Wireless Federated Learning with Local Differential Privacy

Mohamed Seif Ravi Tandon Ming Li
Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ, 85721
Email: {*mseif, tandonr, lim*}@email.arizona.edu

***Abstract*—In this paper, we study the problem of federated learning (FL) over a wireless channel, modeled by a Gaussian multiple access channel (MAC), subject to local differential privacy (LDP) constraints. We show that the superposition nature of the wireless channel provides a dual benefit of bandwidth efficient gradient aggregation, in conjunction with strong LDP guarantees for the users. We propose a private wireless gradient aggregation scheme, which shows that when aggregating gradients from K users, the privacy leakage per user scales as $\mathcal{O}\big(\frac{1}{\sqrt{K}}\big)$ compared to orthogonal transmission in which the privacy leakage scales as a constant. We also present analysis for the convergence rate of the proposed private FL aggregation algorithm and study the tradeoffs between wireless resources, convergence, and privacy.**

I. Introduction

Federated learning (FL) [1] is a framework that enables multiple users to jointly train a learning model. In prototypical FL, a central server interacts with multiple users to train a ML model in an iterative manner as follows: users compute gradients for the ML model on their local data sets, and gradients are subsequently exchanged for model updates. There are several motivating factors behind the surging popularity of FL: a) centralized approaches can be inefficient in terms of storage/computation, and FL provides natural parallelization for training, and can leverage increasing computational power of devices and b) local data at each user is never shared, but only gradient computations from each user are collected. Despite the fact that in FL, local data is never shared by a user, even exchanging gradients in a raw form can leak information, as shown in recent works [2]–[4].

Motivated by these factors, there has been a recent surge in designing FL algorithms with rigorous privacy guarantees. Differential privacy (DP) [5] has been adopted a *de facto* standard notion for private data analysis and aggregation. Within the context of FL, the notion of local differential privacy (LDP) is more suitable in which a user can locally perturb and disclose the data to an *untrusted* data curator/aggregator [6]. LDP has been already adopted and used in current applications, including Google's RAPPOR [7] for website browsing history aggregation, and by Microsoft for privately collecting telemetry data [8]. In the literature, there has been several research efforts to design FL algorithms satisfying LDP [9]–[15]. While LDP provides stronger privacy guarantees (compared to a centralized solution), this comes at the cost of lower utility. In particular, to achieve the same level of privacy attained by a centralized solution, significantly higher amount of noise/perturbation is needed [16]–[20].

Another parallel recent trend is to study the feasibility of FL over wireless channels. As the prototypical computation for FL training involves gradient aggregation from multiple users, the superposition property of the wireless channel can naturally support this operation much more efficiently. This has led to several recent works [21]–[31] under the umbrella of FL at the wireless edge, where distributed users interact with a parameter server (PS) over a shared wireless medium for training ML models. Several methodologies have been proposed to study wireless FL, which can be broadly categorized into either digital or analog aggregation schemes. In digital schemes, quantized gradients from each user are individually transmitted to the PS using orthogonal transmission. For analog schemes, on the other hand, the gradient computations are rescaled and transmitted directly over the air by all users simultaneously. The superposition nature of the wireless medium makes analog schemes more bandwidth efficient compared to digital ones.

In this paper, we focus on the following question: *Can the superposition property of wireless also be beneficial for privacy? If yes, how can we optimally utilize the wireless resources, and what are the tradeoffs between convergence of FL training, wireless resources and privacy?*

Main Contributions: In this paper, we consider the problem of FL training over a flat-fading Gaussian multiple access channel (MAC), subject to LDP constraints. We propose and study analog aggregation schemes, in which each user transmits a linear combination of a) local gradients and b) artificial Gaussian noise, subject to power constraints. The local gradients are processed as a function of the channel gains to *align* the resulting gradients at the PS, whereas the artificial noise parameters are selected to satisfy the privacy constraints. We show that the privacy level per user scales as $\mathcal{O}\big(\frac{1}{\sqrt{K}}\big)$ compared to orthogonal transmission in which the privacy leakage scales as a constant. We also provide the privacy-convergence trade-offs for smooth and convex loss functions through convergence analysis of the distributed gradient descent algorithm. We show that the training error decreases as the number of users increases and converges to the centralized algorithm where all points are available at the

This work has been supported in part by NSF Grants CAREER 1651492, CNS 1715947, CNS 1731164, CNS 1564477, and the 2018 Keysight Early Career Professor Award.

PS. To the best of our knowledge, this is the first result on wireless FL with LDP constraints.

II. System Model & Problem Statement

Wireless Channel Model: We consider a single-antenna wireless FL system with K users and a central PS as shown in Fig. 1. The input-output relationship at time i is

$$y(i) = \sum_{k=1}^{K} h_k x_k(i) + m(i), \tag{1}$$

where $x_k(i)$ is the signal transmitted by user k at time i, and $y(i)$ is the received signal at the PS. Here, $h_k = |h_k| e^{j\phi_k}$ is the complex valued channel coefficient between the k-th user and the PS, and and $m(i)$ is the independent additive zero-mean unit-variance (AWGN) Gaussian noise. The channel coefficients are assumed to be time invariant, and each user can transmit subject to maximum power constraint of P_k. Each user is assumed to know its local channel gains, whereas we assume that the PS has global channel state information.

Federated Learning Problem: Each user k has a private local dataset $\mathcal{D}_k$ of size $|\mathcal{D}_k|$ data points, denoted as $\mathcal{D}_k = \{(\mathbf{u}_i^{(k)}, v_i^{(k)})\}_{i=1}^{|\mathcal{D}_k|}$, where $\mathbf{u}_i^{(k)}$ is the i-th data point and $v_i^{(k)}$ is the corresponding label at user k. Users communicate with the PS through the Gaussian MAC described above in order to train a model by minimizing the loss function $F(\mathbf{w})$, i.e.,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w}) \triangleq \frac{1}{|\mathcal{D}_{\text{total}}|} \sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{D}_k|} f_k((\mathbf{u}_i^{(k)}, v_i^{(k)}); \mathbf{w}),$$

where $\mathbf{w} \in \mathbb{R}^d$ is the parameter vector to be optimized, $f_k(\cdot)$ is the loss function for user k, and $\mathcal{D}_{\text{total}} = \cup_{k=1}^{K} \mathcal{D}_k$ denotes the entire dataset used for training. The minimization of $F(\mathbf{w})$ is carried out iteratively through a distributed gradient descent (GD) algorithm. More specifically, in the t-th training iteration, the PS broadcasts the global parameter vector $\mathbf{w}_t$ from the last iteration to all users. Each user k computes his local gradient over the local $|\mathcal{D}_k|$ data points, i.e., $\mathbf{g}_k(\mathbf{w}_t) = \frac{1}{|\mathcal{D}_k|} \sum_{i=1}^{|\mathcal{D}_k|} \nabla f_k((\mathbf{u}_i^{(k)}, v_i^{(k)}); \mathbf{w})$, and sends back the computed gradient to the PS. For the scope of this paper, we assume that $|\mathcal{D}_k| = |\mathcal{D}|$, therefore $|\mathcal{D}_{\text{total}}| = K|\mathcal{D}|$. The global parameter $\mathbf{w}_t$ is updated according to

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \frac{1}{K} \sum_{k=1}^{K} \mathbf{g}_k(\mathbf{w}_t), \tag{2}$$

where η_t is the learning rate of the distributed GD algorithm at iteration t. The iteration process continues until convergence. In addition, the gradient descent (GD) algorithm for wireless FL should also satisfy local differential privacy (LDP) constraints for each user, as defined next.

Definition 1. *((ϵ, δ)-LDP [32]) A randomized mechanism $\mathcal{M}: \mathcal{X} \to \mathbb{R}^d$ is (ϵ, δ)-LDP if for any pair $x, x' \in \mathcal{X}$ and any measurable subset $\mathcal{O} \subseteq Range(\mathcal{M})$, we have*

$$\Pr(\mathcal{M}(x) \in \mathcal{O}) \leq e^{\epsilon} \Pr(\mathcal{M}(x') \in \mathcal{O}) + \delta. \tag{3}$$

The case of $\delta = 0$ is called pure ϵ-LDP.

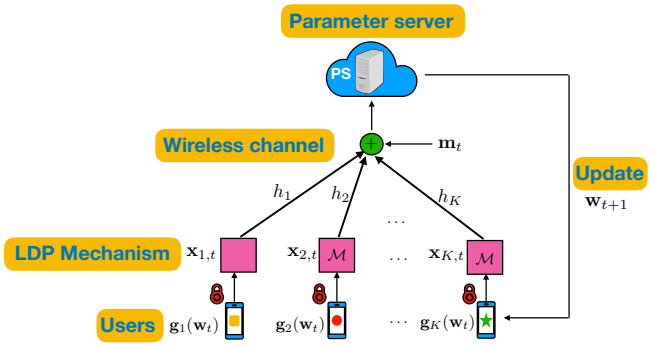

Fig. 1. Illustration of the private wireless FL framework: Users collaborate with the PS to jointly train a machine learning model over a Gaussian MAC. The interaction between the users and the PS must satisfy local differential privacy (LDP) constraints for each user.

Problem Statement. The main goal of this paper is to explore the benefits of wireless gradient aggregation for privacy in FL. In addition, we investigate tradeoffs between the convergence rate of GD, wireless channel conditions and resources (such as power, SNR), subject to the privacy budgets of the users.

III. Main Results & Discussions

In this Section, we present a general gradient aggregation scheme for wireless FL, where each user transmits a linear combination of its local gradients and artificial noise. We then specialize this scheme in which the part of transmission containing gradients are designed in a manner so that this component is aligned at the PS. We analyze this scheme and obtain the privacy leakage under LDP for each user, as a function of the wireless channel conditions, and the transmission parameters. Finally, we present the convergence rate of the private FL algorithm, and maximize the convergence rate by optimizing the local perturbations of each user for privacy.

A. FL Transmission Scheme over Gaussian MAC

The overall FL scheme consists of T training iterations, where each iteration comprises of d uses of the wireless channel described in (1). At each iteration t, each user k transmits the computed gradient vector $\mathbf{g}_k(\mathbf{w}_t) \in \mathbb{R}^d$ together with additive Gaussian noise for privacy. In particular, the transmitted signal of user k at iteration t is given as:

$$\mathbf{x}_{k,t} = e^{-j\phi_k} \left[\underbrace{\frac{\sqrt{\alpha_k P_k}}{L} \mathbf{g}_k(\mathbf{w}_t)}_{\text{local gradient estimate}} + \underbrace{\sqrt{\beta_k P_k} \mathbf{n}_{k,t}}_{\text{local perturbation}} \right]. \tag{4}$$

Here, each user k performs local phase correction (i.e., input is multiplied by $e^{-j\phi_k}$) so that the received channel coefficient is non-negative, i.e., $|h_k|$. We assume that the gradient vectors have a bounded norm, i.e., $\|\mathbf{g}_k(\mathbf{w}_t)\|_2 \leq L, \forall k$, and normalize the gradient vector by L. Also, $\alpha_k \in [0, 1]$ denotes the fraction of power dedicated to the gradient vector $\mathbf{g}_k(\mathbf{w}_t)$, whereas $\beta_k \in [0, 1 - \alpha_k]$ is the fraction of power dedicated to artificial Gaussian noise $\mathbf{n}_{k,t}$, whose elements are i.i.d., and drawn from $\mathcal{N}(0, 1)$. These parameters satisfy $\alpha_k + \beta_k \leq 1$ so that the

maximum power constraint of P_k is satisfied. From (1) and (4), the received signal at the PS can be written as:

$$\begin{aligned}\mathbf{y}_t &= \sum_{k=1}^{K} |h_k| \left[\frac{\sqrt{\alpha_k P_k}}{L} \mathbf{g}_k(\mathbf{w}_t) + \sqrt{\beta_k P_k} \mathbf{n}_{k,t} \right] + \mathbf{m}_t \\ &= \underbrace{\sum_{k=1}^{K} |h_k| \frac{\sqrt{\alpha_k P_k}}{L} \mathbf{g}_k(\mathbf{w}_t)}_{\text{aggregated gradient at PS}} + \underbrace{\sum_{k=1}^{K} |h_k| \sqrt{\beta_k P_k} \mathbf{n}_{k,t}}_{\text{aggregated noise at PS}} + \mathbf{m}_t, \end{aligned} \tag{5}$$

where $\mathbf{m}_t \in \mathbb{R}^d$ is the independent Gaussian noise, whose elements are i.i.d. drawn from $\mathcal{N}(0, \sigma_m^2)$. In order to carry out the summation of the local gradients over-the-air, and receive an unbiased estimate of the true aggregated gradient, all users pick the coefficients α_ks in order to align their transmitted local gradient estimates. Specifically, user k picks α_k so that

$$\frac{|h_k| \sqrt{\alpha_k P_k}}{L} = c, \forall k, \tag{6}$$

where c is a constant. From (6), we obtain $\alpha_k = \frac{c^2 L^2}{|h_k|^2 P_k}$, and using the fact that $\alpha_k \leq 1$, for all k, we can upper bound the constant c as follows: $c \leq \frac{\sqrt{\min_j |h_j|^2 P_j}}{L}$. To maximize the signal power of the aligned gradient, we choose c to match this upper bound, i.e.,

$$c = \frac{\sqrt{\min_j |h_j|^2 P_j}}{L}. \tag{7}$$

Plugging this back in (6), we obtain the choice of α_k as

$$\alpha_k = \frac{\min_j |h_j|^2 P_j}{|h_k|^2 P_k}. \tag{8}$$

The above choice shows that alignment of gradients is effectively limited by the user with the worst effective SNR, i.e., $\min_j |h_j|^2 P_j$. For the alignment scheme described above, the received signal by the PS in iteration t in (5) simplifies to:

$$\mathbf{y}_t = c \sum_{k=1}^{K} \mathbf{g}_k(\mathbf{w}_t) + \sum_{k=1}^{K} |h_k| \sqrt{\beta_k P_k} \mathbf{n}_{k,t} + \mathbf{m}_t. \tag{9}$$

The PS subsequently performs post-processing on $\mathbf{y}_t$ as follows:

$$\begin{aligned}\hat{\mathbf{g}}_t &= \frac{1}{Kc} \times \mathbf{y}_t \\ &= \underbrace{\frac{1}{K} \sum_{k=1}^{K} \mathbf{g}_k(\mathbf{w}_t)}_{\nabla F(\mathbf{w}_t)} + \underbrace{\frac{1}{Kc} \times \left[\sum_{k=1}^{K} |h_k| \sqrt{\beta_k P_k} \mathbf{n}_{k,t} + \mathbf{m}_t \right]}_{\mathbf{z}_t}, \end{aligned} \tag{10}$$

where $\mathbf{z}_t \sim \mathcal{N}(0, \sigma_z^2 \mathbf{I}_d)$ is the effective noise at the PS, and $\sigma_z^2 = \frac{1}{K^2 c^2} \left[\sum_{k=1}^{K} |h_k|^2 \beta_k P_k + \sigma_m^2 \right]$. Thus, we can write $\hat{\mathbf{g}}_t = \nabla F(\mathbf{w}_t) + \mathbf{z}_t$. As $\mathbf{z}_t$ is zero mean, $\hat{\mathbf{g}}_t$ is an unbiased estimate of $\nabla F(\mathbf{w}_t)$, with variance of $\hat{\mathbf{g}}_t$ being equal to σ_z^2.

B. Local Differential Privacy Analysis

We next analyze the privacy level achieved by the transmission scheme for each user, as per the definition of LDP. Recall, that the local perturbation noise is drawn from Gaussian distribution. This well-known technique is known as Gaussian mechanism and can provide rigorous privacy guarantees based on LDP, as defined next.

Definition 2. *(Gaussian Mechanism - Appendix A of [32]) Suppose a user wants to release a function $f(X)$ of an input X subject to (ϵ, δ)-LDP. The Gaussian release mechanism is defined as:*

$$M(X) \triangleq f(X) + \mathcal{N}(0, \sigma^2 \mathbf{I}). \tag{11}$$

If the sensitivity of the function is bounded by Δ_f, i.e., $\|f(x) - f(x')\|_2 \leq \Delta_f$, $\forall x, x'$, then for any $\delta \in (0, 1]$, Gaussian mechanism satisfies (ϵ, δ)-LDP, where

$$\epsilon = \frac{\Delta_f}{\sigma} \sqrt{2 \log \frac{1.25}{\delta}}. \tag{12}$$

In the next Theorem, we make use of the above result, and present the per-user privacy achieved by the proposed wireless FL scheme as a function of the noise power allocation parameters $\{\beta_k\}_{k=1}^K$, transmit powers $\{P_k\}_{k=1}^K$, and the channel coefficients $\{h_k\}_{k=1}^K$.

Theorem 1. *For each user k, the proposed transmission scheme achieves (ϵ_k, δ)-LDP per iteration, where*

$$\epsilon_k = \frac{2\sqrt{\min_j |h_j|^2 P_j}}{\sqrt{\sum_{k=1}^{K} |h_k|^2 \beta_k P_k + \sigma_m^2}} \sqrt{2 \log \frac{1.25}{\delta}}. \tag{13}$$

Proof. The final received signal at the PS from (9) can be expressed as: $\mathbf{y}_t = c \sum_{k=1}^{K} \mathbf{g}_k(\mathbf{w}_t) + Kc\mathbf{z}_t$. We first observe that the variance of the effective Gaussian noise, i.e., variance of $Kc\mathbf{z}_t$ is $\sigma^2 = \sum_{k=1}^{K} |h_k|^2 \beta_k P_k + \sigma_m^2$. In order to invoke the result of the Gaussian mechanism, we next obtain a bound on the sensitivity for user k. To bound the local sensitivity of $c \sum_{k=1}^{K} \mathbf{g}_k(\mathbf{w}_t)$, consider any two different local datasets $\mathcal{D}_k$ and $\mathcal{D}_k'$ at user k, while fixing the datasets (and thus the gradients) of the remaining $(K-1)$ users. The local sensitivity of user k can then be bounded as

$$\begin{aligned}\Delta_k &= \max_{\mathcal{D}_k, \mathcal{D}_k'} ||\mathbf{y}_t - \mathbf{y}_t'||_2 = \max_{\mathcal{D}_k, \mathcal{D}_k'} ||c(\mathbf{g}_k(\mathbf{w}_t) - \mathbf{g}_k'(\mathbf{w}_t))||_2 \\ &\leq c \max_{\mathcal{D}_k, \mathcal{D}_k'} ||\mathbf{g}_k(\mathbf{w}_t)||_2 + ||\mathbf{g}_k'(\mathbf{w}_t)||_2 \overset{(a)}{\leq} 2cL \\ &\overset{(b)}{=} 2\sqrt{\min_j |h_j|^2 P_j}, \end{aligned} \tag{14}$$

where in step (a), we used the fact that $\|\mathbf{g}_k(\mathbf{w}_t)\|_2 \leq L, \forall k$, and (b) follows from (7). Hence, using the sensitivity bound in (14) together with the variance $\sigma^2 = \sum_{k=1}^{K} |h_k|^2 \beta_k P_k + \sigma_m^2$ in (12), we arrive at the proof of Theorem 1. □

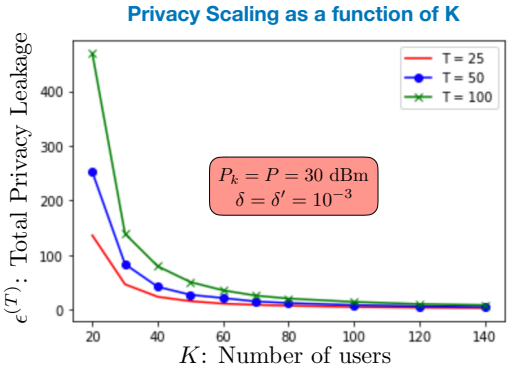

Fig. 2. Total per-user privacy leakage as a function of K, number of users for different values of T, the number of training iterations.

Remark 1. *From Theorem 1, we can observe the privacy benefits of wireless gradient aggregation. We can further upper bound the achievable ϵ_k in Theorem 1 as follows:*

$$\epsilon_k = \frac{2\sqrt{\min_j |h_j|^2 P_j}}{\sqrt{\sum_{k=1}^{K} |h_k|^2 \beta_k P_k + \sigma_m^2}} \sqrt{2\log\frac{1.25}{\delta}} \leq \frac{2\sqrt{\min_j |h_j|^2 P_j}}{\sqrt{\sum_{k=1}^{K} |h_k|^2 \beta_k P_k}} \sqrt{2\log\frac{1.25}{\delta}} \leq \frac{1}{\sqrt{K}} \times \frac{2\sqrt{\min_j |h_j|^2 P_j}}{\sqrt{\min_k |h_k|^2 \beta_k P_k}} \sqrt{2\log\frac{1.25}{\delta}},$$

which shows that asymptotically, the per-user privacy level behaves like $\mathcal{O}(1/\sqrt{K})$. In contrast, privacy achieved by orthogonal transmission can be shown to be:

$$\epsilon_k^{Orthogonal} = \frac{2|h_k|\sqrt{\alpha_k P_k}}{\sqrt{|h_k|^2 \beta_k P_k + \sigma_m^2}} \sqrt{2\log\frac{1.25}{\delta}}, \quad (15)$$

which scales as a constant, and does not decay with K.

Remark 2. *While Theorem 1 shows the per-iteration leakage, we can use advanced composition results for LDP using the Gaussian mechanism to obtain the total privacy leakage when the wireless FL algorithm is used for T iterations. Using existing results in [33], it can be readily shown that the total leakage over T iterations (per-user) of the proposed scheme is $(\epsilon_k^{(T)}, T\delta + \delta')$-LDP for $\delta' \in (0, 1]$ where,*

$$\epsilon_k^{(T)} = \sqrt{2T\log(1/\delta')}\epsilon_k + T\epsilon_k(e^{\epsilon_k} - 1). \quad (16)$$

We illustrate the total per-user privacy leakage as a function of K, the number of users in Fig. 2 for various values of T. As is clearly evident, the leakage provided by wireless FL goes asymptotically to 0 as $K \to \infty$.

C. Convergence rate of private FL

We next analyze the performance of private wireless FL under the assumption that the global loss function $F(\mathbf{w})$ is smooth and strongly convex. Due to privacy requirements and noisy nature of wireless channel, the convergence rate is penalized as shown in the following Theorem.

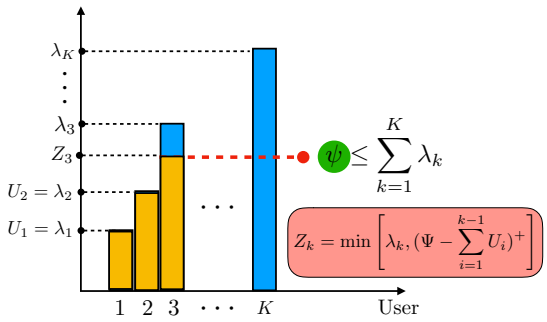

Fig. 3. An example for the iterative solution: $Z_1 + Z_2 + Z_3 \geq \Psi$, $Z_k = 0, k = 4, \cdots, K$.

Theorem 2. *Suppose the loss function F is λ-strongly convex and μ-smooth with respect to $\mathbf{w}^*$. Then, for a learning rate $\eta_t = 1/\lambda t$ and a number of iterations T, the convergence rate of the private wireless FL algorithm is*

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T} \times \left[L^2 + \frac{d}{K^2 c^2}\left[\sum_{k=1}^{K} |h_k|^2 \beta_k P_k + \sigma_m^2\right]\right]. \quad (17)$$

Theorem 2 is proved in Appendix I. We next show that artificial noise parameters $\{\beta_k\}_{k=1}^K$ can be optimized to maximize the convergence rate in (17) while satisfying a desired privacy level (ϵ_k, δ)-LDP at each user.

Theorem 3. *The optimized convergence rate of the private wireless FL algorithm is given as follows:*

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T} \times \left[L^2 + \frac{d}{K^2 c^2}\left[\sum_{k=1}^{K} Z_k + \sigma_m^2\right]\right], \quad (18)$$

where $Z_k = \min\left[\lambda_k, (\Psi - \sum_{i=1}^{k-1} U_i)^+\right]$ *where* $(a)^+ \triangleq \max(0, a)$, $\lambda_k = |h_k|^2 P_k(1 - \alpha_k)$, $\Psi = \max_i \frac{8\min_j |h_j|^2 P_j}{\epsilon_i^2} \log\frac{1.25}{\delta} - \sigma_m^2$, *and* $U_i = |h_i|^2 P_i \beta_i$.

Proof. Maximizing the convergence rate in (17) is equivalent to minimizing the term that depends on $\{\beta_k\}_{k=1}^K$. Therefore, we solve the following optimization problem:

$$\min_{\{\beta_i\}_{k=1}^K} \sum_{k=1}^{K} |h_k|^2 \beta_k P_k \quad \text{such that} \quad 0 \leq \beta_k \leq 1 - \alpha_k, \forall k,$$

$$\& \quad \sum_{k=1}^{K} |h_k|^2 \beta_k P_k \geq \frac{8\min_j |h_j|^2 P_j}{\epsilon_k^2} \log\frac{1.25}{\delta} - \sigma_m^2.$$

For given target privacy levels $\{\epsilon_k\}_{k=1}^K$, this is feasible when

$$\sum_{k=1}^{K} \underbrace{|h_k|^2 P_k(1 - \alpha_k)}_{\lambda_k} \geq \max_i \frac{8\min_j |h_j|^2 P_j}{\epsilon_i^2} \log\frac{1.25}{\delta} - \sigma_m^2.$$

We design $\beta_k, \forall k$ as follows:

$$\beta_k = \frac{Z_k}{|h_k|^2 P_k}, k = 1, \cdots, K. \quad (19)$$

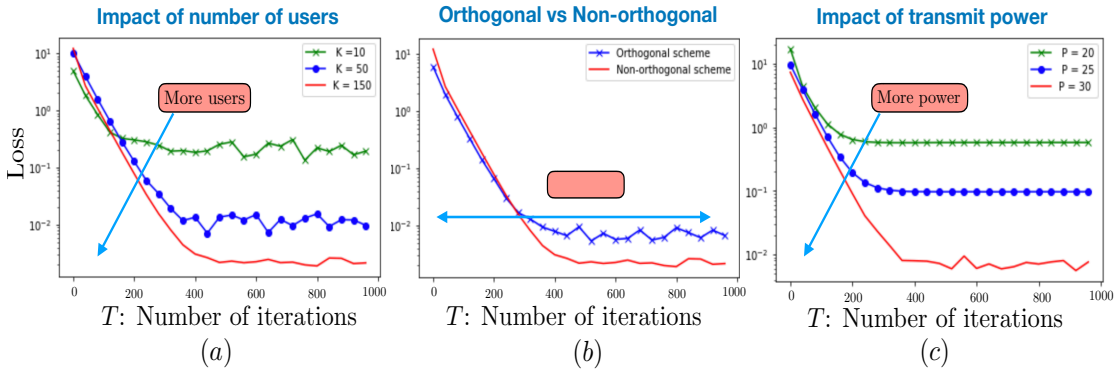

Fig. 4. Impact of a) number of users, b) orthogonal vs non-orthogonal transmission, and c) transmit power, on the training loss as a function of iterations. As we see from the figures, as T increases, the variance term due to the local privacy perturbation and the noisy channel becomes dominant.

where $Z_k = \min\left[\lambda_k, (\Psi - \sum_{i=1}^{k-1} U_i)^+\right], k = 1, \cdots, K$, $\Psi = \max_i \frac{8 \min_j |h_j|^2 P_j}{\epsilon_i^2} \log \frac{1.25}{\delta} - \sigma_m^2$, and $U_i = |h_i|^2 \beta_i P_i$. As seen in Fig. 3, we first rank the left-over powers from the users after aligning the gradients, i.e., $\{\lambda_k\}_{k=1}^K$ in an ascending order. We then allocate the powers Z_k such that a subset of users S satisfies $\sum_{k=1}^S Z_k \geq \psi, S \leq K$, to satisfy privacy constraints. This completes the proof of Theorem 3. □

IV. Simulation Results

In this Section, we provide some simulation results to assess the performance of private wireless FL model. We consider a linear regression task on a synthetic dataset. The regularized loss function at the kth user is given as:

$$f_k(\mathbf{w}) = \frac{1}{|\mathcal{D}_k|} \sum_{i=1}^{|\mathcal{D}_k|} (\mathbf{w}^T \mathbf{u}_i^{(k)} - v_i^{(k)})^2 + \frac{\lambda}{2} \|\mathbf{w}\|_2^2. \quad (20)$$

Our synthetic dataset consists of 3000 i.i.d. samples drawn from $\mathcal{N}(0, \mathbf{I}_{d+1})$, where $\mathbf{u}_i^{(k)} \in \mathbb{R}^d$, $v_i^{(k)} \in \mathbb{R}$ and $d = 30$. We assume that each user has $|\mathcal{D}_k| = 20$ data points. For the GD algorithm, the regularization parameter λ is 10^{-3} and $T = 1000$ training iterations. The channel coefficients are drawn from $\mathcal{CN}(0, 1)$, and the channel noise variance is set to $\sigma_m^2 = 1$. Also, we assume that each user requires the same privacy level $(\epsilon, \delta) = (1.2, 10^{-4})$-LDP.

In Fig. 4(a), we show the impact of the number of users on the training loss for $P_k = 30$ dBm for all k. As we increase the number of users, the training loss decays faster with T. In Fig. 4(b), we compare with the private orthogonal scheme for $KT_2 = T_1 = T$ iterations and $P_k = 30$ dBm for all k. Interestingly, the non-orthogonal scheme is more efficient in terms of the bandwidth and accuracy. In Fig. 4(c), we show the impact of the transmit power on the training loss where the error decays faster with T as we increase the transmit power.

V. Conclusion & Future Directions

We studied the problem of wireless federated learning subject to local differential privacy (LDP) constraints. We showed that the wireless channel provides a dual benefit of bandwidth efficiency together with strong LDP guarantees. Using the proposed wireless aggregation scheme, privacy leakage was shown to scale as $\mathcal{O}(\frac{1}{\sqrt{K}})$ compared to orthogonal transmission in which the privacy leakage scales as a constant. We also analyzed and optimized the convergence rate of the proposed private FL training algorithm and studied the tradeoffs between wireless resources, convergence, and privacy.

There are several interesting directions for future work, such as generalization to multiple-antennas at the users and the PS. In the proposed scheme, all users align their gradients, which limits the effective SNR by a user with the worst channel conditions. A possible direction would be to explore generalizations of this scheme, by selecting and aligning gradients from a smaller subsets of users.

Appendix I: Proof of Theorem 2

To prove the convergence rate of the proposed algorithm, we recall that the gradient estimate at the PS in (10) satisfies: (a) Unbiasedness, i.e., $\mathbb{E}[\hat{\mathbf{g}}_t] = \mathbb{E}[\nabla F(\mathbf{w}_t)]$, since the total additive noise is zero mean; and (b) Bounded second moment, $\mathbb{E}\left[\|\hat{\mathbf{g}}_t\|_2^2\right] \leq G^2$, which we prove as follows:

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\mathbf{g}}_t\|_2^2\right] = \mathbb{E}\left[\|\nabla F(\mathbf{w}_t) + \mathbf{z}_t\|_2^2\right] \\
&= \mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|_2^2\right] + 2\mathbb{E}\left[\nabla F(\mathbf{w}_t)^T \mathbf{z}_t\right] + \mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right] \\
&\overset{(a)}{=} \|\nabla F(\mathbf{w}_t)\|_2^2 + \mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right] \\
&\overset{(b)}{\leq} \frac{1}{K^2} \times \left(\sum_{k=1}^K \|\mathbf{g}_k(\mathbf{w}_t)\|_2\right)^2 + \mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right] \\
&\overset{(c)}{\leq} \frac{1}{K^2} \times (KL)^2 + \frac{d}{K^2 c^2} \left[\sum_{k=1}^K |h_k|^2 \beta_k P_k + 1\right] \\
&\leq L^2 + \frac{d}{K^2 c^2} \left[\sum_{k=1}^K |h_k|^2 \beta_k P_k + 1\right] \triangleq G^2, \quad (21)
\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}\left[\nabla F(\mathbf{w}_t)^T \mathbf{z}_t\right] = 0$, (b) follows from Cauchy-Schwarz inequality, and (c) from the assumption that $\|\mathbf{g}_k(\mathbf{w}_t)\|_2 \leq L$, i.e., the Lipschitz constant $\forall k$. We next invoke standard results [34] on convergence of SGD for μ-smooth and λ-strongly convex loss, which states

$$\mathbb{E}[F(\mathbf{w}_T)] - F(\mathbf{w}^*) \leq \frac{2\mu G^2}{\lambda^2 T}. \quad (22)$$

Plugging G^2 from (21) in (22), we arrive at Theorem 2.

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 *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (S & P)*, May 2017, pp. 3–18.

[3] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 1, pp. 133–152, 2019.

[4] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (S & P)*, May 2019, pp. 691–706.

[5] C. Dwork, "Differential privacy," in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Part II*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds., 2006, pp. 1–12. [Online]. Available: https://doi.org/10.1007/11787006_1

[6] M. Joseph, A. Roth, J. Ullman, and B. Waggoner, "Local differential privacy for evolving data," in *Advances in Neural Information Processing Systems*, 2018, pp. 2375–2384.

[7] G. Fanti, V. Pihur, and Ú. Erlingsson, "Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 3, pp. 41–61, 2016.

[8] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Advances in Neural Information Processing Systems*, 2017, pp. 3571–3580.

[9] A. Triastcyn and B. Faltings, "Federated learning with Bayesian differential privacy," *arXiv preprint arXiv:1911.10071*, 2019.

[10] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[11] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 453–15 462.

[12] C. Wu, F. Zhang, and F. Wu, "Distributed modelling approaches for data privacy preserving," in *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, September 2019, pp. 357–365.

[13] O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, and A. Das, "Differential privacy-enabled federated learning for sensitive health data," *arXiv preprint arXiv:1910.02578*, 2019.

[14] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farhad, S. Jin, T. Q. Quek, and H. V. Poor, "Performance analysis on federated learning with differential privacy," *arXiv preprint arXiv:1911.00222*, 2019.

[15] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed sgd," in *Advances in Neural Information Processing Systems*, 2018, pp. 7564–7575.

[16] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: Local differential privacy in practice," in *Proceedings of the 2018 International Conference on Management of Data*. ACM, 2018, pp. 1655–1658.

[17] D. Wang, M. Gaboardi, and J. Xu, "Empirical risk minimization in non-interactive local differential privacy revisited," in *Advances in Neural Information Processing Systems*, 2018, pp. 965–974.

[18] R. Bassily, "Linear queries estimation with local differential privacy," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 721–729.

[19] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, June 2015, pp. 127–135.

[20] R. Bassily, K. Nissim, U. Stemmer, and A. G. Thakurta, "Practical locally private heavy hitters," in *Advances in Neural Information Processing Systems*, 2017, pp. 2288–2296.

[21] M. M. Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *arXiv preprint arXiv:1901.00844*, 2019.

[22] G. Zhu, Y. Wang, and K. Huang, "Low-latency broadband analog aggregation for federated edge learning," *arXiv preprint arXiv:1812.11494*, 2018.

[23] Q. Zeng, Y. Du, K. K. Leung, and K. Huang, "Energy-efficient radio resource allocation for federated edge learning," *arXiv preprint arXiv:1907.06040*, 2019.

[24] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *arXiv preprint arXiv:1812.11750*, 2018.

[25] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, March 2019.

[26] M. M. Amiri and D. Gunduz, "Federated learning over wireless fading channels," *arXiv preprint arXiv:1907.09769*, 2019.

[27] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *arXiv preprint arXiv:1908.07463*, 2019.

[28] ——, "A sequential gradient-based multiple access for distributed learning over fading channels," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, September 2019, pp. 303–307.

[29] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," *arXiv preprint arXiv:1909.02362*, 2019.

[30] L. U. Khan, N. H. Tran, S. R. Pandey, W. Saad, Z. Han, M. N. Nguyen, and C. S. Hong, "Federated learning for edge networks: Resource optimization and incentive mechanism," *arXiv preprint arXiv:1911.05642*, 2019.

[31] M. M. Amiri and D. Gündüz, "Over-the-air machine learning at the wireless edge," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2019, pp. 1–5.

[32] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[33] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, October 2010, pp. 51–60.

[34] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 1571–1578.