Privacy Amplification for Federated Learning via User Sampling and Wireless Aggregation

Mohamed Seif Wei-Ting Chang Ravi Tandon
Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ, 85721
Email: {*mseif, wchang, tandonr*}@email.arizona.edu

***Abstract*—In this paper, we study the problem of federated learning over a wireless channel with user sampling, modeled by a Gaussian multiple access channel, subject to central and local differential privacy (DP/LDP) constraints. It has been shown that the superposition nature of the wireless channel provides a dual benefit of bandwidth efficient gradient aggregation, in conjunction with strong DP guarantees for the users. Specifically, the central DP privacy leakage has been shown to scale as $\mathcal{O}(1/\sqrt{K})$, where K is the number of users. It has also been shown that user sampling coupled with orthogonal transmission can enhance the central DP privacy leakage with the same scaling behavior. In this work, we show that, by jointly incorporating both wireless aggregation and user sampling, one can obtain even stronger central DP guarantees. We propose a private wireless gradient aggregation scheme, which relies on independently randomized participation decisions by each user. The central DP leakage of our proposed scheme scales as $\mathcal{O}(1/K^{3/4})$. In addition, we show that LDP is also boosted by user sampling.**

Full version of the paper available in [1].

I. Introduction

Federated learning (FL) [2] is a framework that enables multiple users to jointly train a learning model with the help of a parameter server (PS), typically, in an iterative manner. In this paper, we focus on a variation of FL termed federated stochastic gradient descent (FedSGD), where users compute gradients for the machine learning (ML) model on their local datasets, and subsequently exchange the gradients for model updates at the PS. There are several motivating factors behind the surging popularity of FL: (a) centralized approaches can be inefficient in terms of storage/computation, whereas FL provides natural parallelization for training, and (b) local data at each user is never shared, but only the local gradients are collected. However, even exchanging gradients in a raw form can leak information, as shown in recent works [3]–[9]. In addition, exchanging gradients incurs significant communication overhead. Therefore, it is crucial to design training protocols that are both communication efficient and private.

Since the training of FedSGD involves gradient aggregation from multiple users, the superposition property of wireless channels can naturally support this operation. Several recent works [10]–[20] have focused on exploiting the wireless channel to alleviate the communication overhead of FL (see a comprehensive survey [21]). Depending on the transmission strategy, wireless FL can be broadly categorized into digital or analog schemes. In digital schemes, gradients from each user are compressed and transmitted to the PS using a multi-access scheme. However, digital schemes require the PS to decode individual gradients and then aggregate them. For analog schemes, on the other hand, gradient is rescaled at each user to satisfy the power constraint and to mitigate the effect of channel noise. All users then transmit the rescaled gradients via wireless channel simultaneously. Non-orthogonal over the air aggregation makes analog schemes more bandwidth efficient compared to digital ones.

This work has been supported in part by NSF Grants CAREER 1651492, CNS 1715947, CCF 2100013.

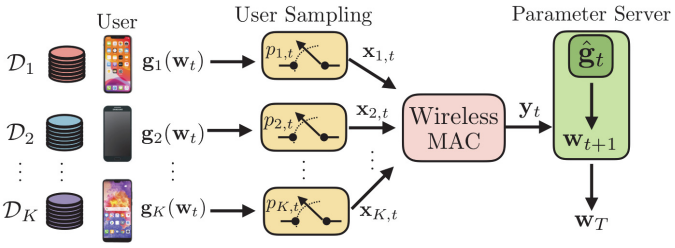

Fig. 1: Illustration of the private wireless FedSGD framework: Users collaborate with the PS to jointly train a ML model over a MAC.

In addition to saving bandwidth and computation, it has been shown in [22]–[24] that wireless FL also naturally provides strong differential privacy (DP) [25] guarantees. Specifically, in [22], it was shown that the superposition nature of the wireless channel provides a stronger privacy guarantee as well as faster convergence in comparison to orthogonal transmission. The privacy level is shown to scale as $\mathcal{O}(1/\sqrt{K})$, where K is the number of users in the wireless FL system. On the other hand, it was shown in [26] that one can obtain a similar scaling of $\mathcal{O}(1/\sqrt{K})$ for privacy leakage through user sampling. The scheme of [26], however, considers orthogonal transmission from the sampled users.

One natural question to ask is the following: *Could we achieve even stronger privacy guarantees by incorporating both user sampling and wireless aggregation? If it does provide stronger guarantee, how much additional gain can be obtained? How can we optimally utilize the wireless resources, and what are the tradeoffs between convergence of FedSGD training, wireless resources and privacy?*

Main Contributions: In this paper, we consider the problem of FedSGD training over Gaussian multiple access channels (MACs), subject to LDP and DP constraints. We propose a

Transmission scheme	Without sampling	With sampling
Orthogonal	$\mathcal{O}(1)$ [28]	$\mathcal{O}(1/\sqrt{K})$ [26]
Wireless Aggregation	$\mathcal{O}(1/\sqrt{K})$ [22]	$\mathcal{O}(1/K^{3/4})$ (This work)

TABLE I: Comparison for central privacy under different settings: (a) orthogonal and (b) wireless aggregation transmissions.

wireless FedSGD scheme with user sampling, where users are sampled uniformly or based on their channel conditions. We then study analog aggregation schemes coupled with the proposed sampling schemes, in which each user transmits a linear combination of (a) local gradient and (b) artificial Gaussian noise. 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. The existing privacy analysis in [26], [27] for FL with user sampling cannot be applied to our problem. The key challenge is that in each training iteration, the effective noise seen at the signal received by the PS over the wireless channel is a function of a random number of sampled users, making the DP/LDP analysis non-trivial. Using concentration inequalities, we are able to prove that the central privacy leakage scales as $\mathcal{O}(1/K^{3/4})$ with wireless aggregation and user sampling. We also provide convergence analysis of the proposed scheme for different sampling schemes in the full version of this paper [1]. To the best of our knowledge, this is the first result on wireless FedSGD with LDP and DP constraints with user sampling (see Table I for comparison of results).

II. System Model

Wireless Channel Model: We consider a wireless FL system with K users and a central PS. Users are connected to the PS through a Gaussian MAC as shown in Fig. 1. Let $\mathcal{K}_t$ denote the random set of users who participate in iteration t. The input-output relationship at the t-th block is

$$\mathbf{y}_t = \sum_{k\in\mathcal{K}_t} h_{k,t}\mathbf{x}_{k,t} + \mathbf{m}_t, \tag{1}$$

where $\mathbf{x}_{k,t} \in \mathbb{R}^d$ is the signal transmitted by user k at the t-th block, and $\mathbf{y}_t$ is the received signal at the PS. Here, $h_{k,t} \in \mathbb{R}$ is the channel coefficient between user k and the PS at iteration t. We assume a block flat-fading channel, where the channel coefficient remains constant within the duration of a communication block. Each user is assumed to know its local channel gain, whereas we assume that the PS has global channel state information (CSI). Each user can transmit subject to average power constraint i.e., $\mathbb{E}\left[\|\mathbf{x}_{k,t}\|_2^2\right] \leq P_k$. $\mathbf{m}_t \in \mathbb{R}^d$ is the channel noise whose elements are independent and identically distributed (i.i.d.) according to Gaussian distribution $\mathcal{N}(0, N_0)$. The random set of participants $\mathcal{K}_t$ can be obtained through various strategies. In this paper, we focus on user sampling, where user k participates in the training at time t according to probability $p_{k,t}$, for $k = 1, \ldots, K$. When $\mathcal{K}_t = [K]$, we recover the conventional federated SGD where every user participates in the training.

For this work, we consider (a) time-invariant uniform sampling; (b) time-variant uniform sampling; and (c) channel aware sampling. We note that sampling strategies based on gradients or losses can potentially leak information about local datasets, hence, require different privacy analysis. Thus, we leave gradient-based sampling strategies to future work.

Federated Learning Problem: Each user k has a private local dataset $\mathcal{D}_k$ with D_k data points, denoted as $\mathcal{D}_k = \{(\mathbf{u}_i^{(k)}, v_i^{(k)})\}_{i=1}^{D_k}$, where $\mathbf{u}_i^{(k)}$ is the i-th data point and $v_i^{(k)}$ is the corresponding label at user k. The local loss function at user k is given by $f_k(\mathbf{w}) = (1/D_k)\sum_{i=1}^{D_k} f(\mathbf{w}; \mathbf{u}_i^{(k)}, v_i^{(k)}) + \Omega R(\mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^d$ is the parameter vector to be optimized, $R(\mathbf{w})$ is a regularization function and $\Omega \geq 0$ is a regularization hyperparameter. 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.,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w}) \triangleq \frac{1}{\sum_{k=1}^K D_k}\sum_{k=1}^K D_k f_k(\mathbf{w}). \tag{2}$$

The minimization of $F(\mathbf{w})$ is carried out iteratively through a distributed SGD algorithm. More specifically, in the t-th training iteration, the PS broadcasts the global parameter vector $\mathbf{w}_t$ to all users. Each user k computes his local gradient using stochastic mini batch $\mathcal{B}_k \subseteq \mathcal{D}_k$, with size b_k points, i.e.,

$$\mathbf{g}_k(\mathbf{w}_t) = \frac{1}{b_k}\sum_{i\in\mathcal{B}_k} \nabla f_k(\mathbf{w}_t; (\mathbf{u}_i^{(k)}, v_i^{(k)})) + \Omega\nabla R(\mathbf{w}_t), \tag{3}$$

where $\mathbf{g}_k(\mathbf{w}_t)$ is the stochastic gradient estimate of user k. The participants, i.e., $k \in \mathcal{K}_t$, next pre-process their $\mathbf{g}_k(\mathbf{w}_t)$ and obtains $\mathbf{x}_{k,t}$, which is subsequently send to the PS. The PS then receives $\mathbf{y}_t$ as defined in (1). Upon receiving $\mathbf{y}_t$, the PS performs post-processing on $\mathbf{y}_t$ to obtain $\hat{\mathbf{g}}_t$, the estimate of the true gradient $\mathbf{g}_t$ which is defined as,

$$\mathbf{g}_t = \frac{1}{\sum_{k=1}^K D_k}\sum_{k=1}^K D_k \nabla f_k(\mathbf{w}_t). \tag{4}$$

The global parameter $\mathbf{w}_t$ is updated using the estimated gradient $\hat{\mathbf{g}}_t$ according to the update rule $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\hat{\mathbf{g}}_t$, where η_t is the learning rate of the distributed SGD algorithm at iteration t. The iteration process continues until convergence.

Typically, in the wireless setting, post-processing is done at the PS to remove impact of the channel, and to ensure unbiased gradient estimates. Post-processing requires the PS to have knowledge of the channel condition, number of participants, and knowing how users are selected to participate. We assume that the PS has global CSI, and knows sampling probabilities $p_{k,t},\ \forall k, t$. However, the number of participants may or may not be known. Thus, in this work, we study both cases where (a) $\mathcal{K}_t$ is known, or (b) $\mathcal{K}_t$ is unknown at the PS.

Wireless FL with User Sampling: Here, we describe the per-iteration operation of the algorithm. At each iteration t, the PS transmits the model $\mathbf{w}_t$ to the users, and each user computes the local gradient using its local dataset (as in (3)). Each user k decides whether or not it wants to participate in the training according to probability $p_{k,t}$. Users then transmit their local

gradients with d channel uses of the wireless channel described in (1). The transmitted signal of user k at iteration t is given as:

$$\mathbf{x}_{k,t} = \begin{cases} \alpha_{k,t}\big(\mathbf{g}_k(\mathbf{w}_t) + \mathbf{n}_{k,t}\big), & \text{w.p. } p_{k,t} \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{n}_{k,t} \sim \mathcal{N}(0, \sigma_{k,t}^2 \mathbf{I}_d)$ is the artificial noise term to ensure privacy, and $\alpha_{k,t}$ is the scaling factor satisfying power constraint at each user. If a user is not participating, it does not transmit anything. 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. The parameters $\alpha_{k,t}$s and $\sigma_{k,t}$s are designed such that the power constraints are satisfied, i.e., $\mathbb{E}\left[\|\mathbf{x}_{k,t}\|_2^2\right] = \alpha_{k,t}^2 \left[\|\mathbf{g}_k(\mathbf{w}_t)\|^2 + d\sigma_{k,t}^2\right] \leq P_k$. From (1) and (5), the received signal at the PS is given as:

$$\mathbf{y}_t = \sum_{k\in\mathcal{K}_t} h_{k,t}\alpha_{k,t}\mathbf{g}_k(\mathbf{w}_t) + \sum_{k\in\mathcal{K}_t} h_{k,t}\alpha_{k,t}\mathbf{n}_{k,t} + \mathbf{m}_t. \tag{6}$$

In order to carry out the summation of the local gradients over-the-air, all users pick the coefficients $\alpha_{k,t}$s to align their transmitted local gradient estimates. Specifically, user k picks $\alpha_{k,t}$ so that

$$h_{k,t}\alpha_{k,t} = 1, \forall k \in \mathcal{K}_t. \tag{7}$$

The PS can perform two different post-processing to get unbiased gradient estimate $\hat{\mathbf{g}}_t$, i.e., $\mathbb{E}\left[\hat{\mathbf{g}}_t\right] = \mathbf{g}_t$ (see Appendix in [1]), based on the knowledge it has about $\mathcal{K}_t$:

Case (a)**:** When $\mathcal{K}_t$ is known at the PS, it obtains the unbiased gradient estimate $\hat{\mathbf{g}}_t$ as follows,

$$\begin{aligned} \hat{\mathbf{g}}_t &= \frac{1}{\zeta_t|\mathcal{K}_t|}\mathbf{y}_t \\ &= \frac{1}{\zeta_t|\mathcal{K}_t|}\sum_{k\in\mathcal{K}_t}\mathbf{g}_k(\mathbf{w}_t) + \frac{1}{\zeta_t|\mathcal{K}_t|}\left[\sum_{k\in\mathcal{K}_t}\mathbf{n}_{k,t} + \mathbf{m}_t\right], \end{aligned} \tag{8}$$

where $\zeta_t = 1 - \prod_{k=1}^{K}(1 - p_{k,t})$.

Case (b)**:** When $\mathcal{K}_t$ (thus $|\mathcal{K}_t|$) is unknown at the PS, it obtains the unbiased gradient estimate $\hat{\mathbf{g}}_t$ as follows,

$$\begin{aligned} \hat{\mathbf{g}}_t &= \frac{1}{\mu_{|\mathcal{K}_t|}}\mathbf{y}_t \\ &= \frac{1}{\mu_{|\mathcal{K}_t|}}\sum_{k\in\mathcal{K}_t}\mathbf{g}_k(\mathbf{w}_t) + \frac{1}{\mu_{|\mathcal{K}_t|}}\left[\sum_{k\in\mathcal{K}_t}\mathbf{n}_{k,t} + \mathbf{m}_t\right], \end{aligned} \tag{9}$$

where $\mu_{|\mathcal{K}_t|} = \mathbb{E}\left[|\mathcal{K}_t|\right] = \sum_{k=1}^{K} p_{k,t}$ is the expected number of participants in iteration t. The PS then updates the model. The process then repeats for T iterations.

Privacy Definitions: We assume the PS is honest but curious. It is honest in the sense that it follows the procedure accordingly, but it might learn sensitive information about users. Therefore, the wireless FedSGD algorithm should satisfy LDP constraints for each user. At the end of the training process, the PS may release the trained model to a third party. Thus, the training algorithm should provide central DP guarantees against any further post-processing or inference. The local and central privacy are formally defined as follows:

Definition 1. *($(\epsilon_\ell^{(k)}, \delta_\ell)$-LDP [29]) Let $\mathcal{X}_k$ be a set of all possible data points at user k. For user k, a randomized mechanism $\mathcal{M}_k : \mathcal{X}_k \to \mathbb{R}^d$ is $(\epsilon_\ell^{(k)}, \delta_\ell)$-LDP if for any $x,\ x' \in \mathcal{X}_k$, and any measurable subset $\mathcal{O}_k \subseteq$ Range$(\mathcal{M}_k)$, we have*

$$\Pr(\mathcal{M}_k(x) \in \mathcal{O}_k) \leq \exp\left(\epsilon_\ell^{(k)}\right)\Pr(\mathcal{M}_k(x') \in \mathcal{O}_k) + \delta_\ell. \tag{10}$$

The case of $\delta_\ell = 0$ is called pure $\epsilon_\ell^{(k)}$-LDP.

Definition 2. *((ϵ_c, δ_c)-DP [29]) Let $\mathcal{D} \triangleq \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_K$ be the collection of all possible datasets of all K users. A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{R}^d$ is (ϵ_c, δ_c)-DP if for any two neighboring datasets D, D' and any measurable subset $\mathcal{O} \subseteq$ Range$(\mathcal{M})$, we have*

$$\Pr(\mathcal{M}(D) \in \mathcal{O}) \leq \exp\left(\epsilon_c\right)\Pr(\mathcal{M}(D') \in \mathcal{O}) + \delta_c. \tag{11}$$

We refer to a pair of datasets $D, D' \in \mathcal{D}$ if D' can be obtained from D by removing one data element x_i for some $i \in [K]$. The case when $\delta_c = 0$ is called pure ϵ_c-DP.

III. Main Results & Discussions

A. Privacy Analysis for wireless FedSGD with User Sampling

In this section, we first derive the central privacy leakage for wireless FedSGD with user sampling. Specifically, we consider non-uniform sampling, where each user can be sampled according to a probability that depends on the channel conditions. We then study a special case, i.e., uniform sampling, to understand the asymptotic behavior of the central privacy as a function of the total number of users. In addition, we show that user sampling is also beneficial for the local privacy level. We also quantify the gain for the local privacy level achieved by user sampling and wireless aggregation where Gaussian mechanism is used at each sampled user. We note that the knowledge of $\mathcal{K}_t$ at the PS does not play a role in the proofs of the privacy guarantees due to the robustness of post-processing of DP. The privacy guarantee of the proposed wireless FedSGD with non-uniform sampling scheme is stated in the following Theorem:

Theorem 1. *(Non-uniform sampling) Suppose each user k participates in the training process at iteration t according to probability $p_{k,t}$, and utilizes local mechanism that satisfies $(\epsilon_{\ell,t}^{(k)}, \delta_\ell)$-LDP if they decided to participate. The central privacy level of the wireless FedSGD with user sampling at iteration t is given as*

$$\begin{aligned} \epsilon_{c,t} &\leq \log\left[1 + \frac{\max_k p_{k,t}}{1-\delta'}\left(e^{\frac{c}{\sqrt{\mu_{|\mathcal{K}_t|} - \beta K}}} - 1\right)\right], \\ \delta_{c,t} &= \delta' + \frac{\max_k p_{k,t}\delta_\ell}{1-\delta'}, \end{aligned} \tag{12}$$

for any $\delta' \in (2e^{-2\mu_{|\mathcal{K}_t|}^2/K}, 1)$ and $\beta = \frac{1}{\sqrt{K}}\sqrt{0.5\log\left(2/\delta'\right)}$, where $\mu_{|\mathcal{K}_t|} = \sum_{k=1}^{K} p_{k,t}$ denotes the expected number of users participating in iteration t, and $c = \frac{2L}{\sigma_{\min}}\sqrt{2\log(1.25/\delta_\ell)}$, where $\sigma_{\min} = \min_{k,t}\sigma_{k,t}$ and L is the Lipschitz constant for the loss function.

Proof Sketch: To derive analyze the central DP leakage at iteration t, we need to compare the distributions of the outputs seen at the PS via MAC for two cases: (a) when user k participates in training, and (b) when user k does not participate in the training. The existing privacy analysis for user sampling (with orthogonal transmissions) in [26], [27] cannot be directly applied to the current problem. The key challenge is that in each training iteration, the effective noise seen at the signal received by the PS over the wireless channel is a function of a random number ($\mathcal{K}_t$) of sampled users. To account for this randomness, we consider two sub-cases, one where $\mathcal{K}_t$ is close to its mean $\mu_{|\mathcal{K}_t|}$, and the complementary event. We bound the terms arising from these sub-cases individually using concentration inequalities, and then arrive at the central DP leakage result $\epsilon_{c,t}$ presented in Theorem 1 by taking the worst case bounds across all users k. The detailed proof can be found in [1].

The privacy parameters in Eq. (12) indicates that the central privacy leakage depends on the user with the highest sampling probability. Intuitively, a user with high sampling probability participates in the training process more often than other users, hence, having most impact on the central privacy leakage.

We note that (12) is a convex function of $\{p_{k,t}\}_{k=1}^{K}$ when $\epsilon_{\ell,t}^{(k)} \leq 1$. If the primary goal is to have strong privacy guarantee and does not need fast convergence, one can solve for the optimal sampling probabilities using the expression in (12). However, it is difficult to obtain a closed form solution of the optimal sampling probability for the non-uniform case. Due to convexity, one can still solve it numerically using convex solvers [30]. In contrast to the non-uniform case, one can solve for the optimal sampling probability for the uniform case analytically and obtain the following p_t^* by first setting $p_{k,t} = p_t, \forall k$ in (12), and obtain the following Lemma:

Lemma 1. *The optimal sampling probability that minimizes the central privacy level for the uniform case is given by*

$$p_t^* = \min\left[1, \frac{2}{\sqrt{K}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}\right]. \tag{13}$$

Using p_t^* and defining $c' = \sqrt{\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}$, one can obtain the following upper bound on the central privacy level,

$$\epsilon_c = \log\left[\frac{2c'}{\sqrt{K}(1-\delta')}\left(e^{\frac{c}{\sqrt[4]{K}\sqrt{c'}}} - 1\right) + 1\right] = \mathcal{O}\left(\frac{1}{K^{3/4}}\right)$$

From Lemma 1, we observe that the central privacy level behaves as $\mathcal{O}(1/K^{3/4})$ as opposed to the $\mathcal{O}(1/\sqrt{K})$ in [22] and [27]. Clearly, when both wireless aggregation and user sampling are employed, we can obtain additional benefit in terms of central privacy (see Table I and Fig. 2). Interestingly, the addition of user sampling in wireless FedSGD also provides benefit for LDP as shown in the following Lemma:

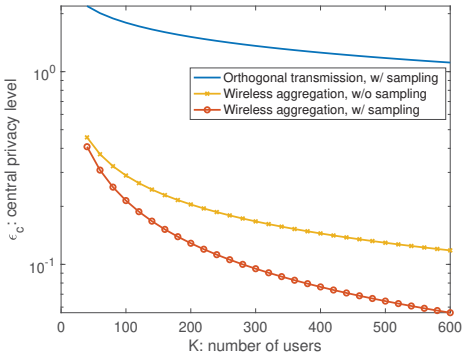

Fig. 2: Comparison for central privacy when $\sigma_{k,t} = 3$. The proposed scheme is shown to outperform other variants when $L = 1$, $N_0 = 3$, $\delta_l = \delta' = 10^{-4}$ and $p = \min\left[1, \frac{2}{\sqrt{K}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}\right]$.

Lemma 2. *For each user k, the proposed transmission scheme achieves $(\epsilon_{\ell,t}^{(k)}, p_{k,t}(\delta_\ell + \delta'))$-LDP per iteration, where*

$$\epsilon_{\ell,t}^{(k)} \leq \frac{1}{\sqrt{1+\kappa_t}} \times \frac{2L}{\sigma_{\min,t}}\sqrt{2\log\left(\frac{1.25}{\delta_\ell}\right)}, \tag{14}$$

where $\sigma_{\min,t} \triangleq \min_k \sigma_{k,t}$, $\kappa_t \triangleq \sum_{i=1,i\neq k}^{K} p_{i,t} - \beta K$, where β and δ' are defined in Theorem 1.

From Lemma 2, we observe the benefits of wireless aggregation. Asymptotically, LDP behaves like $\mathcal{O}(1/\sqrt{1+\kappa_t})$. In contrast, LDP achieved for orthogonal transmission scales as a constant, and does not decay with K. In the full version [1] of this paper, we present additional results on the total central leakage for the entire training process (T iterations) by using composition results for DP [31] [32].

IV. Experiments

In this section, we conduct experiments to assess the performance of the wireless FedSGD with user sampling on MNIST dataset for image classification. We model the instances of fading channels $h_{k,t}$'s via an autoregressive (AR) Rician model [33], where the Rician parameter $\Gamma = 5$ and the temporal correlation coefficient $\rho = 0.1$. The channel noise variance (receiver noise) is set as $N_0 = 1$. The user's transmit signal-to-noise ratio is defined as $\text{SNR}_k = \frac{P_k}{dN_0}$. We use $\sigma_{k,t}^2 = 0.1$ as the perturbation noise. Prior to sending the local gradient to the PS, each user clips the local gradient using the Lipschitz constant chosen empirically with test runs. We use $\delta_\ell = 10^{-5}$ and $\delta' = 2e^{-2\mu_{|\mathcal{K}_t|}^2/K} + 10^{-5}$ to satisfy the constraint on δ' and to avoid it from going to 0. We consider two different sampling schemes described as follows,

Uniform Sampling: Let $p_{k,t} = p, \ \forall k, t$ for any p.

Channel Aware Sampling: Each user computes $p_{k,t} = h_{k,t}/h_{\text{th}}$, where the threshold h_{th} is a hyperparameter which is optimized via cross-validation.

We train a single-layer neural network (with no hidden layer) using MNIST dataset [34], which consists of $60,000$ training and $10,000$ testing samples. The loss function we used is cross-entropy, and ADAM optimizer for training with

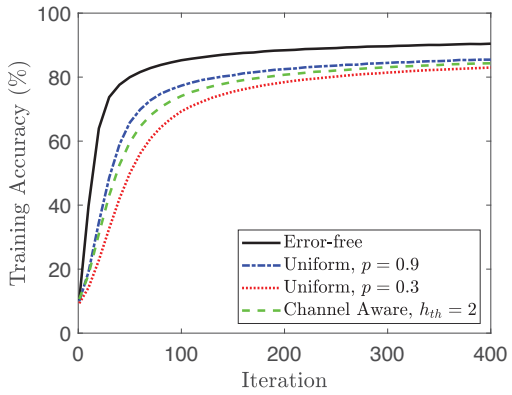

Fig. 3: The impact of the sampling probability on the training accuracy with $\sigma_{k,t}^2 = 0.1$, $L = 1$ and $T = 400$.

	Channel Aware	Uniform	
	$h_{\text{th}} = 2$	$p = 0.3$	$p = 0.9$
$\epsilon_{\ell,\max}$	3.675	5.124	2.46
$\epsilon_{c,\max}$	4.535	5.61	3.132
Avg. $\lvert\mathcal{K}\rvert$	96	60	180
Testing Acc.	85.27%	83.98%	86.42%

TABLE II: Comparison of privacy leakage per iteration with $\sigma_{k,t}^2 = 0.1$, $L = 1$ and $T = 400$ iterations. $\epsilon_{\ell,\max}$ and $\epsilon_{c,\max}$ are the maximum leakages across iterations.

a learning rate of $\eta = 0.001$. The training samples are evenly and randomly distributed across $K = 200$ users. Users are split into three groups where the first group consists of 68 users with $\text{SNR}_k = 2$ dB; the second and third group consist of 66 users in each group with $\text{SNR}_k = 10$ and 30 dB, respectively. We use $h_{\text{th}} = 2$ as the threshold for the channel aware sampling scheme. Empirically, the scaling factor is computed as follows,

$$\alpha_{k,t} = \min\left[\frac{1}{h_{k,t}}, \frac{\sqrt{P_k}}{\sqrt{\|\mathbf{g}_k(\mathbf{w}_t)\|^2 + d\sigma_{k,t}^2}}\right]. \tag{15}$$

In Fig. 3 and 4, we show the impact of sampling probability on the training accuracy. First, we observe that a higher p leads to a higher accuracy for the model. Next, in Table II, we observe that, for the uniform case with $L = 1$, the central DP leakage decreases as p increases, which contradicts with the intuition that higher p leads to higher leakage. However, let $p_{k,t} = p, \forall k, t$ in (12), i.e.,

$$\epsilon_{c,t} \leq \log\left[1 + \frac{p}{1-\delta'}\left(e^{\frac{c}{\sqrt{K(p-\beta)}}} - 1\right)\right], \tag{16}$$

we can see that the behavior of $\epsilon_{c,t}$ depends on two terms: $p/(1-\delta')$ and $\exp(c/\sqrt{K(p-\beta)})$. As p increases, the first term increases and the second term decreases. For a certain range of c, the second term dominates, therefore, $\epsilon_{c,t}$, as a whole, decreases. This is due to the fact that, since perturbation noises get aggregated over the wireless channel, the privacy enhances. Hence, users are encouraged to participate more when c is in that range. In general, c depends on $\sigma_{k,t}, L, \delta_\ell$, and c for Fig. 3 and Table II falls in the range that allows the second term to dominate as p increases. We also demonstrate the case when the first term dominates, i.e., $L = 0.1$ for this set of parameters. We can see that the central DP leakage

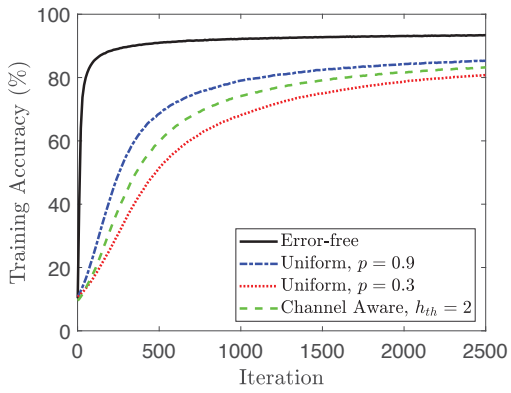

Fig. 4: The impact of the sampling probability on the training accuracy with $\sigma_{k,t}^2 = 0.1$, $L = 0.1$ and $T = 2500$.

	Channel Aware	Uniform	
	$h_{\text{th}} = 2$	$p = 0.3$	$p = 0.9$
$\epsilon_{\ell,\max}$	0.3677	0.5124	0.2460
$\epsilon_{c,\max}$	0.3642	0.2258	0.2317
Avg. $\lvert\mathcal{K}\rvert$	96	60	180
Testing Acc.	84.33%	81.76%	86.25%

TABLE III: Comparison of privacy leakage per iteration with $\sigma_{k,t}^2 = 0.1$, $L = 0.1$ and $T = 2500$ iterations.

increases as p increases from Table III. When c is in this range, the amplification of privacy is not enough to outweigh the disadvantage of participating more. Thus, the intuition that higher p leads to higher leakage holds. From Table II and III, we can see that channel aware sampling achieves 85.27% and 84.33% testing accuracy, which is lower than those of uniform sampling with $p = 0.9$. This is due to the choice of h_{th}. By reducing h_{th}, we can improve the accuracy of the channel aware sampling. Another interesting observation is that, while channel aware sampling suffers slightly from higher central DP leakages, it does achieve relatively high testing accuracy and good local DP leakages with significant less average number of participants compare to uniform sampling with $p = 0.9$. We refer the readers to the full version of this paper [1] for more discussions and experiments.

V. Conclusions

In this work, we showed the privacy benefits of user sampling and wireless aggregation for federated learning. The resulting leakage for central DP was shown to scale as $O(1/K^{3/4})$, improving upon prior results on this topic. As a future work, we would like to study other variations of FL such as FedAvg, where each user performs local model updates through multiple SGD computations, followed by model exchange with the PS. Another interesting direction would be to consider scenarios where the sampling probabilities can depend on the local gradients/losses. These scenarios may require new techniques for privacy analysis than the ones used in this paper, where sampling probabilities are independent of the local data (gradients/local loss function).

REFERENCES

[1] M. Seif, W.-T. Chang, and R. Tandon, "Privacy amplification for federated learning via user sampling and wireless aggregation," *arXiv preprint arXiv:2103.01953*, 2021.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[9] J. Chen and R. Luss, "Stochastic gradient descent with biased but consistent gradient estimators," *arXiv preprint arXiv:1807.11880*, 2018.

[10] 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.

[11] ——, "Federated learning over wireless fading channels," *arXiv preprint arXiv:1907.09769*, 2019.

[12] W.-T. Chang and R. Tandon, "Communication efficient federated learning over multiple access channels," *arXiv preprint arXiv:2001.08737*, 2020.

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

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

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

[21] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *arXiv preprint arXiv:2104.02151*, 2021.

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

[23] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *arXiv preprint arXiv:2006.05459*, 2020.

[24] A. Sonee and S. Rini, "Efficient federated learning over multiple access channel with differential privacy constraints," *arXiv preprint arXiv:2005.07776*, 2020.

[25] 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

[26] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," *Advances in Neural Information Processing Systems*, vol. 31, pp. 6277–6287, 2018.

[27] B. Balle, P. Kairouz, H. B. McMahan, O. Thakkar, and A. Thakurta, "Privacy amplification via random check-ins," *arXiv preprint arXiv:2007.06605*, 2020.

[28] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *IEEE Symposium on Security and Privacy (S&P)*, 2017, pp. 58–77.

[29] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, S. Song, K. Talwar, and A. Thakurta, "Encode, shuffle, analyze privacy revisited: formalizations and empirical evaluation," *arXiv preprint arXiv:2001.03618*, 2020.

[30] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2009.

[31] 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.

[32] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *International conference on machine learning*. PMLR, 2015, pp. 1376–1385.

[33] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[34] Y. LeCun, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1998.