

# Privacy Enhancement in Over-the-Air Federated Learning via Adaptive Receive Scaling

Faeze Moradi Kalarde\*, Ben Liang\*, Min Dong<sup>†</sup>, Yahia A. Eldemerdash Ahmed<sup>‡</sup>, and Ho Ting Cheng<sup>‡</sup>

\*Department of Electrical and Computer Engineering, University of Toronto, Canada

<sup>†</sup>Department of Electrical, Computer and Software Engineering, Ontario Tech University, Canada

<sup>‡</sup>Ericsson Canada, Canada

**Abstract**—In Federated Learning (FL) with over-the-air aggregation, the quality of the signal received at the server critically depends on the receive scaling factors. While a larger scaling factor can reduce the effective noise power and improve training performance, it also compromises the privacy of devices by reducing uncertainty. In this work, we aim to adaptively design the receive scaling factors across training rounds to balance the trade-off between training convergence and privacy in an FL system under dynamic channel conditions. We formulate a stochastic optimization problem that minimizes the overall Rényi differential privacy (RDP) leakage over the entire training process, subject to a long-term constraint that ensures convergence of the global loss function. Our problem depends on unknown future information, and we observe that standard Lyapunov optimization is not applicable. Thus, we develop a new online algorithm, termed AdaScale, based on a sequence of novel per-round problems that can be solved efficiently. We further derive upper bounds on the dynamic regret and constraint violation of AdaScale, establishing that it achieves diminishing dynamic regret in terms of time-averaged RDP leakage while ensuring convergence of FL training to a stationary point. Numerical experiments on canonical classification tasks show that our approach effectively reduces RDP and DP leakages compared with state-of-the-art benchmarks without compromising learning performance.

## I. INTRODUCTION

Federated Learning (FL) leverages the computational capabilities of edge devices by allowing them to collaboratively train a global model on their local data without requiring data to be shared [1]. To enable efficient uplink transmission from devices to the server, over-the-air (OTA) aggregation via analog transmission has emerged as an effective solution [2]–[10]. In each training round of OTA FL, the devices simultaneously transmit their local signals using analog modulation over a shared multiple access channel, enabling natural model aggregation through signal superposition.

Nevertheless, OTA computation is susceptible to aggregation errors introduced by receiver noise and channel distortion. To mitigate these errors, in each training round of OTA FL, each device scales its local signal by a transmit weight, while the server applies a scaling factor to the received signal. The design of the receive scaling factors over training rounds significantly affects the quality of the aggregated signal and thus influences training convergence.

Another concern in FL is privacy leakage, as the local signals sent from the devices reveal information about their underlying data [11]–[15]. To reduce data privacy risks, differential privacy (DP) [16] is commonly employed in FL. In the standard DP framework, each device clips its per-sample gradients and adds artificial noise to the batch-averaged gradients before transmission. However, in OTA FL, adding artificial noise is not necessary, as the inherent receiver noise serves as privacy noise and can provide the desired level of privacy [17]–[23]. Nevertheless, the receive scaling factors determine the effective noise power at the server and, consequently, the level of privacy leakage. Thus, it is essential to design the receive scaling factors to balance the trade-off between training convergence and privacy.

There are two main challenges in designing the receive scaling factors. First, while privacy leakage occurs in each training round, the overall leakage over all rounds is our ultimate concern. Existing works either ignore the overall leakage [17]–[19], or use the Advanced Composition Theorem for DP over all rounds [20]–[23], which is known to be a loose approximation, especially over a large number of rounds [16], [24], [25]. Second, the receive scaling decisions are coupled over the training rounds by the overall privacy leakage and the FL convergence objectives, while the future communication channel state is usually unknown. This necessitates an *online* algorithm to design receive scaling over time. Existing solutions either depend on simplified assumptions about channel conditions [20]–[22] or are heuristic-based [23], lacking theoretical performance guarantees to assess how closely they approximate the optimal solution (see Section II).

In this work, we aim to adaptively design the receive scaling factors for an OTA FL system under time-varying channel conditions. We address the aforementioned challenges, first, by employing the Rényi differential privacy (RDP) framework [24], which allows a simple additive form for the overall privacy leakage, and second, by designing an effective online algorithm that is shown to provide strong performance guarantees with respect to the offline optimum. Our contributions are as follows:

- We formulate an optimization problem whose objective is to minimize the time-averaged RDP leakage of devices over the entire training process after an arbitrary number of rounds  $T$ . The problem formulation includes a constraint to ensure model convergence to a stationary

point of the global loss, along with individual transmit power constraints for each device. We further derive a sufficient condition for convergence and reformulate the problem using this condition as a surrogate convergence constraint.

- The reformulated problem involves both a long-term objective and a long-term constraint, making it difficult to solve due to the lack of knowledge about future channel conditions. Standard Lyapunov optimization techniques are not applicable, as the long-term constraint is unbounded over the feasible set. Instead, we develop a novel online algorithm termed AdaScale, which decomposes the original problem into convex per-round optimization problems that can be solved efficiently using bisection search.
- We further establish an upper bound for the dynamic regret of AdaScale, with respect to an offline optimum that assumes all future information is available. We demonstrate that the proposed method achieves  $\mathcal{O}(T^{\max\{1-\beta, \frac{1-\beta}{2}\}})$  dynamic regret and  $\mathcal{O}(T^{\frac{\beta-1}{2}})$  constraint violation, where  $\beta$  is a tunable parameter. Furthermore, when  $1 < \beta < 2$ , the regret bound diminishes to zero, and FL training converges to a stationary point, as  $T \rightarrow \infty$ .
- We conduct numerical experiments on canonical classification datasets under typical wireless network setting. Our results show that AdaScale is nearly optimal and outperforms state-of-the-art alternatives, effectively reducing both RDP and DP leakages under the same training convergence level.

## II. RELATED WORK

Among existing works on DP in OTA FL, [17]–[19] consider only per-round privacy leakage. They cannot provide proper trade-off between the *overall* privacy leakage and training performance. More recent works evaluate the overall privacy leakage throughout the training process. Among them, [25], [26] analyze privacy leakage only and do not design the receive scaling factors. In contrast, [20]–[23] focus on designing the receive scaling factors to enhance training performance while imposing constraints on the overall privacy leakage.

In [20], an optimal offline solution is obtained when the future channel conditions are known. Otherwise, estimation of the future channel is used to update the offline solution over the training rounds. This work is extended in [21], to consider a reconfigurable intelligent surface (RIS), and in [22], to consider a multi-antenna server. These works provide essentially offline solutions, while our objective is *online* adaptation to the time-varying channels over time.

The work in [23] is the closest to ours. It extends the problem formulation of [21] for active RIS and proposes an online solution. The standard Lyapunov optimization framework is employed to formulate per-round problems, which are solved using an alternating optimization heuristic. However, since the per-round problems are not solved within a bounded optimality gap, and the objective function is not bounded over the feasible

set, the Lyapunov approach does not offer any performance guarantee [27]. In comparison, we propose a novel online solution in AdaScale that is proven to achieve diminishing regret with respect to the offline optimum, while guaranteeing that FL training converges to a stationary point.

Finally, in all aforementioned works, the overall privacy leakage is evaluated using the Advanced Composition Theorem for DP [16], which is known to be loose and can lead to inefficient designs [24], [25]. In this work, we address this limitation through a tighter analysis based on the RDP.

## III. PRELIMINARIES

### A. FL System

We consider a wireless FL system comprising a central server and  $M$  edge devices. Each device, indexed by  $m$ , contains a local training dataset  $\mathcal{D}_m = \{(\mathbf{u}_{m,i}, o_{m,i}) : 1 \leq i \leq n_m\}$ , where  $\mathbf{u}_{m,i}$  is the  $i$ -th data feature vector, and  $o_{m,i}$  is its label. The local data of device  $m$  follows distribution  $p_m$ . The local loss function of device  $m$  is defined as  $f_m(\mathbf{w}) \triangleq \mathbb{E}_{(\mathbf{u}_{m,i}, o_{m,i}) \sim p_m} c_m l(\mathbf{w}; (\mathbf{u}_{m,i}, o_{m,i}))$ , where  $l(\cdot)$  represents a sample-wise loss function,  $c_m \in \mathbb{R}$  is the device loss weight, and  $\mathbf{w} \in \mathbb{R}^d$  contains the model parameters. The edge devices aim to train a global model on the server cooperatively. This requires minimizing a global loss function defined as  $f(\mathbf{w}) = \frac{1}{M} \sum_{m=1}^M f_m(\mathbf{w})$ . The ultimate goal is to determine the optimal model,  $\mathbf{w}^*$ , that minimizes  $f(\mathbf{w})$  in a distributed manner.

In this study, we adopt the conventional Federated Stochastic Gradient Descent (FedSGD) technique [1] for iterative model training in FL. We consider OTA aggregation for uplink transmission from the devices to the server. The FedSGD algorithm with OTA aggregation is described in Section IV-A.

### B. Differential Privacy

Two widely adopted notions of Differential Privacy (DP) in the literature are  $(\epsilon, \delta)$ -DP and  $(\alpha, \epsilon)$ -RDP.

**Definition 1** ( $(\epsilon, \delta)$ -DP [16]). *The randomized mechanism  $M : \mathcal{D} \rightarrow \mathcal{R}$  with domain  $\mathcal{D}$  and range  $\mathcal{R}$  satisfies  $(\epsilon, \delta)$ -DP if, for any two neighboring datasets  $S \in \mathcal{D}$  and  $S' \in \mathcal{D}$ , i.e.,  $S'$  is formed by adding or removing a single element from  $S$ , and for any output set  $\mathcal{R}' \subseteq \mathcal{R}$ ,*

$$\Pr[M(S) \in \mathcal{R}'] \leq e^\epsilon \Pr[M(S') \in \mathcal{R}'] + \delta. \quad (1)$$

**Definition 2** ( $(\alpha, \epsilon)$ -RDP [24]). *The randomized mechanism  $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$  satisfies  $(\alpha, \epsilon)$ -RDP for  $\alpha \in \mathbb{R}$ ,  $\alpha > 1$  if for any neighboring datasets  $S \in \mathcal{D}$  and  $S' \in \mathcal{D}$ , it holds that*

$$D_\alpha(\mathcal{M}(S) \parallel \mathcal{M}(S')) \leq \epsilon, \quad (2)$$

where  $D_\alpha(p_1 \parallel p_2)$  denotes the Rényi divergence of order  $\alpha$  between distributions  $p_1(x)$  and  $p_2(x)$ :

$$D_\alpha(p_1 \parallel p_2) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim p_2} \left[ \left( \frac{p_1(x)}{p_2(x)} \right)^\alpha \right]. \quad (3)$$

**Remark 1** (Conversion from RDP to DP [24]). *If a randomized mechanism  $\mathcal{M}$  satisfies  $(\alpha, \varepsilon_1)$ -RDP for some  $\alpha > 1$ , then for any  $\delta \in (0, 1)$ , it also satisfies  $(\varepsilon_2, \delta)$ -DP, where*

$$\varepsilon_2 = \varepsilon_1 + \log\left(\frac{\alpha - 1}{\alpha}\right) - \frac{\log \delta + \log \alpha}{\alpha - 1}. \quad (4)$$

Next, we present a method to compute the RDP leakage.

**Definition 3** (Sampled Gaussian Mechanism (SGM) [28]). *Let  $u$  be a function mapping subsets of  $\mathcal{D}$  to  $\mathbb{R}^d$ . We define the Sampled Gaussian Mechanism parameterized with the sampling rate  $0 < q \leq 1$  and the noise  $\sigma > 0$  as*

$$SG_{q,\sigma}(\mathcal{D}) \triangleq u(\mathcal{S}) + \mathcal{N}(0, \sigma^2 \mathbb{I}^d), \quad (5)$$

where  $\mathcal{S} = \{x: x \in \mathcal{D} \text{ is sampled with probability } q\}$  is formed by sampling each element of  $\mathcal{D}$  independently at random with probability  $q$  without replacement, and  $\mathcal{N}(0, \sigma^2 \mathbb{I}^d)$  is spherical  $d$ -dimensional Gaussian noise with per-coordinate variance  $\sigma^2$ .

**Definition 4** ( $\ell_2$ -sensitivity). *Let  $u$  be a function with domain  $\mathcal{D}$  and range  $\mathcal{R}$ . The  $\ell_2$ -sensitivity of  $u$  is  $\Delta$  if for any two neighboring datasets  $\mathcal{S} \in \mathcal{D}$  and  $\mathcal{S}' \in \mathcal{D}$ , it holds that*

$$\|u(\mathcal{S}) - u(\mathcal{S}')\|_2 \leq \Delta. \quad (6)$$

**Lemma 1** (RDP leakage of SGM). *For any integer  $\alpha > 1$ , the SGM defined in Definition 3, with mapping  $u(\cdot)$  having  $\ell_2$ -sensitivity  $\Delta$ , satisfies  $(\alpha, \rho_\alpha(q, \sigma_{\text{eff}}))$ -RDP, where  $\sigma_{\text{eff}} \triangleq \frac{\sigma}{\Delta}$  is the effective noise multiplier,  $\rho_\alpha(q, \sigma_{\text{eff}}) \triangleq \frac{A_\alpha(q, \sigma_{\text{eff}})}{\alpha - 1}$ , and*

$$A_\alpha(q, \sigma_{\text{eff}}) \triangleq \ln \left[ \sum_{k=0}^{\alpha} \binom{\alpha}{k} (1-q)^{\alpha-k} q^k \exp\left(\frac{k^2 - k}{2\sigma_{\text{eff}}^2}\right) \right]. \quad (7)$$

*Proof.* The result can be directly derived from [28], and is omitted here for brevity.  $\square$

#### IV. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we summarize the FedSGD [1] algorithm with OTA uplink transmission, describe how we calculate its overall RDP leakage, and formulate our constrained RDP leakage minimization problem.

##### A. FedSGD with OTA Aggregation

At each training round of FedSGD, the server updates the global model based on signals received from the devices. Specifically, round  $t$  involves the following steps:

- 1) **Model broadcast:** The server broadcasts the model parameter vector  $\mathbf{w}_t$  to all devices. As commonly considered in the literature, we assume each device perfectly recovers the model.
- 2) **Local gradient computation:** Each device  $m$  forms a batch  $\mathcal{B}_{m,t}$  according to Poisson sampling.<sup>1</sup> Specifically, each data point is sampled independently with probability  $q_m = \frac{B_m}{n_m}$  from its local dataset  $\mathcal{D}_m$ , where  $B_m$  is

<sup>1</sup>FedSGD is not specific to any sampling method, but we will see later that Poisson sampling is needed for tractable RDP analysis.

the expected batch size. The devices then compute the average of the sample gradients over the batch:

$$\mathbf{g}_{m,t} = \frac{1}{B_m} \sum_{i \in \mathcal{B}_{m,t}} \mathbf{g}_{m,t,i}, \quad (8)$$

where  $\mathbf{g}_{m,t,i} \triangleq c_m \nabla l(\mathbf{w}_t, (\mathbf{u}_{m,i}, o_{m,i})) \in \mathbb{R}^d$ .

- 3) **OTA uplink transmission:** The devices transmit  $\mathbf{g}_{m,t}$  to the server via OTA aggregation [2]. Specifically, all devices select a transmit weight  $a_{m,t} \in \mathbb{C}$  and send  $a_{m,t} \mathbf{g}_{m,t}$  to the server simultaneously using the same frequency resource over  $d$  consecutive time slots.
- 4) **Receiver processing and model update at server:** Denote the channel coefficient of device  $m$  by  $h_{m,t} \in \mathbb{C}$ . The received signal at the server is

$$\mathbf{r}_t = \sum_{m=1}^M h_{m,t} a_{m,t} \mathbf{g}_{m,t} + \mathbf{n}_t, \quad (9)$$

where  $\mathbf{n}_t \sim \mathcal{CN}(0, \sigma_n^2 \mathbb{I}^d)$  is the receiver noise. The server scales the received signal and updates the model by applying one-step gradient descent as<sup>2</sup>

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \lambda \frac{\text{Re}(\mathbf{r}_t)}{\sqrt{\eta_t}}, \quad (10)$$

where  $\eta_t \in \mathbb{R}^+$  is the receive scaling factor,  $\lambda$  is the learning rate, and  $\text{Re}(\cdot)$  returns the real part of a complex variable.

As in [20], [21], [23], we assume that the sample gradient norms are upper bounded by  $G$ , i.e.,  $\|\mathbf{g}_{m,t,i}\| \leq G, \forall m, t, i$ , and we set the device transmit weights proportional to the inverse of the uplink channels. Thus,  $a_{m,t} = \frac{\sqrt{\eta_t}}{M h_{m,t}}, \forall m, \forall t$ . Then, we can rewrite the server processed received signal in (10) as

$$\tilde{\mathbf{r}}_t \triangleq \frac{\text{Re}(\mathbf{r}_t)}{\sqrt{\eta_t}} = \underbrace{\frac{1}{M} \sum_{m=1}^M \mathbf{g}_{m,t}}_{\triangleq \mathbf{s}_t} + \underbrace{\frac{\text{Re}(\mathbf{n}_t)}{\sqrt{\eta_t}}}_{\triangleq \tilde{\mathbf{n}}_t}, \quad (11)$$

which contains two parts: i) the signal  $\mathbf{s}_t$ , and ii) the effective noise at the receiver  $\tilde{\mathbf{n}}_t \sim \mathcal{N}(\mathbf{0}, \frac{\sigma_n^2}{2\eta_t} \mathbb{I}^d)$ .

The average transmit power of device  $m$  in round  $t$  is

$$P_{m,t} = |a_{m,t}|^2 \frac{\mathbb{E}\|\mathbf{g}_{m,t}\|^2}{d} = \frac{\eta_t \mathbb{E}\|\mathbf{g}_{m,t}\|^2}{d M^2 |h_{m,t}|^2} \stackrel{(a)}{\leq} \frac{\eta_t G^2}{d M^2 |h_{m,t}|^2} \mathbb{E}\left[\frac{|\mathcal{B}_{m,t}|^2}{B_m^2}\right] \stackrel{(b)}{=} \frac{\eta_t G^2 k_m^2}{d M^2 |h_{m,t}|^2}, \quad (12)$$

where (a) follows from the fact that based on (8),  $\mathbf{g}_{m,t}$  is the average of sample gradients with norms less than or equal to  $G$ , and thus, by the triangle inequality, we have  $\|\mathbf{g}_{m,t}\| \leq \frac{|\mathcal{B}_{m,t}| G}{B_m}$ ; and (b) follows from  $k_m^2 \triangleq \mathbb{E}\left[\frac{|\mathcal{B}_{m,t}|^2}{B_m^2}\right] = 1 + \frac{(1-q_m)}{B_m}$  due to Poisson sampling.

<sup>2</sup>For more efficient transmission,  $\mathbf{g}_{m,t}$  can be sent via complex signals using both the real and imaginary parts of the signal. This will not change the fundamental process developed subsequently.

## B. RDP Leakage Calculation

Privacy leakage quantifies the information about the device local data samples that the server can extract from the post-processed received signal  $\tilde{\mathbf{r}}_t$ .<sup>3</sup>

1) *Per-Round RDP Leakage*: By comparing (11) with Definition 3, it is evident that for each device  $m$ , the vector  $\tilde{\mathbf{r}}_t$  constitutes an SGM with respect to the local dataset  $\mathcal{D}_m$ . Thus, RDP leakage for each device can be quantified using Lemma 1, by identifying the effective noise multiplier associated with  $\tilde{\mathbf{r}}_t$  for each device. By Definition 4, the  $\ell_2$  sensitivity of  $\mathbf{s}_t$  with respect to the batch of device  $m$  is  $\Delta_{m,t} = \frac{G}{B_m M}$ , since the norm of each sample gradient is upper bounded by  $G$ , and the aggregation of sample gradients is divided by  $M B_m$  according to (8) and (11). Now, given  $\Delta_{m,t}$ , the effective noise multiplier for device  $m$  is

$$\sigma_{m,t} = \frac{\frac{\sigma_n}{\sqrt{2\eta_t}}}{\Delta_{m,t}} = \frac{M B_m \sigma_n}{\sqrt{2\eta_t} G}. \quad (13)$$

Based on Lemma 1, with  $\sigma_{m,t}$  in hand, for any order  $\alpha$ , the RDP leakage for device  $m$  in round  $t$  is  $\rho_\alpha(q_m, \sigma_{m,t})$ .

2) *Overall RDP Leakage*: The RDP leakage of a sequence of randomized mechanisms composed sequentially is given by the sum of the RDP leakages of the individual mechanisms [24]. Thus, the overall RDP leakage over  $T$  rounds for device  $m$  is  $\sum_{t=0}^{T-1} \rho_\alpha(q_m, \sigma_{m,t})$ . As discussed in Section I, this is a main advantage of using RDP instead of DP in our problem formulation.

## C. Problem Formulation

We aim to minimize the overall RDP leakage after  $T$  training rounds, via optimizing the receive scaling factors  $\{\eta_t\}_{t=0}^{T-1}$ , while ensuring a certain level of convergence of the global model:

$$\min_{\{\eta_t\}} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^M \rho_\alpha(q_m, \sigma_{m,t}) \quad (14a)$$

$$\text{s.t.} \quad \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\mathbf{w}_t)\|^2 \leq \gamma, \quad (14b)$$

$$\frac{\eta_t G^2 k_m^2}{dM^2 |h_{m,t}|^2} \leq P_{\max}, \forall m, \forall t, \quad (14c)$$

$$\eta_t > 0, \forall t, \quad (14d)$$

where the  $\mathbb{E}[\cdot]$  is on the randomness of the batch sampling, the noise of sample gradients, and the receiver noise. Constraint (14b) ensures that the system achieves  $\gamma$ -convergence to a stationary point of the global loss function  $f(\mathbf{w})$ , and constraint (14c) limits the average power consumption of devices.

**Remark 2.** We note that bounding  $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\mathbf{w}_t)\|^2$  in (14b) implies a bound on  $\min_{0 \leq t \leq T-1} \mathbb{E} \|\nabla f(\mathbf{w}_t)\|^2$ , which guarantees that at least a model among  $\{\mathbf{w}_t\}$  during the training process will be sufficiently close to a stationary point if  $\gamma$  is chosen to be small enough.

<sup>3</sup>Note that the imaginary part of  $\mathbf{r}_t$  is pure noise, containing no information.

Solving (14) presents significant challenges because constraint (14b) involves the gradient of the global loss function, which is not an explicit function of the optimization variables. Additionally, the global loss function is typically not quantifiable, as it depends on the local data distributions  $\{p_m\}$ , which are unknown. Furthermore, the effective noise multipliers  $\{\sigma_{m,t}\}$  in (14a) and the model sequence  $\{\mathbf{w}_t\}$  in (14b) depend on the channel conditions at each round, which are unknown prior to the start of the round. This necessitates the development of an online solution to address unknown future information. To proceed, we first analyze the convergence of FedSGD with OTA aggregation and substitute (14b) with a more manageable surrogate constraint.

## V. ADAPTIVE RECEIVE SCALAR DESIGN

In this section, we first reformulate problem (14) through the training convergence analysis. We then present an online algorithm to adaptively design the receiver scaling factors  $\{\eta_t\}_{t=0}^{T-1}$  to address the trade-off between privacy and training convergence.

### A. Problem Reformulation via Training Convergence Analysis

Convergence analysis for FedSGD under uniform batch sampling with ideal communication is provided in [29]. Here, we extend this analysis to account for Poisson sampling and OTA aggregation transmission. We then use the resulting convergence bound to reformulate problem (14).

1) *Convergence Analysis*: We consider the following assumptions on the loss function, which are common in the literature of distributed training and first-order optimization [30], [31]:

**A1. Smoothness:**  $\forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ ,

$$f_m(\mathbf{w}) \leq f_m(\mathbf{w}') + \langle \nabla f_m(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle + \frac{L}{2} \|\mathbf{w} - \mathbf{w}'\|^2.$$

**A2. Global minimum:**  $\exists \mathbf{w}^* \in \mathbb{R}^d$ , such that  $\forall \mathbf{w} \in \mathbb{R}^d$ ,

$$f(\mathbf{w}^*) = f^* \leq f(\mathbf{w}). \quad (15)$$

**A3. Unbiased sample gradients with bounded variance:**  $\exists A_1, A_2 \geq 0$ , such that  $\forall \mathbf{w}_t \in \mathbb{R}^d$ ,

$$\mathbf{g}_{m,t,i} = \nabla f_m(\mathbf{w}_t) + \mathbf{z}_{m,t,i}, \quad \mathbb{E}[\mathbf{z}_{m,t,i} | \mathbf{w}_t] = 0 \quad (16)$$

$$\mathbb{E}[\|\mathbf{z}_{m,t,i}\|^2 | \mathbf{w}_t] \leq A_1 \|\nabla f_m(\mathbf{w}_t)\|^2 + A_2. \quad (17)$$

**A4. Bounded similarity:**  $\exists C_1, C_2 \geq 0$ , such that  $\forall \mathbf{w} \in \mathbb{R}^d$ ,

$$\frac{1}{M} \sum_{m=1}^M \|\nabla f_m(\mathbf{w}) - \nabla f(\mathbf{w})\|^2 \leq C_1 \|\nabla f(\mathbf{w})\|^2 + C_2. \quad (18)$$

In the following, we provide our convergence bound for the global loss function under FedSGD with OTA aggregation.

**Theorem 1** (Training convergence). *Assume A1-A4 hold, and the learning rate is set as  $\lambda \leq \frac{1}{4L(C_1+1)(A_1+1)}$ . After  $T$  rounds of FedSGD described in Section IV-A, we have*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\mathbf{w}_t)\|^2 \leq \phi + \frac{L\lambda}{2T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{\eta_t}, \quad (19)$$

where  $\phi$  is defined as

$$\phi \triangleq \frac{2(f(\mathbf{w}_0) - f^*)}{\lambda T} + 2L\lambda \left( 2C_2(A_1 + 1) + A_2 \right). \quad (20)$$

Due to the page limit, we omit its proof. Our bound differs slightly from those in prior works [20], [21], [23], as it accounts for Poisson sampling and is derived under weaker assumptions. Specifically, unlike previous bounds that assume strong convexity or the Polyak-Łojasiewicz condition, our analysis does not require convexity of the loss function.

2) *Problem Reformulation:* To reformulate problem (14), we apply a change of variable and define

$$x_t \triangleq \frac{\eta_t}{h_{\min,t}^2}, \quad (21)$$

where  $h_{\min,t} \triangleq \min_m \frac{|h_{m,t}|}{k_m}$ . We further define  $x_{\max} \triangleq \frac{P_{\max} dM^2}{G^2}$ . Then, constraints (14c) and (14d) convert to

$$0 < x_t \leq x_{\max}, \forall t. \quad (22)$$

Moreover, the effective noise multiplier in (13) can be written in terms of  $x_t$  as

$$\sigma_{m,t} = \frac{MB_m \sigma_n}{\sqrt{2x_t G h_{\min,t}}}. \quad (23)$$

To deal with constraint (14b), first we rewrite the bound in (19) in terms of  $x_t$  as follows:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\mathbf{w}_t)\|^2 &\leq \phi + \frac{L\lambda}{2T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2 x_{\max}} \\ &\quad + \frac{L\lambda}{2T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t} - \frac{1}{x_{\max}} \right). \end{aligned} \quad (24)$$

We note that the second term of the upper bound in (24) does not depend on the decision variables  $\{x_t\}$ . For simplicity, we define these terms as

$$\phi' \triangleq \phi + \frac{L\lambda}{2T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2 x_{\max}}. \quad (25)$$

We replace the left-hand side (LHS) of (14b) with its upper bound given in (24). To ensure that  $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{w}_t)\|^2$  is bounded by  $\gamma$ , it suffices to bound the right-hand side (RHS) of (24) by the same amount. Since the first two terms of RHS of (24) are constant, restricting it by  $\gamma$  implies a bound on the third term  $\frac{1}{T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t} - \frac{1}{x_{\max}} \right)$  by  $\nu$ , where  $\nu = \frac{2(\gamma - \phi')}{\lambda L}$ . Hence, we reformulate problem (14) as

$$\min_{\{x_t\}} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^M \rho_\alpha(q_m, \sigma_{m,t}) \quad (26a)$$

$$\text{s.t.} \quad \frac{1}{T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t} - \frac{1}{x_{\max}} \right) \leq \nu, \quad (26b)$$

$$0 < x_t \leq x_{\max}, \forall t, \quad (26c)$$

where  $\nu$  replaces  $\gamma$  as the hyperparameter to tune the trade-off between training convergence and privacy.

The above problem is still difficult to handle due to the presence of the long-term objective and constraint, and the channel coefficients  $\{h_{m,t}\}$  are unknown prior to the start of the  $t$ -th round. Next, we propose a novel algorithm to solve the problem in an online manner and provide bounds for both its constraint violation and its dynamic regret.

### B. Proposed Algorithm

We start with a conventional virtual queue to keep track of the violation of constraint (26b), which is denoted by  $Q_t \in \mathbb{R}$  with  $Q_0 = 0$ . In each round  $t$ , the server updates the virtual queue as

$$Q_{t+1} = \max \left\{ Q_t + \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t} - \frac{1}{x_{\max}} \right) - \nu, 0 \right\}. \quad (27)$$

If we directly apply standard Lyapunov optimization [27] to solve problem (26), the decision variable at each round would be obtained by solving a per-round optimization problem with objective  $V \sum_{m=1}^M \rho_\alpha(q_m, \sigma_{m,t}) + Q_t \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t} - \frac{1}{x_{\max}} \right)$ . Minimizing this objective is equivalent to minimizing an upper bound on the drift-plus-penalty, if the constraint function is *bounded* within the feasible set [27]. However, this boundedness assumption does not hold for problem (26), as the LHS of constraint (26b) can grow arbitrarily large when  $x_t$  approaches zero. In fact, it is easy to see that directly applying the standard Lyapunov method leads to infinite constraint violation.

This motivates us to modify the standard Lyapunov method by introducing an *additional term* into the per-round objective. This modification prevents the solution from collapsing to zero and avoids unbounded constraint violations. As we will show in Section VI, the inclusion of this additional term makes the per-round optimization problem equivalent to minimizing an upper bound on the drift-plus-penalty, even in the presence of unbounded constraints. This enables us to establish performance guarantees.

Specifically, we consider a different form of the per-round optimization as follows. In round  $t$ , the server solves an optimization problem to design its receiver scaling factor as

$$\begin{aligned} \min_{x_t} \quad & V \sum_{m=1}^M \rho_\alpha(q_m, \sigma_{m,t}) + Q_t \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t} - \frac{1}{x_{\max}} \right) \\ & + \frac{1}{2} \left( \frac{d\sigma_n^2}{h_{\min,t}^2} \right)^2 \left( \frac{1}{x_t} - \frac{1}{x_{\max}} \right)^2 \end{aligned} \quad (28a)$$

$$\text{s.t.} \quad 0 < x_t \leq x_{\max}. \quad (28b)$$

where  $V \in \mathbb{R}^+$  is a predefined constant. Note that  $\rho(q_m, \sigma_{m,t})$  depends on  $x_t$  through (23).

Problem (28) is a single-variable optimization problem. The following proposition establishes that it is convex for integer values of  $\alpha$ .

**Proposition 1.** *For any integer  $\alpha$ , problem (28) is convex.*

*Proof.* Since the constraint (28b) is linear in  $x_t$ , it suffices to show that the objective function in (28a) is convex in  $x_t$ . This can be verified by observing that all three terms of this

---

**Algorithm 1** AdaScale at round  $t$ 

---

**Inputs:**  $\sigma_n, \{q_m\}, \{B_m\}, G, M, d, P_{\max}$ .**Output:**  $\eta_t$ 

- 1: Server solves (28) using bisection search.
  - 2: Server updates its virtual queue based on (27).
  - 3: Server sets  $\eta_t = x_t \min_m \frac{|h_{m,t}|^2}{k_m^2}$ .
  - 4: Server transmits  $\eta_t$  to the devices; devices use it to set their transmit weights.
- 

function are convex in  $x_t$ . The details require only elementary algebraic analysis and are omitted due to the page limit.  $\square$

Based on Proposition 1, for any integer  $\alpha$ , problem (28) can be solved by setting the derivative of the objective function to zero and identifying its root. If the root lies within the interval  $(0, x_{\max}]$ , it corresponds to the optimal solution; otherwise, the optimal solution is given by  $x_{\max}$ . However, due to the complexity of the objective function, finding a closed-form expression for the root is not feasible. Therefore, we employ the bisection algorithm, based on the derivative of the objective function, to numerically compute the point at which the derivative of the objective function equals zero. The detailed procedure is standard and is omitted to avoid redundancy.

We refer to our proposed algorithm, which adaptively designs the receive scaling factor by solving (28) and updating the virtual queue based on (27), as Adaptive receive Scaling (AdaScale). It is summarized in Algorithm 1.

**Remark 3.** *Throughout this work, we consider integer values of  $\alpha$ . Nevertheless, since the RDP leakage is a monotonically increasing function of  $\alpha$ , upper and lower bounds on the leakage for a non-integer  $\alpha$  can be obtained by evaluating the RDP expression at the closest integers.*

### C. Computational Complexity

Solving (28) using the bisection algorithm requires evaluating the derivative of (28a), which has a constant computational cost of  $\mathcal{O}(1)$ . To reach a solution within a distance of  $\tau$  from the optimum, the algorithm requires at most  $\log_2(\frac{x_{\max}}{\tau})$  iterations. Therefore, in each training round, the overall computational complexity for obtaining a solution within  $\tau$ -vicinity of the optimum is  $\mathcal{O}(\log_2(\frac{x_{\max}}{\tau}))$ .

Despite the low computational complexity of this algorithm, we next show that it has strong performance guarantees, in terms of constraint violation and dynamic regret.

## VI. THEORETICAL PERFORMANCE ANALYSIS

We analyze the performance of AdaScale in this section. We note that even though our analysis uses the familiar notion of drift, it is substantially different from the conventional Lyapunov stability analysis, and it leads to novel constraint violation and dynamic regret bounds. To begin, let  $\hat{x}_t$  denote the optimization variable at round  $t$  obtained using AdaScale, and let  $\hat{\sigma}_{m,t}$  denote the corresponding effective noise multiplier, obtained by substituting  $\hat{x}_t$  for  $x_t$  in (23).

### A. Upper Bound on $R$ -Slot Drift

For any positive integer  $R \leq T$ , we define the  $R$ -slot drift of the virtual queue as

$$\Delta_R(t) \triangleq \frac{1}{2}Q_{t+R}^2 - \frac{1}{2}Q_t^2. \quad (29)$$

Using (29) and noting that the initial value of the queue is set to zero, we can rewrite the  $R$ -slot drift at time  $t = 0$  as  $\Delta_R(0) = \frac{1}{2}Q_R^2$ , which implies

$$Q_R = \sqrt{2\Delta_R(0)}. \quad (30)$$

We start with an upper bound on the one-slot drift in the lemma below.

**Lemma 2** (One-slot drift bound). *The one-slot drift for AdaScale is upper bounded by*

$$\begin{aligned} \Delta_1(t) &\leq Q_t \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right) \\ &\quad + \frac{1}{2} \left( \frac{d\sigma_n^2}{h_{\min,t}^2} \right)^2 \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right)^2 + \frac{1}{2} \nu^2. \end{aligned} \quad (31)$$

*Proof.* The proof is similar to one in standard Lyapunov optimization, by squaring both sides of an inequality obtained from the virtual queue updating equation. The details are omitted due to the page limit.  $\square$

We sum both sides of (31) over  $t$  from 0 to  $R-1$  to obtain

$$\begin{aligned} \Delta_R(0) &\leq \sum_{t=0}^{R-1} Q_t \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right) \\ &\quad + \frac{1}{2} \sum_{t=0}^{R-1} \left( \frac{d\sigma_n^2}{h_{\min,t}^2} \right)^2 \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right)^2 + \frac{R\nu^2}{2}. \end{aligned} \quad (32)$$

### B. Upper Bound on Virtual Queue

**Lemma 3** (Virtual queue upper bound). *Under AdaScale, the virtual queue is upper bounded by*

$$Q_t \leq Q_T^{\max}, 0 \leq t \leq T, \quad (33)$$

where  $Q_T^{\max} \triangleq \left( 2V \sum_{t=0}^{T-1} \sum_{m=1}^M \rho_\alpha(q_m, \sigma_{m,t}^{\min}) + T\nu^2 \right)^{\frac{1}{2}}$  and  $\sigma_{m,t}^{\min} \triangleq \frac{\sigma_n M B_m}{G h_{\min,t} \sqrt{2x_{\max}}}$ .

*Proof.* From (32), we have

$$\begin{aligned} V \sum_{t=0}^{R-1} \sum_{m=1}^M \rho_\alpha(q_m, \hat{\sigma}_{m,t}) + \Delta_R(0) &\leq \\ V \sum_{t=0}^{R-1} \sum_{m=1}^M \rho_\alpha(q_m, \hat{\sigma}_{m,t}) + \frac{1}{2} \sum_{t=0}^{R-1} \left( \frac{d\sigma_n^2}{h_{\min,t}^2} \right)^2 \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right)^2 \\ &\quad + \sum_{t=0}^{R-1} Q_t \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right) + \frac{R\nu^2}{2}. \end{aligned} \quad (34)$$

Since AdaScale solves (28) optimally, and the RHS of (34) is the summation of the objective function of (28) (up to a constant) over rounds, AdaScale achieves the minimum value of the RHS of (34). In particular, considering  $x_t = x_{\max}, \forall t$ ,

as a feasible solution to (28) and its resultant RDP leakage  $\sigma_{m,t}^{\min}, \forall t$ , we obtain

$$\begin{aligned} & V \sum_{t=0}^{R-1} \sum_{m=1}^M \rho_{\alpha}(q_m, \hat{\sigma}_{m,t}) + \Delta_R(0) \\ & \leq V \sum_{t=0}^{R-1} \sum_{m=1}^M \rho_{\alpha}(q_m, \sigma_{m,t}^{\min}) + \frac{R\nu^2}{2}. \end{aligned} \quad (35)$$

This implies

$$\Delta_R(0) \leq V \sum_{t=0}^{R-1} \sum_{m=1}^M \rho_{\alpha}(q_m, \sigma_{m,t}^{\min}) + \frac{R\nu^2}{2}, 1 \leq R \leq T. \quad (36)$$

Using (30) together with (36), we can provide an upper bound on the queue length as

$$Q_R \leq \left( 2V \sum_{t=0}^{R-1} \sum_{m=1}^M \rho_{\alpha}(q_m, \sigma_{m,t}^{\min}) + R\nu^2 \right)^{\frac{1}{2}} \quad (37)$$

$$\stackrel{(a)}{\leq} Q_T^{\max}, \quad 1 \leq R \leq T, \quad (38)$$

where (a) follows from the fact that the RHS of (37) is an increasing function of  $R$ .  $\square$

### C. Constraint Violation Bound

The following theorem provides an upper bound on the amount of violation with respect to the constraint (26b).

**Theorem 2** (Constraint violation bound). *Under AdaScale, the constraint violation of problem (26) is upper bounded as*

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right) - \nu \leq \frac{Q_T^{\max}}{T}. \quad (39)$$

*Proof.* We have

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right) - \nu \stackrel{(a)}{\leq} \frac{1}{T} \sum_{t=0}^{T-1} (Q_{t+1} - Q_t) \quad (40)$$

$$\stackrel{(b)}{\leq} \frac{Q_T^{\max}}{T}, \quad (41)$$

where (a) follows from  $Q_t + \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right) - \nu \leq Q_{t+1}$  based on (27), and (b) follows from Lemma 3.  $\square$

### D. Dynamic Regret Bound

Let  $\{x_t^*\}$  denote the offline optimal solution to (26) when all future information is available and  $\sigma_{m,t}^* \triangleq \frac{\sigma_n M B_m}{G h_{\min,t} \sqrt{2x_t^*}}$ . We aim to derive an upper bound on the dynamic regret, which is the difference in the time-averaged RDP leakage achieved under AdaScale and that of  $\{x_t^*\}$ .

**Theorem 3** (Dynamic regret bound). *The dynamic regret of AdaScale is upper bounded as*

$$\begin{aligned} & \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^M \left( \rho_{\alpha}(q_m, \hat{\sigma}_{m,t}) - \rho_{\alpha}(q_m, \sigma_{m,t}^*) \right) \\ & \leq \frac{Q_T^{\max} \nu}{V} + \frac{T\nu^2}{2V} + \frac{\nu^2}{2V}. \end{aligned} \quad (42)$$

*Proof.* We use a similar argument as in the proof of Lemma 3. Using (34) with  $R = T$ , and considering  $x_t = x_t^*, \forall t$ , as a feasible solution to (28),

$$\begin{aligned} & V \sum_{t=0}^{T-1} \sum_{m=1}^M \rho_{\alpha}(q_m, \hat{\sigma}_{m,t}) + \Delta_T(0) \\ & \leq V \sum_{t=0}^{T-1} \sum_{m=1}^M \rho_{\alpha}(q_m, \sigma_{m,t}^*) + \frac{1}{2} \sum_{t=0}^{T-1} \left( \frac{d\sigma_n^2}{h_{\min,t}^2} \right)^2 \left( \frac{1}{x_t^*} - \frac{1}{x_{\max}} \right)^2 \\ & \quad + \sum_{t=0}^{T-1} Q_t \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t^*} - \frac{1}{x_{\max}} \right) + \frac{T\nu^2}{2}. \end{aligned} \quad (43)$$

We now provide an upper bound on the RHS of (43). The second term in the RHS of (43) can be upper bounded as

$$\begin{aligned} & \frac{1}{2} \sum_{t=0}^{T-1} \left( \frac{d\sigma_n^2}{h_{\min,t}^2} \right)^2 \left( \frac{1}{x_t^*} - \frac{1}{x_{\max}} \right)^2 \\ & \stackrel{(a)}{\leq} \frac{1}{2} \left( \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t^*} - \frac{1}{x_{\max}} \right) \right)^2 \stackrel{(b)}{\leq} \frac{T^2 \nu^2}{2}, \end{aligned} \quad (44)$$

where (a) follows from the fact  $\|\mathbf{y}\|_2 \leq \|\mathbf{y}\|_1$ , if all entries of  $\mathbf{y} \in \mathbb{R}^T$  are positive, and (b) is due to the fact that  $\{x_t^*\}$  meets the constraint (26b). Additionally, the third term on the RHS of (43) can be further upper bounded as

$$\begin{aligned} & \sum_{t=0}^{T-1} Q_t \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t^*} - \frac{1}{x_{\max}} \right) \stackrel{(a)}{\leq} Q_T^{\max} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{x_t^*} - \frac{1}{x_{\max}} \right) \\ & \stackrel{(b)}{\leq} T Q_T^{\max} \nu, \end{aligned} \quad (45)$$

where (a) follows the result in Lemma 3, and (b) is due to the fact that  $\{x_t^*\}$  meets the constraint (26b).

Applying the upper bounds in (44) and (45) on (43), and dividing both sides by  $TV$  and noting that  $\Delta_T(0) \geq 0$  completes the proof.  $\square$

### E. Discussion on Bounds

In the following, we first present Corollary 1 to simplify the bounds in Theorems 2 and 3 and elucidate their scaling w.r.t.  $T$ . We then draw connection with the convergence of FL training in Corollary 2.

**Corollary 1.** *Assume the minimum channel norm is bounded above, i.e.,  $\min_m |h_{m,t}| \leq h_{\text{ub}}, \forall t$ . Setting  $V \propto T^{\beta}$  for any  $\beta \in \mathbb{R}$ , the constraint violation bound in Theorem 2 and the dynamic regret bound in Theorem 3 reduce to the following:*

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{d\sigma_n^2}{h_{\min,t}^2} \left( \frac{1}{\hat{x}_t} - \frac{1}{x_{\max}} \right) - \nu \leq \mathcal{O}(T^{\frac{\beta-1}{2}}). \quad (46a)$$

$$\begin{aligned} & \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^M \left( \rho(q_m, \hat{\sigma}_{m,t}) - \rho(q_m, \sigma_{m,t}^*) \right) \\ & \leq \mathcal{O}(T^{\max\{1-\beta, \frac{1-\beta}{2}\}}). \end{aligned} \quad (46b)$$

*Proof.* Setting  $V \propto T^{\beta}$  in the bounds of Theorems 2 and 3, and upper bounding  $Q_T^{\max}$  using  $h_{\min,t} \stackrel{(a)}{\leq} \min_m |h_{m,t}| \leq h_{\text{ub}}$ , we obtain the results in (46a) and (46b).  $\square$

In Corollary 1, the parameter  $\beta$  balances the trade-off between utility and privacy. Specifically,  $\beta > 1$  yields a diminishing bound for the regret, while  $\beta < 1$  results in a diminishing bound for the constraint violation. Although these two regions of  $\beta$  do not overlap, the following corollary establishes that, when the minimum channel norm is bounded both below and above, and  $1 < \beta < 2$ , AdaScale achieves diminishing dynamic regret and ensures convergence to a stationary point of the global loss function.

**Corollary 2.** *Assume the minimum channel norm is bounded both below and above, i.e.,  $h_{lb} \leq \min_m |h_{m,t}| \leq h_{ub}, \forall t$ . Setting  $V \propto T^\beta$ , with  $1 < \beta < 2$ , yields a diminishing regret bound, when  $T \rightarrow \infty$ . Moreover, by setting  $\lambda \propto \frac{1}{\sqrt{T}}$ ,  $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\mathbf{w}_t)\|^2$  converges to zero when  $T \rightarrow \infty$ .*

*Proof.* Utilizing the result in (46b), it is clear that setting  $\beta > 1$  results in a diminishing time-averaged regret bound when  $T \rightarrow \infty$ . Further substituting  $\lambda \propto \frac{1}{\sqrt{T}}$  into (24), we observe that the first two terms on the RHS of (24) are  $\mathcal{O}(\frac{1}{\sqrt{T}})$ , while the third term is also  $\mathcal{O}(\frac{1}{\sqrt{T}})$  since  $\frac{h_{lb}}{\sqrt{2}} \leq \frac{\min_m |h_{m,t}|}{\max_m k_m} \leq h_{min,t}$  as  $k_m \leq \sqrt{2}, \forall m$ . Additionally, substituting  $\lambda$  into the fourth term, and using the result in (46a), we conclude that the fourth term is  $\mathcal{O}(T^{\frac{\beta-2}{2}})$ . Since  $\beta < 2$ , all the terms converge to zero as  $T \rightarrow \infty$ , which completes the proof.  $\square$

## VII. NUMERICAL EXPERIMENTS

We evaluate the effectiveness of AdaScale in reducing privacy leakage during OTA FL training for classification tasks on the MNIST [32] and CIFAR-10 [33] datasets.<sup>4</sup>

We consider  $M = 10$  devices, and set the maximum power limit to  $P_{\max} = 23$  dBm. Assuming a bandwidth of 100 kHz, we set the noise power to  $\sigma_n^2 = -90$  dBm, which accounts for both thermal noise and additional interference at the receiver. The distance of device  $m$  from the server is randomly generated, i.e.,  $d_m \sim \text{Uniform}[r_{\min}, r_{\max}]$  with  $r_{\min} = 10$  m,  $r_{\max} = 200$  m. The path loss follows the COST Hata model, i.e.,  $\text{PL}_m[\text{dB}] = 33.44 + 35.22 \log_{10}(d_m)$  [35], [36]. The channel between device  $m$  and the server in round  $t$  is generated as  $h_{m,t} \sim \mathcal{CN}(0, \frac{1}{\text{PL}_m})$ , which is i.i.d. across rounds.<sup>5</sup> We use the following benchmarks for comparison:

- **Optimal:** Assuming full knowledge of future information, problem (26) becomes a convex problem that can be solved optimally. The resulting solution serves as a lower bound on the achievable privacy leakage.
- **EqualAlloc:** This benchmark uniformly allocates  $\nu$  across all rounds to satisfy the constraint in (26b). Thus, it sets  $x_t = \frac{x_{\max}}{1 + \frac{x_{\max} \nu h_{2_{\min,t}}}{d \sigma_n^2}}, \forall t$ .
- **EstimFuture:** This method finds the MMSE estimation of the squared norm of future channels. Using these

estimations, the convex problem (26) is solved at each round to design  $x_t$ , given the remaining constraint budget.

- **Method in [20]:** This approach aims to enhance training convergence while bounding the DP leakage. To address the unknown future channels, it employs MMSE estimation of the squared channel norm.
- **Method in [23]:** This approach has the same aim as that of [20]. It is an online algorithm based on standard Lyapunov optimization.

Since, in practice, the upper bound on the norm of sample gradients  $G$  is unknown, we follow the convention in the DP literature [20], [21], [23] and apply a clipping operation to the sample gradients using a predefined threshold  $C$ . Specifically, each sample gradient is replaced by its clipped version as  $\mathbf{g}_{m,t,i} \leftarrow \mathbf{g}_{m,t,i} \min\left(1, \frac{C}{\|\mathbf{g}_{m,t,i}\|}\right)$ . Correspondingly, in our solution formulation,  $G$  is replaced by  $C$ .

Since problem (26) minimizes the overall RDP leakage subject to a long-term convergence constraint bounded by  $\nu$ , we consider different values of  $\nu$  as a measure of convergence and compare the resulting RDP and DP leakages across different methods for each  $\nu$ . Specifically, to evaluate the RDP leakage for a given order  $\alpha$ , we compute  $\rho_m = \sum_{t=0}^{T-1} \rho_\alpha(q_m, \sigma_{m,t})$  for each device, where the  $m$ -th device satisfies  $(\alpha, \rho_m)$ -RDP. The average RDP leakage across all devices is then reported as  $\frac{1}{M} \sum_{m=1}^M \rho_m$ . To evaluate DP leakage, we fix  $\delta = 10^{-5}$ , and compute  $\varepsilon_m$  for each device, where the  $m$ -th device satisfies  $(\varepsilon_m, \delta)$ -DP.<sup>6</sup> We then report the average DP leakage across all devices as  $\frac{1}{M} \sum_{m=1}^M \varepsilon_m$ . The Opacus library [40] is used to compute  $\rho_m$  and  $\varepsilon_m$ .

**Hyperparameter tuning:** For each value of  $\nu$ , we tune the hyperparameters of each method so that the LHS of constraint (26b) matches  $\nu$ . This approach allows for fair comparison of different methods under the *same* learning performance. Specifically, for AdaScale, we tune the parameter  $V$ ; for the method in [20], we tune the privacy budget  $\varepsilon_{\text{budget}}$ ; and for the method in [23], we tune both the privacy budget  $\varepsilon_{\text{budget}}$  and the objective multiplier  $V$  used in the Lyapunov framework. We set  $\alpha = 3$  for both ‘‘EstimFuture’’ and AdaScale in all experimental settings.<sup>7</sup>

### A. MNIST Dataset with I.I.D. Data Distribution

In MNIST, each data sample is a labeled grey-scaled handwritten digit image of size  $\mathbb{R}^{28} \times \mathbb{R}^{28}$  pixels, with a label indicating its class. There are 60,000 training and 10,000 test samples. We consider training a CNN whose architecture is detailed in [39], [41], with  $d = 26,010$  parameters.

An equal number of data samples from different classes are uniformly and randomly distributed among the devices. The batch size for each device is set to 60. We set the number of training epochs to 5 and thus the number of rounds is  $T = 500$ . The learning rate is constant throughout the training and set to

<sup>4</sup>The simulation codes and hyperparameter values are available at [34].

<sup>5</sup>The i.i.d. assumption is more realistic than a correlated channel model in FL, since the channel coherence time is typically much less than 200 milliseconds even for a fixed device in a wireless environment [37], [38], while a training round of FL typically has duration on the order of seconds and minutes or more.

<sup>6</sup>The value of  $\delta$  used in our experiments is commonly adopted in the literature for the datasets considered [39]. In principle,  $\delta$  should be chosen to be on the order of  $\frac{1}{2n}$ , where  $n$  denotes the dataset size.

<sup>7</sup>We observed that, for a fixed  $\nu$ , changing  $\alpha$  has a negligible impact on the privacy leakage as measured by DP.

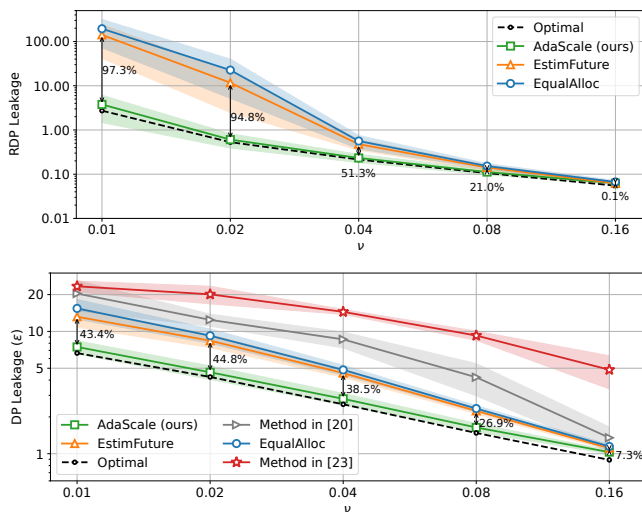


Fig. 1. RDP and DP leakage vs.  $\nu$  for MNIST. Range of  $\nu$  corresponds to test accuracies between 90% and 95%.

$\lambda = 0.5$ . The SGD optimizer with a weight decay of  $10^{-4}$  is utilized for training. The clipping threshold is set to  $C = 1.0$ . We consider several values of  $\nu$  ranging from 0.01 to 0.16, which correspond to test accuracies between 95% and 90%.

Fig. 1 illustrates the average overall RDP and DP leakages across devices plotted against  $\nu$ . The results are averaged over three realizations, and the shaded regions around each curve represent the 95% confidence intervals. Note that when evaluating RDP leakage, we consider only the first two benchmarks, “EqualAlloc” and “EstimFuture,” as the other two methods (from [20] and [23]) do not account for RDP leakage in their formulations and exhibit significantly higher leakage, making them incomparable. In contrast, for the evaluation of DP leakage, all four benchmarks are included in the comparison.

Fig. 1 shows that AdaScale reduces the RDP leakage compared with benchmarks across different values of  $\nu$ . Furthermore, AdaScale performs close to the offline Optimal benchmark. As  $\nu$  increases, the gap between AdaScale and the benchmarks narrows, since the benchmarks also approach near-optimal performance. However, it is important to note that the more desirable regime corresponds to smaller values of  $\nu$ , which correspond to higher learning accuracy, where AdaScale’s advantage becomes more pronounced.

Figure 1 further shows that all methods incur higher DP leakage as  $\nu$  decreases, which aligns with the results on RDP. We observe that although AdaScale is primarily designed with RDP as objective, it can effectively improve privacy in terms of the DP metric as well, outperforming state-of-the-art benchmarks and performs closely to the optimal offline solution. Again, this improvement becomes clearer for smaller values of  $\nu$ , which correspond to higher learning accuracies.

### B. CIFAR-10 Dataset with Non-I.I.D. Data Distribution

In CIFAR-10, each data sample consists of a colored image of size  $\mathbb{R}^3 \times \mathbb{R}^{32} \times \mathbb{R}^{32}$  and a label indicating the class of the

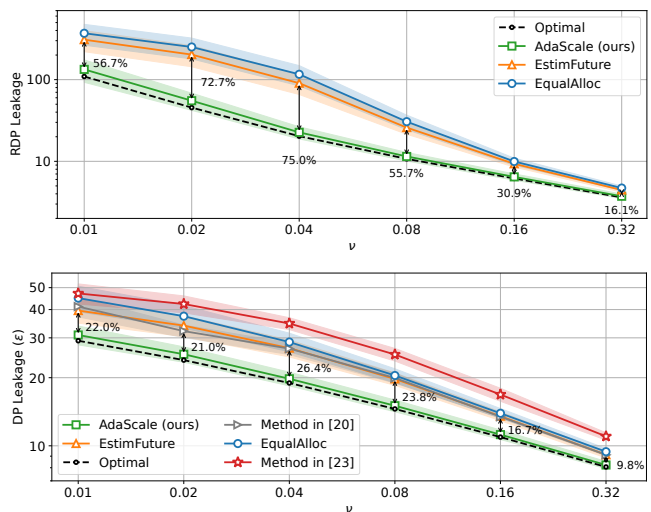


Fig. 2. RDP and DP leakage vs.  $\nu$  for CIFAR-10. Range of  $\nu$  corresponds to test accuracies between 60% and 65%.

image. There are 50,000 training and 10,000 test samples. We train the CNN described [42] with approximately 500,000 parameters using the cross-entropy loss.

The training data is distributed across devices in a non-i.i.d. manner, with each device containing 5000 samples only from two classes. The batch size is set to 400, and the training is conducted over 60 epochs, resulting in  $T = 720$ . We set  $C = 2.0$ . The learning rate is set to  $\lambda = 0.25$ , and the SGD optimizer with a momentum of 0.9 is used. We consider  $\nu$  from 0.01 to 0.32, which corresponds to a test accuracy between 65% and 60%.

Fig. 2 illustrates the RDP and DP leakages for various methods. As shown, for this more challenging learning task, AdaScale still effectively reduces both RDP and DP leakages across different values of the convergence level  $\nu$ , and its performance is close to that of the optimal offline solution.

## VIII. CONCLUSION

In this work, we have investigated adaptive design of the receive scaling factors in an OTA FL system under dynamic wireless channel conditions, to reduce the overall privacy leakage during training. Unlike previous works, we aimed to minimize the overall RDP leakage directly while ensuring a specific level of convergence for the global loss function. We propose AdaScale, a novel online algorithm with per-round optimization problems that can be efficiently solved. Through novel bounding techniques, we derive upper bounds on the dynamic regret and constraint violation of the proposed algorithm, establishing that it achieves diminishing dynamic regret in time-averaged RDP leakage while ensuring convergence to a stationary point of the global loss function. Numerical experiments show that our approach performs nearly optimally and effectively reduce both RDP and DP leakages compared with state-of-the-art benchmarks under the same learning performance.

## REFERENCES

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. on Artificial Intelligence and Statistics*, 2017.
- [2] 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, 2019.
- [3] 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, 2020.
- [4] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, 2020.
- [5] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, 2021.
- [6] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7595–7609, 2021.
- [7] J. Wang, B. Liang, M. Dong, G. Boudreau, and H. Abou-Zeid, "Joint online optimization of model training and analog aggregation for wireless edge learning," *IEEE/ACM Trans. Netw.*, vol. 32, no. 2, pp. 1212–1228, 2023.
- [8] M. Kim, A. L. Swindlehurst, and D. Park, "Beamforming vector design and device selection in over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7464–7477, 2023.
- [9] F. M. Kalarde, B. Liang, M. Dong, Y. Ahmed, and H. T. Cheng, "Power minimization in federated learning with over-the-air aggregation and receiver beamforming," in *Proc. Int. ACM Conf. Modeling, Analys. and Simul. of Wireless and Mobile Sys. (MSWiM)*, 2023.
- [10] F. M. Kalarde, M. Dong, B. Liang, Y. A. E. Ahmed, and H. T. Cheng, "Beamforming and device selection design in federated learning with over-the-air aggregation," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1710–1723, 2024.
- [11] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, 2020.
- [12] K. Wei, J. Li, C. Ma, M. Ding, C. Chen, S. Jin, Z. Han, and H. V. Poor, "Low-latency federated learning over wireless channels with differential privacy," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 290–307, 2022.
- [13] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, and K.-Y. Lam, "Local differential privacy-based federated learning for internet of things," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8836–8853, 2021.
- [14] N. Mohammadi, J. Bai, Q. Fan, Y. Song, Y. Yi, and L. Liu, "Differential privacy meets federated learning under communication constraints," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 22 204–22 219, 2022.
- [15] M. Kim, O. Günlü, and R. F. Schaefer, "Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2021, pp. 2650–2654.
- [16] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*. Hanover, MA, USA: Now Publishers, 2014.
- [17] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *Proc. IEEE Int. Symp. on Infor. Theory (ISIT)*, 2020.
- [18] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private aircomp federated learning with power adaptation harnessing receiver noise," in *IEEE Global Commun. Conf. (GLOBECOM)*, 2020.
- [19] N. Yan, K. Wang, C. Pan, K. K. Chai, F. Shu, and J. Wang, "Over-the-air federated averaging with limited power and privacy budgets," *IEEE Trans. Commun.*, vol. 72, no. 4, pp. 1998–2013, 2023.
- [20] 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, 2020.
- [21] Y. Yang, Y. Zhou, Y. Wu, and Y. Shi, "Differentially private federated learning via reconfigurable intelligent surface," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 19 728–19 743, 2022.
- [22] H. Liu, J. Yan, and Y.-J. A. Zhang, "Differentially private over-the-air federated learning over MIMO fading channels," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8232–8247, 2024.
- [23] Y. Shi, Y. Yang, and Y. Wu, "Federated edge learning with differential privacy: An active reconfigurable intelligent surface approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, 2024.
- [24] I. Mironov, "Rényi differential privacy," in *IEEE 30th Computer Security Foundations Symposium (CSF)*, 2017.
- [25] M. S. E. Mohamed, W.-T. Chang, and R. Tandon, "Privacy amplification for federated learning via user sampling and wireless aggregation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3821–3835, 2021.
- [26] B. Hasircioğlu and D. Gündüz, "Private wireless federated learning with anonymous over-the-air computation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2021.
- [27] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. CA, USA: Morgan & Claypool, 2010.
- [28] I. Mironov, K. Talwar, and L. Zhang, "Rényi differential privacy of the sampled Gaussian mechanism," *arXiv preprint arXiv:1908.10530*, 2019.
- [29] A. Sahu, A. Dutta, A. M. Abdelmoniem, T. Banerjee, M. Canini, and P. Kalnis, "Rethinking gradient sparsification as total error minimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.
- [30] D. Wang, M. Ye, and J. Xu, "Differentially private empirical risk minimization revisited: Faster and more general," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
- [31] X. Chen, S. Z. Wu, and M. Hong, "Understanding gradient clipping in private SGD: A geometric perspective," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
- [32] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits," [Online]. Available: <http://yann.lecun.com/exdb/mnist/>, 1998.
- [33] A. Krizhevsky and G. E. Hinton, "Learning multiple layers of features from tiny images," Master's thesis, University of Toronto, 2009.
- [34] F. M. Kalarde, "AdaScale: Privacy Enhancement in OTA-FL," <https://github.com/faezemoradik/AdaScale-Privacy-Enhancement-in-OTA-FL>, git, 2025, GitHub repository, accessed on: Jan. 8, 2026.
- [35] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE*. Hoboken, NJ, USA: John Wiley & Sons, 2010.
- [36] P. Kumar, B. Patil, and S. Ram, "Selection of radio propagation model for long-term evolution (LTE) network," *Int. J. Eng. Res. Gen. Sci.*, vol. 3, no. 1, pp. 373–379, 2015.
- [37] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2001.
- [38] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 933–946, 2002.
- [39] F. Tramer and D. Boneh, "Differentially private learning needs better features (or much more data)," in *Int. Conf. on Learning Representations (ICLR)*, 2021.
- [40] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov, "Opacus: User-friendly differential privacy library in PyTorch," *arXiv preprint arXiv:2109.12298*, 2021.
- [41] N. Papernot, A. Thakurta, S. Song, S. Chien, and Ú. Erlingsson, "Tempered sigmoid activations for deep learning with differential privacy," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2021.
- [42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.