an inherent privacy-preserving mechanism to guarantee DP levels for each edge device [34]. Besides, as shown in Fig. 3, three cases characterize the learning accuracy of the FL system. Essentially, the power allocation scheme, i.e., the selection of $\eta_t(i)$ under different $\epsilon$ causes the difference in learning accuracy. To make it precise, Fig. 4 illustrates this process in one learning round. Specifically, two power
Power alignment for each edge device in the $t$-th learning round ($K = 5, I = 5$)

![Diagram of power allocation]

Fig. 4. Power allocation under different privacy levels in one learning round. The blue lines indicate the privacy limitations.

alignment schemes are developed for comparison, where one represents the maximum transmit power concerning privacy requirement $\{\epsilon, \tau_i\}$, and the other indicates the case without privacy, i.e., the Minimum column in Fig. 4. We note that $\tau_i$, which represents the impact of $\epsilon$ across $I$ communication blocks, is the key component of Theorem 3. Concretely, $\tau_i$ indicates the number of communication blocks restricted by privacy in one learning round, i.e., the number of yellow bars in Fig. 4. Consequently, different $\epsilon$ results in a different selection of $\tau_i$, yielding the following three cases.

1) Case (a): Due to the extremely strict privacy level, all communication blocks in this case are restricted by privacy, i.e., $\tau_i = I$.

2) Case (b): In this case, several communication blocks are restricted by privacy while others are limited by transmit power, i.e., $\tau_i \in [I]$. Besides, cases (a) and (b) must satisfy the equality in (27b), which is achieved by adaptively selecting $\{\beta, \tau_i\}$ through enumeration.

3) Case (c): In this case, due to the quite loose privacy level, all communication blocks are limited by the transmit power constraints, i.e., $\tau_i = 0$.

We reveal the benefits achieved by RIS to the above three cases. Due to the reconfigurable capability of RIS, we note that the RIS-enabled FL system is able to establish better channel conditions compared to the FL systems without RIS. Besides, according to (32) and (34), the RIS-enabled FL system can enjoy higher transmit power and enhance the power of received signals at the edge server, thereby resulting in high learning accuracy. However, when the privacy level is strict, i.e., case (a), the learning accuracy is restricted by privacy and cannot be improved by RIS. We will further explore the significant impacts brought by RIS on the learning accuracy via simulations in Section V.

B. Design of Phase-Shift Matrix

On the other hand, for given the artificial noise and power scalar $\{\alpha_k(i), \eta_i(i)\}$, problem $P$ becomes the following nonconvex feasibility detection problem $P_2$:

\begin{align}
\text{find } & \Theta_i(i) \\
\text{subject to } & \frac{\eta_i(i) D^2 c^2_{k,t}(i)}{|m_t^T(i) \Theta_i(i) h_t^d_{k,t}(i) + h_k^d_{k,t}(i)|^2} \leq P_0 \quad \forall i, k, t, \\
& |\Theta_i(i)(n, n)| = 1 \quad \forall i, n, t.
\end{align}

For analytical ease, by denoting $v_n(i) = \psi^{(n)}(i)$ and $c^H_{k,t}(i) = m_t^T(i) \Theta_i(i) h_t^d_{k,t}(i) = c^H_{k,t}(i) v_i(i)$, where $v_i(i) = [v_{1,i}(i), \ldots, v_{N,i}(i)]^T$. Therefore, problem (35) is further transformed into problem $P_{2.1}$.

\begin{align}
\text{find } & v_i(i) \\
\text{subject to } & \frac{\eta_i(i) D^2 c^2_{k,t}(i)}{|c^H_{k,t}(i) v_i(i) + h_k^d_{k,t}(i)|^2} \leq P_0 \quad \forall k, t, \\
& |v_{n,i}(i)| = 1 \quad \forall n \in [N].
\end{align}

Nevertheless, problem $P_{2.1}$ is still nonconvex and inhomogeneous. To develop an efficient algorithm, by introducing an auxiliary variable $\hat{v}_i(i)$, problem (36) can be equivalently reformulated as a homogeneous nonconvex quadratically constrained quadratic programming (QCQP) problem [24], [49], which is given by the following problem $P_{2.2}$.

\begin{align}
\text{find } & \hat{v}_i(i) \\
\text{subject to } & \hat{v}_i^H(i) R_{k,t}(i) \hat{v}_i(i) + |h_k^d_{k,t}(i)|^2 \geq \eta_i(i) D^2 c^2_{k,t}(i) / P_0 \quad \forall i, k, t, \\
& |\hat{v}_{n,i}(i)| = 1 \quad \forall n \in [N + 1]
\end{align}

where

\[
R_{k,t}(i) = \begin{bmatrix}
    c_{k,t}(i) h_t^d_{k,t}(i) c_{k,t}(i)^H & c_{k,t}(i) h_t^d_{k,t}(i) h_k^d_{k,t}(i) \\
    c_{k,t}(i) h_t^d_{k,t}(i) h_k^d_{k,t}(i)^H & 0
\end{bmatrix}, \quad \hat{v}_i(i) = \begin{bmatrix} v_i(i) \\ \iota_i(i) \end{bmatrix}.
\]

Let $\hat{v}_i^*(i) = [\hat{v}_i^T(i), \iota_i^*(i)]^T$ denote a feasible $\hat{v}_i(i)$ to problem (37). Then, a feasible solution $v_i^*(i)$ to problem (36) can be immediately recovered by setting $v_i^*(i) = \hat{v}_i^*(i) / \iota_i^*(i)$. A feasible solution $\Theta_i^*(i)$ to problem (35) can thus be expressed as $\Theta_i^*(i) = \text{diag}(v_i^*(i))$.

To solve problem (37), a natural method is to formulate it as a semidefinite programming (SDP) problem by matrix lifting [16], [50]. Since $\hat{v}_i^H(i) R_{k,t}(i) \hat{v}_i(i) = \text{Tr}(R_{k,t}(i) \hat{v}_i(i) \hat{v}_i^H(i))$, we denote $V_i(i)$ as the lifting matrix of $\hat{v}_i(i)$, where $V_i(i) = \hat{v}_i(i) \hat{v}_i^H(i)$ and rank $(V_i(i)) = 1$. Therefore, problem $P_{2.1}$ can be further transformed into the following problem $P_{2.3}$.

\begin{align}
\text{find } & V_i(i) \\
\text{subject to } & \text{Tr}(R_{k,t}(i) V_i(i)) + |h_k^d_{k,t}(i)|^2 \\
& \geq \eta_i(i) D^2 c^2_{k,t}(i) / P_0 \quad \forall i, k, t, \\
& V_i(i)(n, n) = 1 \quad \forall n \in [N + 1], \\
& \text{rank}(V_i(i)) = 1 \quad \forall i, t, \\
& V_i(i) \geq 0 \quad \forall i, t
\end{align}

where $V_i(i)(n, n)$ denotes the $(n, n)$-entry of matrix $V_i(i)$. However, the resulting problem $P_{2.3}$ is still nonconvex due to the rank-one constraints in (38d). Fortunately, a feasible solution to this problem can be generated by simply dropping the rank-one constraints via the SDR technique [50]. The resulting SDP problem can be solved efficiently by existing convex optimization solvers such as CVX [51]. Although the Gaussian randomization technique in SDR may generate
Algorithm 1: Two-Step Alternating Minimization for Solving Problem $\mathcal{P}$ (27) in the RIS-Enabled FL System

**Input:** The phase-shift matrix $\Theta_0$, the privacy level $(\epsilon, \delta)$, and the maximum iteration number $J$.

1. Initialize $j \leftarrow 1$ to denote the number of iterations.
2. **repeat**
   1. Given $\Theta_{j-1}$, obtain solution $(\eta_j, \sigma_{k,j})$ by solving problem $\mathcal{P}_1$ (28).
   2. Given $(\eta_j, \sigma_{k,j})$, obtain solution $\Theta_j$ by solving problem $\mathcal{P}_2$ (35).
   3. Update $j \leftarrow j + 1$.
3. **until** $(2/\theta_j)$ is below a given threshold or $j > J$;

**Output:** $\eta_j(i) \leftarrow \eta_j$, $\Theta_j(i) \leftarrow \Theta_j$, $\sigma_{k,j}(i) \leftarrow \sigma_{k,j}$.

A suboptimal solution, we can still observe that a significant learning performance gain of the RIS-enabled FL system can be achieved through numerical experiments.

Without causing confusion, we omit parameter $i$ for notational ease, e.g., $\Theta_i(i)$ as $\Theta_i$. We present the two-step iterative framework in Algorithm 1 for solving problem $\mathcal{P}$. The computational cost of the proposed algorithm consists of solving a sequence of search programs for $\mathcal{P}_1$ and SDR programs [50] for $\mathcal{P}_2$. From the complexity analysis of a typical interior-point method, the worst case complexity of Algorithm 1 is obtained as $\mathcal{O}(JIT \max[K, N+1]^4 \sqrt{N+1} \log(1/o))$, where $o > 0$ denotes the solution accuracy of SDR.

C. Discussion on the Causal Approximation

Based on the above observations, Algorithm 1 solves the nonconvex problem $\mathcal{P}$ with the knowledge of all channel responses $\{h_{d,k,j}(i), m_j(i), h_{k,j}(i)\}$ and local gradient information $\{\zeta_{k,j}(i), \gamma_j\}$, yielding an impractical noncausal system. We thus present some feasible prediction methods to approximate this noncausal system.

At the beginning of the training process, we need to address the following two questions for our proposed scheme.

1. How to approximate the future gradient information?
2. How to predict the future channel responses?

The first question aims at providing upper bounds for the $l_2$-sensitivity defined in (23) and the local gradients. To this end, there exist multiple adaptive methods for practical use, such as the L-Lipschitz-based technique [52], [53], and the adaptive clipping method [54] with a well-designed critical threshold based on the current gradients.

For the second question, this is essentially related to wireless channel prediction. To estimate channels accurately, many promising methods have been proposed. For instance, Liu and Simeone [34] specialized an accumulated mechanism where the future channel responses are assumed to be consistent with the current ones. But this assumption fails to exploit the correlation among channel blocks. To explore the inherent relationship between each communication block, several prediction techniques have been developed in [55] and [56], including the Gauss–Markov process channel modeling method [57] and Wiener or Kalman filtering-based approach [57], [58]. Besides, the deep learning methods turn out to be powerful to improve channel prediction accuracy [59]–[61]. We thus leave the design of effective causal models as our future work.

V. SIMULATION RESULTS

In this section, we present the simulation results to demonstrate the advantages of the RIS-enabled FL systems. Besides, the effectiveness of our proposed two-step alternating minimization framework will also be illustrated. Simulations are conducted using MATLAB R2021b and the code is available at https://github.com/MengCongWo/FL_Privacy_blockcrossing.

A. Simulation Settings

We propose to gain insights into the impact of deploying RIS on learning accuracy based on ridge regression, whose samplewise loss function is given as

$$f(x, y; \theta) = \frac{1}{2} \left\| \theta^T x - y \right\|^2 + \nu \left\| \theta \right\|^2$$

where $\nu = 5 \times 10^{-5}$ denotes the penalty coefficient. We randomly generate a data set of scale $|D| = 10^8$ and set the model dimension $d$ to be 10. Specifically, the training samples $x$ are drawn i.i.d. according to $N(0, I_d)$ while the corresponding true label $y$ is given by $y = x(2) + 3x(5) + 0.2z_0$, where the data noise $z_0$ is drawn i.i.d. from $N(0, 1)$. We uniformly divide the total data set $D$ into $K$ local data sets, and let the ratio $r = \max_{k \in [K]} |D_k| / |D| \in [0.1, 1)$ (40) denote the system heterogeneity representing various storage capacities of the edge devices and larger $r$ yields higher heterogeneity. The global loss function $F$ is $\mu$-strongly convex, $L$-Lipschitz smooth, and differentiable, where $\mu$ and $L$ are specified by the smallest and largest eigenvalues of the data Gramian matrix $X^T X / |D| + 2 \mu I_d$ with the data matrix $X = [x_1, \ldots, x_D]$. Besides, the optimal $\theta^*$ of (39) is $\theta^* = (X^T X + 2 \mu I_d)^{-1} X^T y$ with true label vector $y = [y_1, \ldots, y_D]^T$. As for the definitions of $\gamma_j$ and $\zeta_{k,j}(i)$, we use the simple upper bounds as in [34, Sec. V-A]. Furthermore, we use the normalized optimality gap defined as $[F(\theta_{T+1}) - F(\theta^*)] / F(\theta^*)$ to measure the learning accuracy.

The wireless channels are assumed to suffer from Rice fading [24], and the channel coefficients are given by

$$\theta = \sqrt{\frac{\kappa}{1 + \kappa}} \theta_{\text{LoS}} + \sqrt{\frac{1}{1 + \kappa}} \theta_{\text{NLoS}}$$

where $\kappa$ represents the Rician factor, $\theta_{\text{LoS}}$ denotes the deterministic line-of-sight (LoS) component, and $\theta_{\text{NLoS}}$ denotes the non-line-of-sight (NLoS) component. For simplicity, we set $\theta_{\text{LoS}} = 1$ and generate $\theta_{\text{NLoS}}$ by the autoregressive (AR) scheme, which is defined as

$$\theta_{\text{NLoS}}(i) = \rho \theta_{\text{NLoS}}(i - 1) + \sqrt{1 - \rho^2} \theta(i)$$

where $\rho$ denotes the correlation coefficient and $\theta(i)$ is drawn based on an innovation process satisfying $\theta(i) \sim N(0, I)$. Hence, the channel coefficients in the $i$th communication block
of learning round $t$ are given by $m_i(t) = \theta_{IB}, h_{k,i}^t(i) = \theta_{DI}$, and $h_{k,i}^t(i) = \theta_{DB}$, whose Rician factors are given by $\kappa_{IB}, \kappa_{DI}$, and $\kappa_{DB}$, respectively. The parameter $\rho$ is set to 1 for simplicity since $\rho$ has no discernible effect on the performance in the case where perfect CSI is available for all edge devices.

Under the above simulation settings, we mainly compare the learning accuracy under the following three RIS schemes.

1) **RIS-Enabled FL System With DP**: In this scheme, Algorithm 1 is applied to solve problem $\mathcal{P}$.

2) **RIS-Enabled FL System Without DP**: In this scheme, we remove the privacy constraints for the baseline without DP, i.e., only the second term in (32) exists.

3) **FL System Without RIS**: In this scheme, we set $\Theta_i(i) = \emptyset$ for the baseline without RIS but with DP [34].

The system parameters used in the simulations are summarized in Table II for reference. We also simulate a practical IoT-based affective computing FL scenario [62] with high-dimensional and nonconvex settings, which will be presented in Section V-C.

### Table II

| Parameter | Value |
|-----------|-------|
| $K$       | 10    |
| $D_k$     | 1000  |
| $\kappa_{IB}$ | 5 |
| $\epsilon$ | 20 |
| $N$       | 30    |
| $r$       | 0.1   |
| $T$       | 30    |
| $\kappa_{DI}$ | 0 |
| $\delta$  | 0.01  |
| $\kappa_{DB}$ | 5 |
| $\rho$    | 1     |
| $\kappa_{IB}$ | 5 |
| $\kappa_{DI}$ | 0 |
| $\kappa_{DB}$ | 5 |
| $\epsilon$ | 20 |
| $\delta$  | 0.01  |

### B. Learning Accuracy Under Different Conditions

To investigate the improvements of learning accuracy achieved by RIS, we compare the learning accuracy of the three RIS schemes based on the ridge regression model from five aspects: privacy levels $\epsilon$, SNR levels, the number of learning rounds $T$, the impact of system heterogeneity $r$, and the number of RIS elements $N$. We also simulate a high-dimensional setting based on the CIFAR-10 data set and a nonconvex setting based on the MNIST data set.

We first illustrate the relationship between learning accuracy and privacy level $\epsilon$ in Fig. 5. As an extension of Fig. 3, we demonstrate that with the relaxation of the privacy level, i.e., for larger values of $\epsilon$, the RIS-enabled FL system achieves a higher learning performance gain compared to the case without RIS which is restricted by unfavorable wireless channel propagations. Instead, RIS has no impact on the learning performance when the privacy level is strict, i.e., the privacy field presented in Theorem 3. Besides, comparing the learning accuracy of the RIS-enabled FL system with DP and the case without DP, we note that the privacy guarantee is achieved at the expense of losing learning accuracy, i.e., higher privacy (smaller value of $\epsilon$) indicates lower learning accuracy.

In Fig. 6, we focus on the learning accuracy of the RIS-enabled FL systems in the noisy wireless environment, i.e., various levels of SNR. It shows that with the increase of SNR, the DP-restricted schemes achieve the same learning accuracy, i.e., around $4 \times 10^{-3}$, under the same privacy and power constraints. Besides, we note that the RIS-enabled FL systems achieve a notable improvement in learning accuracy compared to the FL systems without RIS, especially when SNR is low. This indicates that the RIS-enabled FL systems are more adaptable to noisy wireless channels. Furthermore, Fig. 6 shows that the RIS-enabled FL system with DP achieves close accuracy with the case without DP when SNR < 20. This indicates that privacy guarantee can be ensured freely [34].

The impact brought by the number of learning rounds $T$ or total communication blocks on the learning accuracy is illustrated in Fig. 7, where the oscillation comes from the newly introduced wireless noise in each communication block. We observe that without considering the effect of noise, the number of learning rounds $T$ has no evident effect on the accuracy after running enough number of training rounds for gradient descent due to the adaptive selection of $\beta$ and $t_i$ performed in Theorem 3. To make it precise, $\beta$ can flexibly adjust the selection of $t_i(i)$ to ensure that it satisfies all constraints with the increase of $T$. Moreover, the RIS-enabled FL systems achieve higher accuracy and better robustness to the newly added wireless channel noise.

Fig. 8 shows the learning accuracy of three RIS schemes among various degrees of system heterogeneity $r$. For simplicity, one edge device is assigned more data points while the rest are evenly distributed to the remaining edge devices [34]. Simulation results show the negative influence brought by the
Fig. 7. Learning accuracy of three RIS schemes under different numbers of learning rounds.

Fig. 8. Learning accuracy of three RIS schemes under various degrees of heterogeneity.

Fig. 9. Learning accuracy of the RIS-enabled FL system with DP under different privacy levels $\epsilon$.

Fig. 10. Test accuracy of the three RIS schemes with the MNIST data set.

Fig. 11. Relative test accuracy of the three RIS schemes with the CIFAR-10 data set.

high-heterogenous system on accuracy. Besides, compared to the FL systems without RIS, the RIS-enabled FL systems are more robust due to the interaction between channel condition and system heterogeneity in (32). Furthermore, Fig. 8 serves as a promising result for the design of anti-heterogeneous FL systems by deploying RIS to weaken heterogeneity.

Fig. 9 illustrates the impacts of the number of reflecting elements $N$ at RIS on the learning accuracy under three fields presented in Theorem 3. From Fig. 9, we verify that the learning accuracy cannot be significantly improved when the privacy level is extremely strict (privacy field), while the benefits of deploying RIS emerge when the privacy level is relaxing. Besides, we note that with the increase of $N$, the reconfigurable capability of RIS enhances, yielding high learning accuracy. However, due to the existence of privacy constraints, this improvement of learning accuracy may slow down when $N$ becomes large.

Figs. 10 and 11 show the learning accuracy of three RIS schemes on the MNIST and CIFAR-10 data sets, respectively. Specifically, we train a logistic regression model on the CIFAR-10 data set for high-dimensional settings and a neural network on MNIST for nonconvex settings. The neural network model parameter $\theta$ with dimension $d = 79510$ is comprised of the first-layer parameter $W_1 \in \mathbb{R}^{100 \times 785}$ and the second-layer parameter $W_2 \in \mathbb{R}^{10 \times 101}$. For a high-privacy and noisy environment, we reset the privacy level $\epsilon$ to be 10, the SNR level to be 5 dB, the number of communication blocks in one learning round $I$ to be 10, and the Rice factor of $\varrho_{DB}$ to be 0 (the Rayleigh channel). Figs. 10 and 11 indicate that the RIS-enabled FL system with DP enjoys higher accuracy and better robustness to the wireless channel noise.

The test accuracy of the three RIS schemes with the MNIST data set.

The relative test accuracy of the three RIS schemes with the CIFAR-10 data set.
TABLE III
CLASSIFICATION ACCURACY OF VARIOUS FL SYSTEMS FOR STRESS DETECTION DATA SET UNDER DIFFERENT PRIVACY AND SNR LEVELS

|                                   | High privacy, Low SNR \((e = 10, SNR = 3\, dB)\) | High privacy, High SNR \((e = 10, SNR = 10\, dB)\) | Low privacy, Low SNR \((e = 80, SNR = 3\, dB)\) | Low privacy, High SNR \((e = 80, SNR = 10\, dB)\) |
|-----------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| RIS-enabled FL system with DP     | 0.7308                                          | 0.7308                                          | 0.7620                                          | 0.7716                                          |
| RIS-enabled FL system without DP  | 0.7692                                          | 0.7716                                          | 0.7692                                          | 0.7716                                          |
| FL system without RIS             | 0.6827                                          | 0.7260                                          | 0.6851                                          | 0.7428                                          |

compared to the FL system without RIS. Besides, the accuracy of the RIS-enabled FL system is close to the case without DP. Furthermore, we can observe that our proposed methods can achieve good performance even for nonconvex learning tasks.

C. Differentially Private FL for Wearable IoT-Based Biomedical Monitoring

Driven by the massive amounts of biomedical data generated from widespread IoT edge devices, affective computing has become a field of interest in biomedical informatics [63]. Affective computing is a promising technique that recognizes a person’s emotional state based on the physiological signals and images, e.g., photoplethysmogram (PPG)-based heart activity, to detect the link between physical health and emotional states. To integrate distributed health data analysis and privacy protection, Can and Ersoy [62] established a practical IoT-based wearable biomedical monitoring scenario and provided a first attempt on the application of privacy-preserving FL to IoT-based affective computing. Specifically, 32 public school teachers from Turkey participated in the IoT-based stress detection experiment, where each teacher was required to wear wrist-worn wearable IoT devices and experience three different emotional sessions (baseline, lecture, and exam). After each session, the heart activity features collected by wearable IoT devices and the mental state report from each teacher were gathered to construct the stress detection data set. Besides, the collected data are distributed unevenly across \(K = 26\) private edge devices, which are orchestrated by an edge server to collaboratively train a global stress detection model. Inspired by this, we simulate a practical wireless IoT-based stress detection FL scenario with DP guarantees to test the learning performance of three RIS schemes.

We train a neural network on the stress detection data set whose model parameter \(\theta\) with dimension \(d = 12,002\) consists of the first-layer parameter \(W_1 \in \mathbb{R}^{100 \times 17}\), the second-layer parameter \(W_2 \in \mathbb{R}^{100 \times 101}\), and the third-layer parameter \(W_3 \in \mathbb{R}^{2 \times 101}\). The classification accuracy on the stress detection data set under different levels of privacy and SNR is presented in Table III. It can be observed that the RIS-enabled FL system with DP is capable of outperforming the FL system without RIS in terms of classification accuracy and robustness to wireless noise. The RIS-enabled system with DP can also achieve similar performance compared with the system without DP.

VI. CONCLUSION

In this article, we developed an RIS-enable differentially private FL system by leveraging the reconfigurability of channel propagation via RIS and the property of waveform superposition via AirComp. We theoretically characterized the convergence behavior of the over-the-air FL algorithm, for which a system optimization problem was established to achieve better learning accuracy under the privacy and transmit power constraints. We further proposed a two-step low-rank optimization framework to minimize the learning optimality gap, by jointly optimizing the power allocation, artificial noise, and reflecting coefficients of RIS during the learning procedure. Through convergence analysis and system optimization, we revealed that the RIS-enabled FL system is able to achieve higher system SNR and boost the receive signal power, thereby improving learning accuracy performance while satisfying the privacy requirements. Numerical results also demonstrated that the proposed FL system can achieve higher learning accuracy and privacy than the benchmarks.

APPENDIX A
PROOF OF THEOREM 1

For each edge device \(k\), we note that the local gradient \(g_{k,t}\) is divided into \(I\) \(e\)-dimensional subsignals. Based on (14), we arrive at the equivalent channel noise vector

\[
\hat{w}_t = \left[\frac{w_{1}^T(1)}{\sqrt{\eta_i(1)}}, \ldots, \frac{w_{I}^T(I)}{\sqrt{\eta_i(I)}}\right]^T.
\]

Then, the estimated \(\hat{g}_t\) at the edge server can be expressed as

\[
\hat{g}_t = g_t + \frac{1}{KD} \sum_{k=1}^{K} \text{Re}[n_{k,t}] + \frac{1}{KD} \text{Re}[\hat{w}_t].
\]

Hence, under Assumption 2, we have the following inequality:

\[
F(\theta_{t+1}) \leq F(\theta_t) - \lambda g_t^T \hat{g}_t + \frac{L_2}{2} \|g_t\|^2 - \lambda \left|\sum_{k=1}^{K} \text{Re}[n_{k,t}]\right| - \frac{\lambda}{KD} \text{Re}[g_t]^T \hat{w}_t + \frac{\lambda}{KD} \text{Re}[g_t]^T \hat{w}_t.
\]

Recalling that \(n_{k,t} = [n_{k,t}^T(1), \ldots, n_{k,t}^T(I)]^T\) with \(n_{k,t}(i) \sim \mathcal{CN}(0, \sigma_{k,t}^2(i) I_d)\) and given \(\lambda = 1/L\), by taking the expectations over the additive noise (including the artificial and wireless channel noise) on both sides of the above inequality, we obtain

\[
\mathbb{E}[F(\theta_{t+1})] \leq F(\theta_t) + \left(\frac{L}{2} - \lambda\right) \|g_t\|^2 + \frac{L_2}{2} \times \frac{1}{(KD)^2} \mathbb{E}\left[\sum_{k=1}^{K} \text{Re}\left([n_{k,t}^T(1), \ldots, n_{k,t}^T(I)]^T\right) + \text{Re}[\hat{w}_t]\right]^2,
\]

\[
\leq F(\theta_t) - \frac{1}{2L} \|g_t\|^2 + \frac{e}{4L(KD)^2} \left(\sum_{k=1}^{K} \sum_{i=1}^{I} \sigma_{k,t}^2(i) + \sum_{i=1}^{I} N_0 \eta_i(i)\right).
\]
According to (16), by setting $\theta = \theta^*$ and $\theta' = \theta_i$, we have
\begin{equation}
\frac{1}{2}||\nabla F(\theta_i)||^2 \geq \mu\left[F(\theta_i) - F^*\right].
\end{equation}
Therefore, subtracting the optimal value $F^*$ at both sides yields
\begin{equation}
\mathbb{E}[F(\theta_{i+1}) - F^*] \leq \left(1 - \frac{\mu}{L}\right)[F(\theta_i) - F^*] + \frac{e}{4L(KD)^2}\left(\sum_{k=1}^{K} \sum_{i=1}^{I} \sigma_{k,i}^2(i) + \sum_{i=1}^{I} N_0 / \eta(i)\right).
\end{equation}
Finally, the expected result is obtained by applying (46) iteratively through $T$ learning rounds.

**APPENDIX B**

**PROOF OF THEOREM 2**

The proving process is based on the advanced literature [30], [34], [40]. We focus on the privacy constraint of the $k$th edge device based on the aggregated signals $r = [r_i]_{i=1}^{T}$ at the edge server. It is worth noting that without the traditional lightweight assumption, our transmission model faces a more practical scenario where one learning round consists of $I$ communication blocks.

According to (8), the privacy loss after $T$ learning rounds can be expressed as
\begin{equation}
\mathcal{L}_{D_k}D'_k(r) = \ln\left(\prod_{i=1}^{T} \mathbb{P}\left[r_i | r_{i-1}, \ldots, r_1; D_k\right]\right)
\end{equation}
\begin{equation}
= \sum_{i=1}^{T} \ln\left(\frac{\mathbb{P}\left[r_i | r_{i-1}, \ldots, r_1; D'_k\right]}{\mathbb{P}\left[r_i | r_{i-1}, \ldots, r_1; D_k\right]}\right).
\end{equation}

The effective noise $q_i$ in (20) is a complex Gaussian random vector with statistically independent elements. Besides, recalling that the noise vector $w_i(i) \sim \mathcal{CN}(0, N_0 I_d)$ and $n_{k,i}(i) \sim \mathcal{CN}(0, \sigma_{k,i}^2(i) I_d)$, we obtain the pseudo-covariance matrix of $q_i$, which is given by
\begin{equation}
J_i = \mathbb{E}\left[q_i(q_i)^T\right] = 0
\end{equation}
and the covariance matrix of $q_i$ can be expressed as
\begin{equation}
\Sigma_i = \mathbb{E}\left[q_i(q_i)^H\right] = \begin{bmatrix}
\Lambda_i(1) & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \Lambda_i(I)
\end{bmatrix}
\end{equation}
where
\begin{equation}
\Lambda_i(i) = \left(\eta(i) \sum_{k=1}^{K} \sigma_{k,i}^2(i) + N_0\right) I_d \quad \forall i \in [I]
\end{equation}
denotes the diagonal positive semidefinite covariance matrix of $q_i(i)$, which indicates that $\Sigma_i$ is a diagonal positive semidefinite matrix.

According to [64, Appendix A], it is easy to verify that $q_i$ is a complex random vector with circular symmetry property, whose probability density function (pdf) is given by
\begin{equation}
p(q_i) = \frac{1}{\pi^d \det(\Sigma_i)} \exp\left(-q_i^H \Sigma_i^{-1} q_i\right).
\end{equation}
Therefore, for the analysis of sensitivity, we further denote the effective noise based on data set $D_k$ as $z_{k,i} = [z'_{k,i}(1), \ldots, z'_{k,i}(I)]^T$ with
\begin{equation}
z_{k,i}(i) = r_i(i) - \sqrt{\eta(i)} d_k g_{k,i}(i; D_k) - v_{k,i}(i)
\end{equation}
where $v_{k,i}(i)$ denotes the second constant term in (19) and $g_{k,i}(i; D_k)$ represents the updated gradient based on $D_k$; thus, we can obtain
\begin{equation}
\mathbb{P}\left[r_i | r_{i-1}, \ldots, r_1; D_k\right] = \frac{1}{\pi^d \det(\Sigma_i)} \exp\left(-z_{k,i}^H \Sigma_i^{-1} z_{k,i}\right).
\end{equation}
Furthermore, according to (22), $z_{k,i} + u_{k,i}$ represents the effective noise based on data set $D'_k$. Hence, substituting them into (47), we have
\begin{equation}
\mathcal{L}_{D_k}D'_k(r)
\end{equation}
\begin{equation}
= \sum_{i=1}^{T} \mathbb{E}\left[\left(z_{k,i} + u_{k,i}\right)^H \Sigma_i^{-1} (z_{k,i} + u_{k,i}) - (z_{k,i})^H \Sigma_i^{-1} z_{k,i}\right]
\end{equation}
\begin{equation}
= \sum_{i=1}^{T} \left[2\text{Re}\left(u_{k,i}^H \Sigma_i^{-1} z_{k,i}\right) + u_{k,i}^H \Sigma_i^{-1} u_{k,i}\right].
\end{equation}
Based on (9), given the DP parameter pair $(\epsilon, \delta)$, the privacy violation probability bound can be expressed as
\begin{equation}
\mathbb{P}\left(\left|\mathcal{L}_{D_k}D'_k(r)\right| > \epsilon\right)
\end{equation}
\begin{equation}
\leq \mathbb{P}\left(\left|\text{Re}\left(u_{k,i}^H \Sigma_i^{-1} z_{k,i}\right)\right| > \frac{\epsilon}{2} - \frac{1}{2} \sum_{i=1}^{T} u_{k,i}^H \Sigma_i^{-1} u_{k,i}\right)
\end{equation}
\begin{equation}
= 2\mathbb{P}\left(\sum_{i=1}^{T} \text{Re}\left(u_{k,i}^H \Sigma_i^{-1} z_{k,i}\right) > \frac{\epsilon}{2} - \frac{1}{2} \sum_{i=1}^{T} u_{k,i}^H \Sigma_i^{-1} u_{k,i}\right)
\end{equation}
\begin{equation}
\leq 2\mathbb{P}(\gamma_k > c_k)
\end{equation}
where $(a)$ comes from the inequality $\mathbb{P}(|X + | > \epsilon) \leq \mathbb{P}(|X| > \epsilon - \epsilon)$ for $\epsilon \geq 0$. We note that $z_{k,i} \sim \mathcal{CN}(0, \Sigma_i)$ and $\Sigma_i$ is a diagonal matrix. Hence, it is easy to verify that $\Sigma_i^{-1/2} z_{k,i} \sim \mathcal{CN}(0, I_d)$, which leads to $u_{k,i}^H \Sigma_i^{-1} z_{k,i} \sim \mathcal{CN}(0, u_{k,i}^H \Sigma_i^{-1} u_{k,i})$. As a result, considering the fact that any $z_{k,i}$ is statistical independent of each other, $\gamma_k$ is a random variable according to $\mathcal{N}(0, (1/2) \sum_{i=1}^{T} u_{k,i}^H \Sigma_i^{-1} u_{k,i})$.

According to (24) and by denoting
\begin{equation}
\xi_i \leq \eta(i) \sum_{k=1}^{K} \sigma_{k,i}^2(i) + N_0 \quad \forall i, t
\end{equation}
as a lower bound of the power of effective noise, we have the following inequalities:
\begin{equation}
v_k^2 \leq \sum_{i=1}^{T} \left(2\eta_i^2 \max_{i \in [I]}(\sigma_{k,i}^2(i))\right) \leq \sum_{i=1}^{T} \left(\frac{\eta_i^2}{2\xi_i}\right) \geq \frac{1}{2} \sum_{i=1}^{T} u_{k,i}^H \Sigma_i^{-1} u_{k,i},
\end{equation}
\begin{equation}
\tilde{c}_k \leq \frac{\epsilon}{2} - \sum_{i=1}^{T} \frac{2\eta_i^2}{\xi_i} \max_{i \in [I]}(\eta(i)) \leq \frac{\epsilon}{2} - \sum_{i=1}^{T} \left(\frac{\eta_i^2}{2\xi_i}\right) \leq c_k
\end{equation}
where $(a)$ and $(c)$ come from the upper bound in (24), and $(b)$ is obtained by (49). Based on this, we introduce another
random variable $\mathcal{X}$ according to $\mathcal{C}(0, \nu^2)$. Thus, we arrive at the following inequalities:

$$\Pr(\check{\psi}_k > c_k) \leq \Pr(\mathcal{X} > c_k) \leq \Pr(\mathcal{X} > \check{c}_k).$$

(52)

Finally, according to Mill’s inequality, we can obtain

$$\Pr(\check{\psi}_k > c_k) \leq \frac{\mathbb{E}_k}{\pi \hat{c}_k} \exp \left(-\frac{\check{c}^2_k}{2\hat{c}^2_k}\right) < \delta$$

followed by solving an equation based on the monotonicity property of function $\mathcal{C}(x) = \sqrt{\pi} xe^{x^2}$, then we arrive at the expected result in (25).

**APPENDIX C**

**PROOF OF THEOREM 3**

To start, by using the change of variables, we define the following new relationships:

$$a'_{k,t} = \sigma_{k,t}^2(i), \quad b'_t = \eta_t^{-1}(i), \quad c'_t = \xi_t \bar{b}'_t = \xi_t \eta_t^{-1}(i).$$

(53)

Thus, it is easy to verify that $a'_{k,t} \geq 0$ and $b'_t, c'_t > 0$. Besides, we also denote

$$d_t = \frac{\xi_t}{\max_i(\eta_t(i))} = \xi_t \min_i(b'_t) \quad \forall t$$

(54)

which yields an inherent constraint $d_t \leq c'_t \quad \forall i, t$.

For short, we omit the superscript and subscript when describing the optimization variables, e.g., $a'_{k,t}$ as $a$. Therefore, problem $\mathcal{P}_1$ can be transformed into problem $\mathcal{P}_{1,1}$

$$\min_{a, b, c, d} \sum_{t=1}^T \left( 1 - \frac{\mu}{L} \right)^{-1} \left( \sum_{k=1}^K \sum_{i=1}^l a'_{k,t} + N_0 \sum_{i=1}^l b'_t \right)$$

subject to

$$\sum_{i=1}^l 4\nu_t^2 \frac{d_t}{d_t} - R_{dp} \leq 0$$

(55a)

$$\sum_{i=1}^l c'_t - \sum_{k=1}^K a'_{k,t} - N_0 b'_t \leq 0 \quad \forall i, t$$

(55b)

$$D^2_{\xi'}^2(i) + \bar{e} a'_{k,t} \leq P_0|\bar{h}_k(i)|^2 b'_t \quad \forall i, k, t$$

(55c)

$$d_t - c'_t \leq 0 \quad \forall i, t$$

(55d)

$$a'_{k,t} \geq 0, b'_t \geq 0, c'_t \geq 0, d_t \geq 0 \quad \forall i, k, t$$

(55e)

which is a pure convex problem and can be tackled by KKT conditions. Above all, the Lagrange function is given as

$$\mathcal{L} = \sum_{t=1}^T \left( 1 - \frac{\mu}{L} \right)^{-1} \left( \sum_{k=1}^K \sum_{i=1}^l a'_{k,t} + N_0 \sum_{i=1}^l b'_t \right)$$

$$+ \beta \left( \sum_{i=1}^l 4\nu_t^2 \frac{d_t}{d_t} - R_{dp} \right) - \sum_{k=1}^K \sum_{i=1}^l \psi_{k,t}^{i} a'_{k,t}$$

$$+ \sum_{i=1}^l \sum_{t=1}^T c'_t \left( c'_t - \sum_{k=1}^K a'_{k,t} - N_0 b'_t \right) + \sum_{i=1}^l \sum_{t=1}^T u'_t(d_t - c'_t)$$

$$+ \sum_{k=1}^K \sum_{i=1}^l \sum_{t=1}^T \psi_{k,t}^{i} \left( D^2_{\xi'}^2(i) + \bar{e} a'_{k,t} - P_0|\bar{h}_k(i)|^2 b'_t \right)$$

where $\beta, c'_t, u'_t, \psi_{k,t}^{i}, \psi_{k,t}^{i}, \geq 0$ represent the Lagrange multipliers and for simplicity, we omit some obviously zero terms. We denote variables with \(\sim\) as the optimal solutions satisfying KKT conditions, e.g., $\tilde{a}_{k,t}$ denotes the optimal $a'_{k,t}$.

Hence, we can obtain the following relationships:

$$\frac{\partial \mathcal{L}}{\partial d_t} = \left(1 - \frac{\mu}{L}\right)^{-1} - t'_t + e\psi_{k,t}^{i} - \psi_{k,t}^{i} = 0$$

(56a)

$$\frac{\partial \mathcal{L}}{\partial b'_t} = N_0 \left(1 - \frac{\mu}{L}\right)^{-1} - N_0 t'_t - P_0 \sum_{k=1}^K \psi_{k,t}^{i} |\bar{h}_k(i)|^2 = 0$$

(56b)

$$\frac{\partial \mathcal{L}}{\partial c'_t} = t'_t - u'_t = 0$$

(56c)

$$\frac{\partial \mathcal{L}}{\partial a'_{k,t}} = -\beta 4\nu_t^2 \frac{d_t}{d_t} + \sum_{i=1}^l a'_t = 0$$

(56d)

which indicate that at least one of $\tilde{t}'_t$ and $\tilde{\psi}_{k,t}^{i}$ is greater than 0. Based on this, we mainly focus on the following three conditions and elaborate their optimality.

**Power Field:** On the one hand, we first focus on a simple case where the inequality in (55b) strictly holds, i.e., the privacy constraint is not dominant. Hence, according to complementary slackness condition, we immediately get

$$\tilde{\beta} = \tilde{u}'_t = \tilde{c}'_t = 0 \quad \forall i, t.$$ 

(57)

Furthermore, we have $\tilde{\psi}_{k,t}^{i} > 0$ based on (56a) which also implies $\tilde{a}_{k,t} = 0$. Hence, the power constraint (55d) becomes the unique one that restricts the learning accuracy. Based on this, we arrive at the unique optimal solution in this case, i.e.,

$$\tilde{b}'_t = \frac{1}{P_0 \max_{i,k} |\bar{h}_k(i)|^2} D^2_{\xi'}^2(i)$$

(58)

On the other hand, we turn to the difficult multisolusion case where the privacy constraint (55b) is stringent with $\tilde{\beta} \neq 0$, which results in the following equality:

$$\sum_{i=1}^l 4\nu_t^2 \frac{d_t}{d_t} - R_{dp} = 0$$

(59)

Motivated by the previous condition, we set $\tilde{a}_{k,t} = 0$, which saves the communication and power resources of the edge server. Hence, we arrive at the following two conditions.

**Privacy Field:** We first focus on the case where only the privacy constraint takes effect, which indicates (55d) strictly holds, i.e.,

$$\tilde{b}'_t > \frac{1}{P_0 \max_{i,k} |\bar{h}_k(i)|^2} D^2_{\xi'}^2(i)$$

(60)

yielding $\tilde{\psi}_{k,t}^{i} = 0$. Hence, based on (56b) and (56c), we obtain

$$\tilde{a}'_t = \tilde{c}'_t = (1 - \mu/L)^{-1} \quad \forall i, t.$$ 

(61)

Besides, combining (56d) and the complementary slackness condition, we further get

$$\tilde{d}_t = \tilde{c}'_t = N_0 \tilde{b}'_t = \frac{2\nu_t}{\sqrt{\tilde{\beta}}} \left(1 - \frac{\mu}{L}\right)^{\frac{1}{2}}$$

(62)

which indicates $\tilde{b}'_t$ remains constant during one learning round, and for short, we denote $\tilde{b}'_t = \tilde{b}' \quad \forall i$. In summary, the optimal conditions of $\tilde{b}'_t$ are given by (59), (60), and (62).
Privacy-Power Field: Now, we turn to analyze the case where the dual constraints of privacy and power take effect, i.e., there exist some \( \tilde{\tau}^i \neq \tilde{d}^i \) which lead to \( \tilde{\mu}^i = \gamma^i = 0 \). Besides, from (56b), we can verify that there exist \( \psi^i_{k,t} \neq 0 \) based on the power constraints of all edge devices. Remarkably, \( \tilde{b}^i \) remains constant across all edge devices in the \( \tau \)th communication block of learning round \( t \), so it must meet all power constraints and strictly satisfy a certain one, i.e.,

\[
\tilde{b}^i = \frac{1}{P_0} \max_{k} D^2 \psi^i_{k,t}(t) \| h^i_{k,t}(t) \|^2. \tag{63}
\]

As for the other \( \tilde{\mu}^i \neq 0 \), we consider it as a generalized version of (61), i.e., some communication blocks are restricted by privacy while the others are power. For analytical ease, we introduce parameter \( \tau_r \) to represent the number of communication blocks limited by privacy in the \( \tau \)th learning round, i.e., \( \tilde{\mu}^i_{k,t} \neq 0 \). Hence, (56d) can be reformulated as

\[
\tau_r \left( 1 - \frac{\mu}{L} \right) = \frac{4 \gamma^i}{\beta} \tilde{\tau}^i. \tag{64}
\]

For instance, if all communication blocks are restricted by privacy, \( \tau_r \) is set to \( I \) which is the same as (62), i.e., Privacy field. Therefore, considering the above two cases, we arrive at the expected result by elaborately selecting \( \tau_r \) in each learning round \( t \) and parameter \( \tilde{\beta} \) to strictly meet (59).

Finally, according to (53), we obtain the expected result.

References

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, “The roadmap to 6G: AI empowered wireless networks,” IEEE Commun. Mag., vol. 57, no. 8, pp. 84–90, Aug. 2019.

[2] B. Mcmahon, E. Moore, D. Ramage, S. Hampson, and B. A. Y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. Int. Conf. Artif. Intell. Stat. (AISTATS), vol. 54, 2017, pp. 1273–1282.

[3] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, “Edge artificial intelligence for 6G: Vision, enabling technologies, and applications,” IEEE J. Sel. Areas Commun., vol. 40, no. 1, pp. 5–36, Jan. 2022.

[4] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, “Communication-efficient edge AI: Algorithms and systems,” IEEE Commun. Surveys Tuts., vol. 22, no. 4, pp. 2167–2191, 4th Quart., 2020.

[5] D. C. Nguyen, M. Ding, P. N. Pathirana, J. Seneviratne, J. Li, and H. V. Poor, “Federated learning for Internet of Things: A comprehensive survey,” IEEE Commun. Surveys Tuts., vol. 23, no. 3, pp. 1622–1658, 3rd Quart., 2021.

[6] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amin, “A survey on federated learning for resource-constrained IoT devices,” IEEE Internet Things J., vol. 9, no. 1, pp. 1–24, Jan. 2021, doi: 10.1109/JIOT.2020.3095077.

[7] J. Wang et al., “A field guide to federated optimization,” 2021, arXiv:2107.06917.

[8] N. Rieke et al., “The future of digital health with federated learning,”IEEE J. Sel. Areas Commun., vol. 37, no. 3, pp. 50–60, May 2019.

[9] M. H. Poor, “Federated learning for Internet of Things: A comprehensive survey,” IEEE Commun. Surveys Tuts., vol. 23, no. 3, pp. 1273–1282.

[10] M. M. 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.

[11] 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, Aug. 2021.

[12] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, “Federated machine learning for intelligent IoT via reconﬁgurable intelligent surface,”IEEE Netw., vol. 34, no. 5, pp. 16–22, Sep./Oct. 2020.

[13] Z. Wang et al., “Federated learning via intelligent reﬂecting surface,” IEEE Trans. Wireless Commun., vol. 21, no. 2, pp. 808–822, Feb. 2022.

[14] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, “Reconﬁgurable intelligent surfaces for federated learning: A uniﬁed communication-learning design approach,” IEEE Trans. Wireless Commun., vol. 20, no. 11, pp. 7595–7609, Nov. 2021, doi: 10.1109/TWC.2021.3086116.

[15] X. Yuan, Y.-J. A. Zhang, Y. Shi, W. Yan, and H. Liu, “Reconﬁgurable-intelligent-surface empowered wireless communications: Challenges and opportunities,” IEEE Wireless Commun., vol. 28, no. 2, pp. 136–143, Apr. 2021.

[16] Q. Wu and R. Zhang, “Intelligent reﬂecting surface enhanced wireless network via joint active and passive beamforming,” IEEE Trans. Wireless Commun., vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[17] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, “Reconﬁgurable intelligent surfaces for energy efﬁciency in wireless communication,”IEEE Trans. Wireless Commun., vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[18] M. D. Renzo et al., “Smart radio environments empowered by reconﬁgurable intelligent surfaces: How it works, state of research, and the road ahead,”IEEE J. Sel. Areas Commun., vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[19] B. Hitaj, G. Atieniese, and F. Perez-Cruz, “Deep models under the GAN: Information leakage from collaborative deep learning,” in Proc. ACM SIGSAC Conf. Comput. Commun. Security, 2017, pp. 603–618.

[20] L. Zhu and S. Han, “Deep leakage from gradients,” in Federated Learning, Cham, Switzerland: Springer, 2020, pp. 17–31.

[21] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, “Inverting the model of federated learning—How easy is it to break privacy in federated learning?”, 2020, arXiv:2003.14053.

[22] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,”Found. Trends Theor. Comput. Sci., vol. 9, nos. 3–4, pp. 211–407, 2014.

[23] A. M. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. T. Suresh, “Shufﬂed model of federated learning: Privacy, communication and accuracy trade-offs,”IEEE J. Sel. Areas Commun. Inf. Theory, vol. 2, no. 1, pp. 464–478, Mar. 2021.

[24] N. Wu, F. Farokhi, D. Smith, and M. A. Kaafar, “The value of collaboration in convex machine learning with differential privacy,” in Proc. IEEE Security Privacy (SSP), 2020, pp. 304–317.

[25] N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. McMahan, “cpsGD: Communication-efﬁcient and differentially-private distributed SGD;,” in Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Assoc., 2018, pp. 7564–7575.

[26] D. Liu and O. Simeone, “Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control,”IEEE J. Sel. Areas Commun., vol. 39, no. 1, pp. 170–185, Jan. 2021.

[27] B. Hasrscouglu and D. Gündüz, “Private wireless federated learning with anonymous over-the-air computation,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2021, pp. 5195–5199.

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