Privacy Amplification for Federated Learning via User Sampling and Wireless Aggregation

Mohamed Seif Eldin Mohamed, *Graduate Student Member, IEEE*, Wei-Ting Chang, *Graduate Student Member, IEEE*, and Ravi Tandon, *Senior Member, IEEE*

***Abstract*— In this paper, we study the problem of federated learning over a wireless channel with user sampling, modeled by a fading 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 $\mathcal{O}(1/K^{1/2})$, 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 privacy 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. We also present analysis for the convergence rate of the proposed scheme and study the tradeoffs between wireless resources, convergence, and privacy theoretically and empirically for two scenarios when the number of sampled participants are (a) known, or (b) unknown at the parameter server.**

***Index Terms*— Federated learning, wireless aggregation, differential privacy, user sampling.**

I. Introduction

FEDERATED learning (FL) [1] is a framework that enables multiple users to jointly train a machine learning (ML) 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 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 demonstrated in recent works [2]–[8]. 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 [9]–[20] have focused on exploiting the wireless channel to alleviate the communication overhead of FL. 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. Digital schemes were proposed in [9]–[11], where in [9] the gradient vectors are first sparsified and quantized locally at the users by setting the desired number of top elements in magnitude to one value before transmissions. In [10], the authors modify the digital scheme in [9] to allow only the user with the best channel condition to transmit. In [11], the authors tailor the quantization scheme to the capacity region of the underlying MAC, which allows the gradient vectors to be quantized according to both informativeness of the gradients and the channel conditions. However, digital schemes require the PS to decode individual gradients and then aggregate them.

For analog schemes, on the other hand, gradients are 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. There have been several recent works focusing on the design of analog schemes for wireless FL. In [12], [13], wireless aggregation is done by aligning the gradients through power control or beamforming. The communication efficiency is further enhanced by incorporating user scheduling. In addition to power control, [9], [10], [14] project the gradients to lower dimension prior to transmissions to improve communication efficiency, where [14] also utilizes user scheduling and only allows users with good channel conditions to transmit. In [15], the authors focus on minimizing the energy consumption of users in wireless FL by formulating and solving an optimization problem

Manuscript received February 26, 2021; revised September 10, 2021; accepted September 22, 2021. Date of publication October 6, 2021; date of current version November 22, 2021. This work was supported in part by the NSF under Grant CAREER 1651492, Grant CNS 1715947, and Grant CCF 2100013; and in part by the 2018 Keysight Early Career Professor Award. This article was presented in part at the IEEE International Symposium on Information Theory (ISIT) 2021. *(Corresponding author: Mohamed Seif Eldin Mohamed.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Arizona, Tucson, AZ 85721 USA (e-mail: mseif@email.arizona.edu; wchang@email.arizona.edu; tandonr@arizona.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2021.3118408.

Digital Object Identifier 10.1109/JSAC.2021.3118408

subject to latency constraints. In [16], the authors proposed a gradient-based multiple access algorithm that let users transmit analog functions using common shape waveforms to mitigate the impact of fading. In [17], the authors provide convergence analysis for wireless FL with non-i.i.d. data. Based on the bound on the convergence rate, the authors of [17] optimize the frequency of global aggregation based on the data, model, and system dynamics.

There is a large body of recent work focusing on the design of differentially private FL. Differential privacy (DP) [21] 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 [22]. In the literature, there have been several research efforts to design FL algorithms satisfying LDP [23], [24], which require significant amount of perturbation noise to ensure privacy guarantees. However, the amount of noise can be further reduced when employing user sampling [25], where users are sampled by the PS to participate in the training in each iteration. However, sampling schemes can be challenging in practice since they require coordination between the PS and users, and may not be feasible if the PS is untrustworthy. Hence, decentralized sampling schemes that do not depend on the PS for coordination are desirable. To reduce the dependency on the PS, Balle *et. al.* [26] recently proposed a Random Check-in protocol. More specifically, users have the choice to decide whether or not to participate in the training process, and when to participate during the training process. It is worth noting that the above works focus on orthogonal transmission and do not take the impact of the communication channels into account while performing privacy analysis.

In addition to saving bandwidth and computation, it has been shown in [27]–[29] that wireless FL also naturally provides strong differential privacy (DP) [30] guarantees. Specifically, in [27], 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 [25] that one can obtain a similar scaling of $\mathcal{O}(1/\sqrt{K})$ for privacy leakage through user sampling. The scheme of [25], however, considers orthogonal transmission from the sampled users.

One natural question to ask is *whether one could provide even stronger privacy guarantees by incorporating user sampling to the private wireless FedSGD scheme. 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 fading multiple access channels (MACs), subject to LDP and DP constraints. We propose a 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 [25], [26] 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 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. To the best of our knowledge, this is one of the first results on wireless FedSGD with LDP and DP constraints with user sampling (see Table I for comparison).

TABLE I

COMPARISON FOR CENTRAL PRIVACY UNDER: (1) ORTHOGONAL AND (2) WIRELESS AGGREGATION TRANSMISSIONS

Transmission scheme	Without sampling	With sampling
Orthogonal	$\mathcal{O}(1)$ [31]	$\mathcal{O}(1/\sqrt{K})$ [25]
Wireless Aggregation	$\mathcal{O}(1/\sqrt{K})$ [27]	$\mathcal{O}(1/K^{3/4})$ (Lemma 1)

We would also like to mention a recent concurrent work [32], in which the authors studied the impact of user sampling on central DP for wireless FL. Moreover, they have proposed a wireless transmission scheme that is also robust against CSI attacks from the PS. It is assumed in that the identities of sampled users are shared between participating devices through a side channel, and never shared with the PS. While the problem is similar in spirit, the main differences of work compared to their are: 1) In our system, we do not require the users to share information about participation in any round. 2) We study both local and central DP guarantees and the associated tradeoffs (including scaling laws) as a function of users. 3) We also present convergence rates analysis for the proposed learning algorithm.

Notations: Boldface uppercase letters denote matrices (e.g., $\mathbf{A}$), boldface lowercase letters are used for vectors (e.g., $\mathbf{a}$), we denote scalars by non-boldface lowercase letters (e.g., x), and sets by capital calligraphic letters (e.g., $\mathcal{X}$). $[K] \triangleq [1, 2, \cdots, K]$ represents the set of all integers from 1 to K. The set of natural numbers, integer numbers, real numbers and complex numbers are denoted by $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$, respectively.

II. SYSTEM MODEL

A. Wireless Channel Model

We consider a single-antenna wireless FL system with K users and a central PS. Users are connected to the PS through a fading 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

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

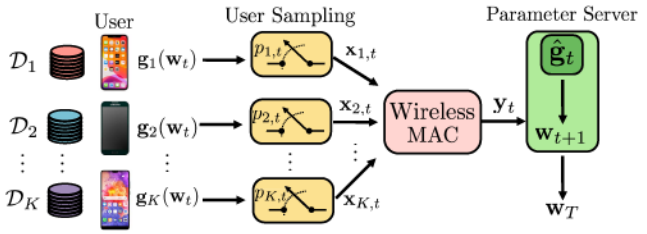


Fig. 1. Illustration of the private wireless FedSGD framework: Users collaborate with the PS to jointly train a machine learning model over a fading MAC.

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, $h_{k,t} \geq 0$ is the channel coefficient between the k-th user 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. 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 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 FedSGD where every user participates in the training.

For this work, we consider (a) time-invariant uniform sampling, where the sampling probability remains the same across users and iterations; (b) time-varying uniform sampling, where the sampling probability remains the same across users but varies across iterations; and (c) channel aware sampling, where sampling probabilities for each user can depend on the local channel gain between the user and the PS. We note that sampling strategies based on gradients or losses can potentially leak information about local datasets, hence, require analysis for privacy. Thus, we leave gradient-based sampling strategies to future work.

B. 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

$$f_k(\mathbf{w}) = \frac{1}{D_k} \sum_{i=1}^{D_k} f(\mathbf{w}; \mathbf{u}_i^{(k)}, v_i^{(k)}) + \Omega R(\mathbf{w}),$$

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 fading MAC described above in order to train a model by minimizing the loss function $F(\mathbf{w})$, i.e.,

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

The minimization of $F(\mathbf{w})$ is carried out iteratively through a distributed stochastic gradient descent (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 (i.e., $|\mathcal{B}_k| = b_k$), i.e.,

$$\mathbf{g}_k(\mathbf{w}_t) = \frac{1}{b_k} \sum_{i \in \mathcal{B}_k} \nabla f_k(\mathbf{w}_t; (\mathbf{u}_i^{(k)}, v_i^{(k)})) + \Omega \nabla R(\mathbf{w}_t), \quad (2)$$

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 $\mathbf{x}_{k,t}$, as explained below. Then, the participants send their $\mathbf{x}_{k,t}$'s to the PS, where the PS 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,

$$\mathbf{g}_t = \frac{1}{\sum_{k=1}^K D_k} \sum_{k=1}^K D_k \mathbf{g}_k(\mathbf{w}_t). \quad (3)$$

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

Typically, in the wireless setting, the post-processing done at the PS involves removing channel effects, averaging the aggregated local gradients, and/or multiplying a constant to maintain the unbiasedness. These post-processing steps depend on the PS's knowledge of the channel condition, number of participants, and knowing how users are selected to participate. As mentioned above, the PS has global CSI. In addition, we assume that the PS knows the sampling probabilities $p_{k,t}$, $\forall k, t$. However, the number of participants may or may not be known at the PS. Thus, in this work, we study both cases, where (a) $|\mathcal{K}_t|$ is known, and (b) $|\mathcal{K}_t|$ is unknown, at the PS.

C. Wireless FL With User Sampling

The training continues for a total of T iterations, where the users are synchronized with the PS. Here, we describe the per-iteration operation of the algorithm. At the beginning of 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 according to (2). Each user k participates in the training with probability $p_{k,t}$. Users then transmit their local gradients with d channel uses of the wireless channel described in (1) in the pre-determined time slot. The transmitted signal of user k at iteration t is given as:

$$\mathbf{x}_{k,t} = \begin{cases} \alpha_{k,t} \left(\mathbf{g}_k(\mathbf{w}_t) + \mathbf{n}_{k,t}\right), & \text{w.p. } p_{k,t} \\ \mathbf{0}, & \text{otherwise} \end{cases} \quad (4)$$

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 (4), the received signal at the PS can be written as:

$$\mathbf{y}_t = \sum_{k\in\mathcal{K}_t} h_{k,t}\alpha_{k,t}\mathbf{g}_k(\mathbf{w}_t) + \underbrace{\sum_{k\in\mathcal{K}_t} h_{k,t}\alpha_{k,t}\mathbf{n}_{k,t} + \mathbf{m}_t}_{\mathbf{z}_t}, \quad (5)$$

where $\mathbf{z}_t \sim \mathcal{N}(0, \sigma_{z_t}^2\mathbf{I}_d)$ is the effective noise, and $\sigma_{z_t}^2 = \sum_{k\in\mathcal{K}_t} h_{k,t}^2\alpha_{k,t}^2\sigma_{k,t}^2 + N_0$. In order to carry out the summation of the local gradients over-the-air, all users pick the coefficients $\alpha_{k,t}$s in order to align their transmitted local gradient estimates. Specifically, user k picks $\alpha_{k,t}$ so that

$$h_{k,t}\alpha_{k,t} = \gamma_t, \quad \forall k \in \mathcal{K}_t, \quad (6)$$

where γ_t is the alignment constant picked for ensuring that the power constraints are satisfied. For the alignment scheme described above, the received signal at the PS at iteration t in (10) simplifies to $\mathbf{y}_t = \sum_{k\in\mathcal{K}_t}\gamma_t\mathbf{g}_k(\mathbf{w}_t) + \mathbf{z}_t$. The PS can perform two different post-processing operations to get unbiased gradient estimate $\hat{\mathbf{g}}_t$, i.e., $\mathbb{E}[\hat{\mathbf{g}}_t] = \mathbf{g}_t$ (see Appendix E), based on the knowledge it has: (a) when $|\mathcal{K}_t|$ is known at the PS; (b) when $|\mathcal{K}_t|$ is unknown at the PS.

Case (a)**:** When $|\mathcal{K}_t|$ is known at the PS, it performs an update when $|\mathcal{K}_t| \neq 0$. It obtains the gradient estimate $\hat{\mathbf{g}}_t$ by dividing the received signal by the alignment constant, ζ_t and the number of participants. When $|\mathcal{K}_t| = 0$, $\hat{\mathbf{g}}_t$ is set to $\mathbf{0}$, and an update is skipped. Hence, the gradient estimate is given as,

$$\hat{\mathbf{g}}_t = \begin{cases} \dfrac{1}{\gamma_t\zeta_t|\mathcal{K}_t|}\mathbf{y}_t, & \text{if } |\mathcal{K}_t| \neq 0, \\ \mathbf{0}, & \text{if } |\mathcal{K}_t| = 0, \end{cases} \quad (7)$$

where

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

and $\zeta_t = 1 - \prod_{k=1}^{K}(1-p_{k,t})$ is chosen to ensure unbiasedness of the estimated aggregated gradient.

Case (b)**:** When $|\mathcal{K}_t|$ is unknown at the PS, it obtains the unbiased gradient estimate $\hat{\mathbf{g}}_t$ by dividing the received signal by the alignment constant and the expected number of participants as follows,

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

where $\mu_{|\mathcal{K}_t|} = \mathbb{E}[|\mathcal{K}_t|] = \sum_{k=1}^{K} p_{k,t}$ is the expected number of participants in iteration t. The PS then update the models and repeats this process for T iterations. We summarize our transmission scheme in Algorithm 1.

Remark 1: One can divide the local gradient estimate $\mathbf{g}_k(\mathbf{w}_t)$ *by* $p_{k,t}$ *to obtain unbiased local estimate for the full gradient* $\mathbf{g}_t$*, i.e.,*

$$\mathbf{x}_{k,t} = \begin{cases} \alpha_{k,t} \times \left(\dfrac{\mathbf{g}_k(\mathbf{w}_t)}{p_{k,t}} + \mathbf{n}_{k,t}\right), & \textit{w.p. } p_{k,t} \\ \mathbf{0}, & \textit{otherwise.} \end{cases} \quad (9)$$

The received signal at the PS at iteration t *is*

$$\mathbf{y}_t = \gamma_t\sum_{k\in\mathcal{K}_t}\frac{\mathbf{g}_k(\mathbf{w}_t)}{p_{k,t}} + \underbrace{\sum_{k\in\mathcal{K}_t} h_{k,t}\alpha_{k,t}\times\mathbf{n}_{k,t} + \mathbf{m}_t}_{\mathbf{z}_t}, \quad (10)$$

It can be readily shown that the PS gets a sum of unbiased local gradient estimates. However, this approach requires more pre-processing at the user which further limits the transmit power scaling, i.e.,

$$\alpha_{k,t}^2 \leq \frac{P_k}{1/p_{k,t}^2 \times \|\mathbf{g}_k(\mathbf{w}_t)\|^2 + d\sigma_{k,t}^2}. \quad (11)$$

In contrast, our scheme requires only the full gradient to be unbiased and in this case the power constraint is more relaxed, i.e.,

$$\alpha_{k,t}^2 \leq \frac{P_k}{\|\mathbf{g}_k(\mathbf{w}_t)\|^2 + d\sigma_{k,t}^2}. \quad (12)$$

Algorithm 1 Differentially Private Wireless FedSGD Scheme With User Sampling

1: Initialize $\mathbf{w}_1$ at the PS;
2: **for** iteration $t = 1, \ldots, T$ **do**
3: PS broadcasts the global model $\mathbf{w}_t$ to all users;
4: **for** each user in parallel **do**
5: Compute $\mathbf{g}_k(\mathbf{w}_t)$ according to (4);
6: Transmit $\mathbf{x}_{k,t} = \alpha_{k,t}\left(\mathbf{g}_k(\mathbf{w}_t) + \mathbf{n}_{k,t}\right)$ with probability $p_{k,t}$ and $\mathbf{x}_{k,t} = 0$ otherwise to the PS;
7: **end for**
8: PS receives $\mathbf{y}_t$ and recovers $\hat{\mathbf{g}}_t$ according to (9) for known $|\mathcal{K}_t|$ case and (10) for unknown $|\mathcal{K}_t|$ case;
9: PS updates global model $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\hat{\mathbf{g}}_t$;
10: **end for**
11: PS returns $\mathbf{w}_{t+1}$;

D. Privacy Definitions

We assume that the PS is honest but curious. It is honest in the sense that it follows the FL procedure faithfully, but it might be interested in learning sensitive information about users. Therefore, the SGD algorithm for wireless FL 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 DP are formally defined as follows:

Definition 1 ($(\epsilon_\ell^{(k)}, \delta_\ell)$*-LDP [33]): 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*

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

The setting when $\delta_\ell = 0$ *is referred as pure* $\epsilon_\ell^{(k)}$*-LDP.*

Definition 2 ((ϵ_c, δ_c)*-DP [33]): 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

$$\Pr(\mathcal{M}(D) \in \mathcal{O}) \leq \exp(\epsilon_c) \Pr(\mathcal{M}(D') \in \mathcal{O}) + \delta_c. \quad (13)$$

We refer to a pair of datasets $D, D' \in \mathcal{D}$ if D' can be obtained from D by removing the whole dataset of a user k. The setting when $\delta_c = 0$ is referred as 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 DP leakage for wireless FedSGD with user sampling. Specifically, we consider two sampling strategies: (a) non-uniform sampling; and (b) both time-varying and none time-varying uniform sampling. For non-uniform sampling, 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. The privacy guarantee of the proposed wireless FedSGD with non-uniform sampling 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 $(\epsilon_{\ell,t}^{(k)}, \tilde{\delta}_\ell)$-LDP if they decided to participate. The central privacy level of the wireless FedSGD with user sampling at iteration t is given as

$$\begin{aligned}\epsilon_{c,t} &\leq \log\left[1 + \frac{\max_k p_{k,t}}{1-\delta'}\left(e^{\max_{k,t}\epsilon_{\ell,t}^{(k)}} - 1\right)\right] \\ &\overset{(a)}{=} \log\left[1 + \frac{\max_k p_{k,t}}{1-\delta'}\left(e^{\frac{c}{\sqrt{\mu_{|\mathcal{K}_t|} - \beta K}}} - 1\right)\right], \\ \delta_{c,t} &= \delta' + \frac{\max_k p_{k,t}\delta_\ell}{1-\delta'}, \end{aligned} \quad (14)$$

for any $\delta' \in (2e^{-2\mu_{|\mathcal{K}_t|}^2 K}, 1)$ and $\beta = \frac{1}{\sqrt{K}}\sqrt{0.5\log(2/\delta')}$, 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. In (14), $\epsilon_{\ell,t}^{(k)}$ is the effective local privacy level of user k due to sampling and wireless aggregation. Step (a) follows from the Gaussian mechanism which will become clearer in the sequel.

The proof of the Theorem can be found in Appendix B. The privacy parameters in (14) 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 with lower probabilities, thereby having most impact on the central privacy leakage. For the case with uniform sampling probability, the privacy parameters can be simplified to the following (the proof of Corollary follows directly from Theorem 1):

Corollary 1 (Uniform Sampling): Suppose each user decides to participate with probability $p_{k,t} = p_t$, and the local mechanism satisfies $(\epsilon_{\ell,t}^{(k)}, \tilde{\delta}_\ell)$-LDP for each user k. The central privacy level of the wireless FedSGD with user sampling is given as

$$\begin{aligned}\epsilon_{c,t} &\leq \log\left[1 + \frac{p_t}{1-\delta'}\left(e^{\frac{c}{\sqrt{K(p_t-\beta)}}} - 1\right)\right], \\ \delta_{c,t} &= \delta' + \frac{p_t\delta_\ell}{1-\delta'}, \end{aligned} \quad (15)$$

for any $\delta' \in (2e^{-2p^2K}, 1)$ and $\beta = \frac{1}{\sqrt{K}}\sqrt{0.5\log(2/\delta')}$, where $c = \frac{2L}{\sigma_{\min}}\sqrt{2\log(1.25/\delta_\ell)}$.

We note that both (14) (respectively, (15)) is a convex function of $\{p_{k,t}\}_{k=1}^K$ (respectively, p_t) 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 expressions in (14) and (15). 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. In contrast to the non-uniform case, one can solve for the optimal sampling probability for the uniform case as stated in the following Lemma.

Lemma 1: For any $p_t > \beta, \beta = \frac{1}{\sqrt{K}}\sqrt{0.5\log(2/\delta')}$, the optimal sampling probability that minimizes the upper bound on ϵ_c in (14) is given by

$$p_t^* = \min\left[1, \frac{2}{\sqrt{K}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}\right] \quad (16)$$

for sufficiently large K. By plugging p_t^* back into (15), one can obtain the following upper bound on the central DP,

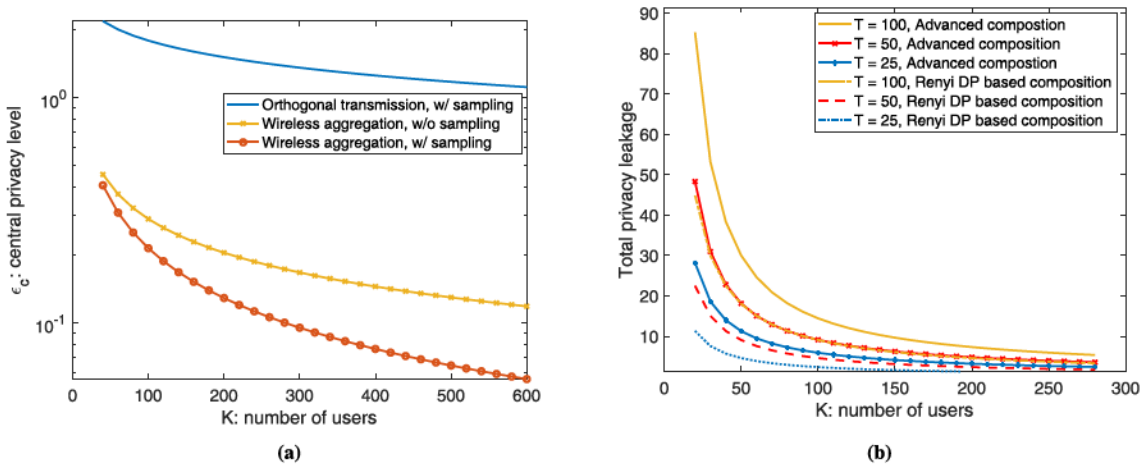
$$\begin{aligned}\epsilon_c &= \log\left[\frac{2\sqrt{\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}}{\sqrt{K}(1-\delta')}\left(e^{\frac{c}{\sqrt[4]{K\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}}} - 1\right) + 1\right] \\ &= \mathcal{O}\left(\frac{1}{K^{3/4}}\right). \end{aligned}$$

The proof of Lemma 1 is presented in Appendix C. From Lemma 1, we observe that the central privacy level behaves as $\mathcal{O}(1/K^{3/4})$ as opposed to the $\mathcal{O}(1/\sqrt{K})$ for wireless FL without sampling [27] and $\mathcal{O}(1/\sqrt{K})$ for FL with orthogonal transmission and user sampling [26] (see Table I). Clearly, when both wireless aggregation and user sampling are employed, we can obtain additional benefit in terms of central privacy. We also plot the central privacy level of the proposed scheme against other variations (see Fig. 2a).

We next analyze the local privacy level achieved by the FedSGD transmission scheme.

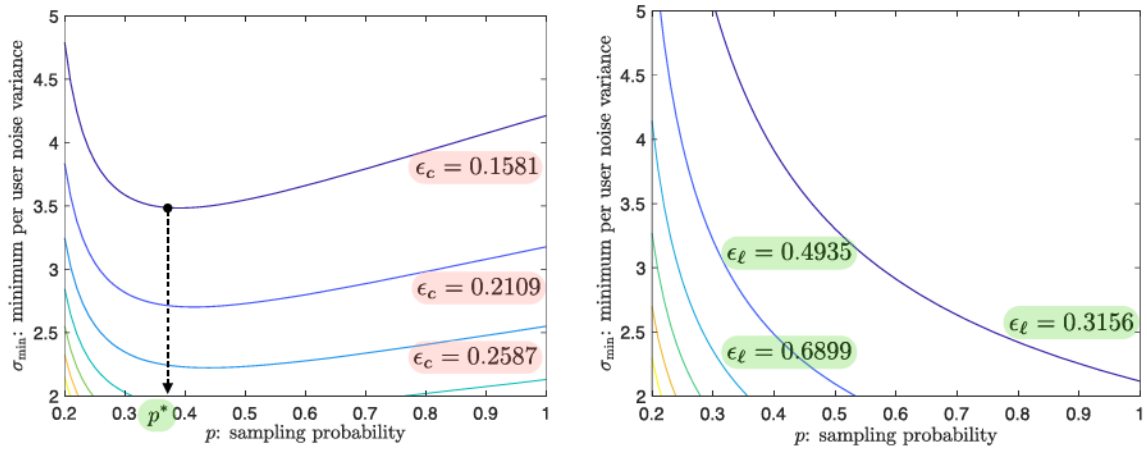
Lemma 2: For each user k, the proposed transmission scheme achieves $(\epsilon_{\ell,t}^{(k)}, p_{k,t}(\delta_\ell + \delta'))$-LDP per iteration, where

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



(a) (b)

Fig. 2. (a) Comparison for central privacy, where wireless aggregation with sampling is shown to outperform other variants; (b) Total privacy leakage as a function of K, number of users for different values of T, the number of training iterations, where $L = 1$, $\sigma_{k,t} = N_0 = 3$, $\gamma_t = 1$, $\delta_l = \delta' = 10^{-4}$ and $p = \frac{2}{\sqrt{K}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}$ for both figures.



(a) (b)

Fig. 3. (a) Contour plot of central DP leakage as a function of $\sigma_{\min}$ and p. It can be seen that there are many operating points to achieve a specific central DP, and there exists a tradeoff between $\sigma_{\min}$ and p for a fixed ϵ_c; (b) Contour plot of LDP leakage as a function $\sigma_{\min}$ and p, where the local leakage is a monotonically decreasing function of p, where $L = 1$, $\delta_l = \delta' = 10^{-4}$ and $K = 200$ for both figures.

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.*

The proof is presented in Appendix D.

Remark 2: From Lemma 2, we can observe the privacy benefits of wireless gradient aggregation. Asymptotically, the local privacy level behaves like $\mathcal{O}(1/\sqrt{1+\kappa_t})$. *In contrast, the local privacy achieved by orthogonal transmission scales as a constant, and does not decay with* K *[27].*

Remark 3: In our privacy analysis we assume that the sampling probability is strictly greater than zero, i.e., $p_t > 0$. *We also assume that* $p_t > \beta$, *so that the exponent term is non-negative. These additional assumptions (lower bounds on sampling probability) can also be interpreted as indirect constraints over utility (e.g., predictive accuracy of the trained model).*

In this paper, we have considered both (stronger) LDP and (weaker) central DP privacy. If one only considers the stronger LDP, then it is clear as shown in Fig. 3b, that to achieve a certain local privacy budget ϵ_ℓ, *the amount of noise each user adds is a decreasing function of the sampling probability. The intuition behind this is that as sampling probability increases, the wireless channel also allows the artificial noises of the sampled users to aggregate, thus providing a boost in LDP. On the other hand, let us now consider the (weaker) central DP. For a central privacy leakage budget of* ϵ_c, *as shown in Fig. 3a, the amount of noise added by each user is interestingly a non-monotonic function of the sampling probability.*

With the utility constraints (lower bounds on sampling probability) in place, depending on other parameters (e.g., Lipschitz constant, δ_ℓ, δ', K*), the optimal probability that provides the strongest central DP guarantee might not be the minimum possible probability (i.e.,* $p_t = \beta$*) anymore. For example, for* $L = 1$, $\delta_l = \delta' = 10^{-4}$ *and* $K = 200$, *we can see in Fig. 3a that there are many ways to achieve certain*

ϵ_c. One can then tune $\sigma_{\min}$ and p_t for a specific scenario. As an example, in a power-constrained setting, one would like to keep $\sigma_{\min}$ as small as possible, then one would pick the p^* as shown in 3a. This would however, result in higher LDP leakage. Alternatively, one can consider increasing p_t to achieve better LDP at the expense of adding more noise.

While Theorem 1 shows the per-iteration leakage, we can use advanced composition results for DP using the Gaussian mechanism to obtain the total privacy leakage when the wireless FL algorithm is used for T iterations. When the sampling probability is time-varying, using existing results in [34], it can be readily shown that the total leakage over T iterations of the proposed scheme is $(\epsilon_c^{(T)}, \delta_c^{(T)})$-DP for $\tilde{\delta} \in (0,1]$ where $\epsilon_c^{(T)}$ and $\delta_c^{(T)}$ can be found as follows,

$$\epsilon_c^{(T)} = \sum_{t\in[T]} \frac{(e^{\epsilon_{c,t}}-1)\epsilon_{c,t}}{(e^{\epsilon_{c,t}}+1)} + \sqrt{2\log(\frac{1}{\tilde{\delta}})\sum_{t\in[T]}\epsilon_{c,t}^2}, \quad (17)$$

$$\overset{(a)}{\leq} \left(e^{\sqrt{\frac{c}{\min_t \mu_{|\mathcal{K}_t|}-\beta K}}}-1\right)^2 \times \frac{\sum_{t\in[T]}(\max_k p_{k,t})^2}{2(1-\delta')^2} + \sqrt{2\log(\frac{1}{\tilde{\delta}})}\left(e^{\sqrt{\frac{c}{\min_t \mu_{|\mathcal{K}_t|}-\beta K}}}-1\right) \times \frac{\sqrt{\sum_{t\in[T]}(\max_k p_{k,t})^2}}{1-\delta'}, \quad (18)$$

where step (a) follows from the fact that $e^x + 1 \geq 2$, where $x \geq 0$ and $\log(1+x) \leq x$. Also,

$$\delta_c^{(T)} = 1-(1-\tilde{\delta})\prod_{t=1}^{T}(1-\delta_{c,t}) = 1-(1-\tilde{\delta})\prod_{t=1}^{T}\left(1-\left(\delta'+\frac{\max_k p_{k,t}\delta_\ell}{1-\delta'}\right)\right)$$

By examining the expression in (18), we can see that, for a given T, $\min_t \mu_{|\mathcal{K}_t|} - \beta K$ grows as K increases. Therefore, the exponential term approaches 1 as K increases, and (18) goes to 0 as the number of users increases. For the case when the sampling probability is time-invariant, using existing results in [35], it can be readily shown that the total leakage over T iterations of the proposed scheme is $(\epsilon_c^{(T)}, T\delta_c + \tilde{\delta})$-DP for $\tilde{\delta} \in (0,1]$ where $\epsilon_c^{(T)} = \sqrt{2\,T\log(1/\tilde{\delta})}\epsilon_c + T\epsilon_c(e^{\epsilon_c}-1)$. We can expect the same behavior to hold true for the time-invariant case since the result in [34] is more general than the result in [35]. We illustrate the total central privacy leakage for the uniform sampling time-invariant case as a function of K in Fig. 2b for various values of T. As is clearly evident, the leakage provided by wireless FedSGD goes asymptotically to 0 as $K \to \infty$. It is worth noting that total privacy leakage over T iterations can be further tightened using existing techniques such as Rényi DP composition [36] (see Fig. 2b). More specifically, after T iterations, the leakage in this case will be

$$\epsilon_c^{(T)} = \frac{T\epsilon_c + \log(1/\delta_c)}{\alpha - 1}, \quad \delta_c^{(T)} = \delta_c,$$

where α is a hyper-parameter that typically ranges 1 to 64 [37].

B. Convergence Rate of Private FL

In this section, we analyze the performance of private wireless FedSGD under the assumption that the global loss function $F(\mathbf{w})$ is smooth and strongly convex,[1] and the data across users is i.i.d. Specifically, we consider two scenarios when (a) $|\mathcal{K}_t|$ is unknown and (b) $|\mathcal{K}_t|$ is known to the PS. We take both privacy and wireless aggregation into account while deriving the bounds. Interestingly, we show that the unknown $|\mathcal{K}_t|$ case always outperforms the known $|\mathcal{K}_t|$ case. Therefore, it is not necessary for the PS to know $|\mathcal{K}_t|$. We confirm this observation in the experiment section as well. Due to privacy requirements and noisy nature of wireless channel, the convergence rate is penalized as shown in the following Theorem.

Theorem 2 (Unknown $|\mathcal{K}_t|$ With Non-Uniform Sampling): 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 FedSGD algorithm is

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T^2}\sum_{t=1}^{T}\left[\frac{L^2\left(\mu_{|\mathcal{K}_t|}^2 + \sigma_{|\mathcal{K}_t|}^2\right)}{\gamma_t^2 \mu_{|\mathcal{K}_t|}^2} + \frac{d}{\gamma_t^2 \mu_{|\mathcal{K}_t|}^2}\left[\max_k \sigma_{k,t}^2 \times \mu_{|\mathcal{K}_t|} + N_0\right]\right], \quad (19)$$

where $\mu_{|\mathcal{K}_t|} = \sum_{k=1}^{K} p_{k,t}$ and $\sigma_{|\mathcal{K}_t|}^2 = \sum_{k=1}^{K} p_{k,t}(1-p_{k,t})$.

Theorem 2 is proved in Appendix E. From the above result, we observe that the convergence rate depends on: (a) the total number of users K, (b) the number of model parameters d, (c) worst amount of perturbation noise across user per iteration, and (d) the sampling probabilities $p_{k,t}$s. When the p_t^* from (16) is used, the convergence rate becomes the following.

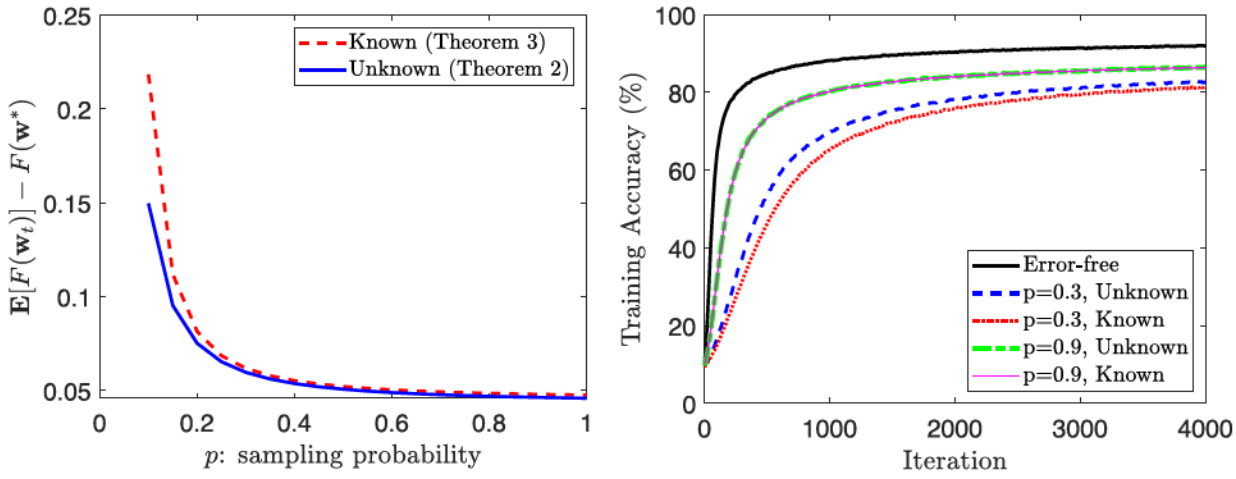
Corollary 2 (Convergence Under Optimal p_t^* From (16)): Under the same assumptions as Theorem 2, the convergence rate for the case when the optimal sampling probability p_t^* from (16) is

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T^2}\sum_{t=1}^{T}\left[\frac{L^2(2\sqrt{K}-\alpha)}{\gamma_t^2 \alpha K} + \frac{d}{\gamma_t^2 \alpha^2 K}\left[\alpha\sqrt{K}\max_k \sigma_{k,t}^2 + N_0\right]\right], \quad (20)$$

where $\alpha = 2\sqrt{\frac{1}{2}\log\frac{2}{\delta'}}$.

It can be seen that the constant in front of both bounds scale as $\mathcal{O}(1/T)$. However, the second parts of the expressions depends on the sampling probabilities. We can see from (20) that the first term in the bracket is constant and that the second term scales as $\mathcal{O}(1/\sqrt{K})$. Since p_t^* is obtained when privacy is

[1] By assuming smooth and strongly convex global loss function, we are able to show convergence to the optimal point. One can also show convergence for non-convex loss functions to a stationary point by following similar steps in [38] and showing that the expectation of the gradient norm, $\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right]$, diminishes as the number of iterations goes to infinity.



(a) Theoretical bounds on optimality gap. **(b)** Empirical.

Fig. 4. Comparisons of convergence bounds and training accuracy with uniform sampling for both cases: (1) unknown $|\mathcal{K}_t|$ or (2) known $|\mathcal{K}_t|$, where $K = 20$, $L = 2$, $T = 4000$, $\lambda = 0.2$, $\mu = 0.9$, $d = 30$, $N_0 = 1$, $\sigma^2_{k,t} = 0.1, \gamma_t = 1$ and $\delta_\ell = \delta' = 10^{-5}$. Each user has transmit $\text{SNR}_k = 10$ dB and (b) is trained on MNIST dataset.

prioritized, (20) is potentially the worst bound of the two. One can potentially select sampling probabilities for (19) to obtain even better scaling than $\mathcal{O}(1/\sqrt{K})$. We next present the convergence results for the case when $\mathcal{K}_t$ is known at the PS.

Theorem 3 (Known $\mathcal{K}_t$ With Non-Uniform Sampling): 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 FedSGD algorithm is given as

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T^2} \sum_{t=1}^{T} \left[\frac{L^2}{\gamma_t^2 \zeta_t} + \frac{d}{\gamma_t^2 \zeta_t^2} \left[\max_k \sigma^2_{k,t} \times \mathbb{E}\left[\frac{1}{|\mathcal{K}_t|}\right] + \mathbb{E}\left[\frac{1}{|\mathcal{K}_t|^2}\right] N_0 \right] \right], \quad (21)$$

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

Theorem 3 depends on $\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|}\right]$ and $\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|^2}\right]$. Note that $\mathcal{K}_t$ is a binomial random variable. It is difficult to obtain closed form expressions for $\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|}\right]$ and $\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|^2}\right]$. However, it is possible to approximate them using Taylor series approximation, specifically, we approximate $\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|}\right]$ using Taylor's series around $\mathbb{E}\left[|\mathcal{K}_t|\right]$ for upto second degree as follows:

$$\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|}\right] \approx \mathbb{E}\left[\frac{1}{\mathbb{E}\left[|\mathcal{K}_t|\right]} - \frac{1}{\mathbb{E}^2\left[|\mathcal{K}_t|\right]}\left(|\mathcal{K}_t| - \mathbb{E}\left[|\mathcal{K}_t|\right]\right) + \frac{1}{\mathbb{E}^3\left[|\mathcal{K}_t|\right]}\left(|\mathcal{K}_t| - \mathbb{E}\left[|\mathcal{K}_t|\right]\right)^2\right] = \frac{1}{\mu_{|\mathcal{K}_t|}} + \frac{\sigma^2_{|\mathcal{K}_t|}}{\mu^3_{|\mathcal{K}_t|}}. \quad (22)$$

Similarly for $\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|^2}\right]$, we approximate it around $\mathbb{E}\left[|\mathcal{K}_t|\right]$ as follows:

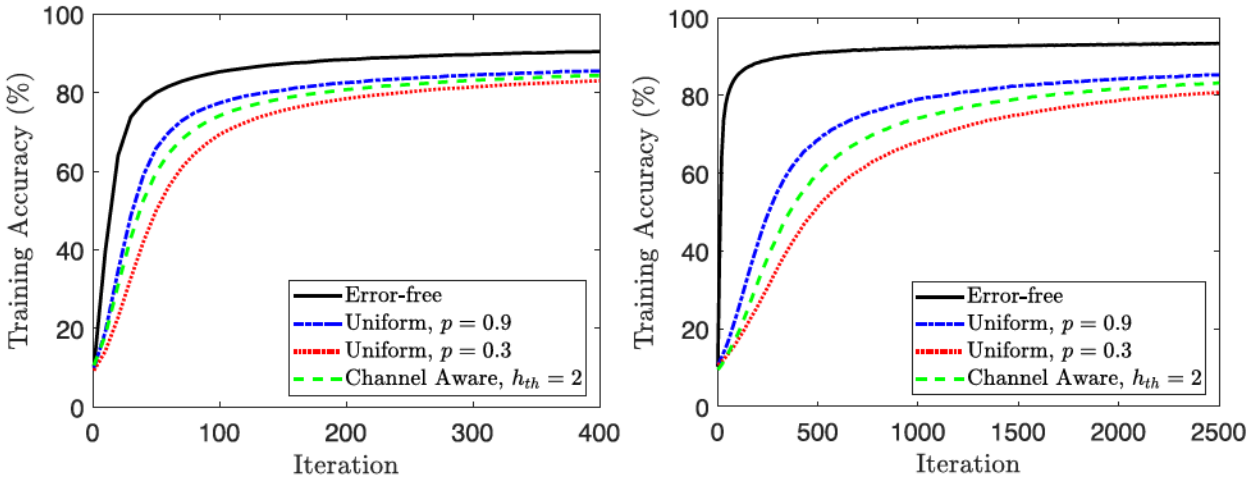
$$\mathbb{E}\left[\frac{1}{|\mathcal{K}_t|^2}\right] \approx \frac{1}{\mu^2_{|\mathcal{K}_t|}} + \frac{3\sigma^2_{|\mathcal{K}_t|}}{\mu^4_{|\mathcal{K}_t|}}. \quad (23)$$

By plugging (22) and (23) back to Theorem 3 for the uniform sampling case, and setting $p_{k,t} = p, \forall k, t$, $\mu_{|\mathcal{K}_t|} = Kp$ and $\sigma^2_{|\mathcal{K}_t|} = Kp(1-p), \forall t$, we obtain,

$$\mathbb{E}\left[F(\mathbf{w}_T)\right] - F(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T^2} \sum_{t=1}^{T} \left[\frac{L^2}{\gamma_t^2 \zeta} + \frac{d}{Kp\gamma_t^2 \zeta^2} \times \left[\sigma^2_{\max,t}(1 + (1-p)^2) + (1 + 3(1-p)^2)\frac{N_0}{Kp} \right] \right],$$

where $\zeta = 1 - (1-p)^K$ and $\sigma^2_{\max,t} = \max_k \sigma^2_{k,t}$.

We note that this bound behaves similarly to the bound in Theorem 2 with $p_{k,t} = p, \forall k, t$ when either T or K is large. Therefore, the proposed scheme performs similarly when $|\mathcal{K}_t|$ is known or unknown. This can be seen in Fig. 4 where the curves are obtained for $K = 200$ users, and $T = 4000$ iterations. We also show this empirically in Fig. 4 using MNIST dataset. It can be seen that for the same sampling probability p, schemes with unknown $|\mathcal{K}_t|$ are always better than schemes with known $|\mathcal{K}_t|$. The difference between two approaches is only at the scaling of the aggregated gradient. This observation indicates that as long as the direction of the aggregated gradient is preserved and the scaling is not drastically different, the performance of the SGD algorithm will not deviate much [39]. This is due to the fact that the magnitude of the gradient at a particular iteration is always corrected in the following iterations as long as the direction is correct.



(a) $L = 1$, $T = 400$. **(b)** $L = 0.1$, $T = 2500$.

Fig. 5. The impact of the sampling probability on the training accuracy for single-layer neural network trained on MNIST dataset with $\sigma_{k,t}^2 = 0.1$.

TABLE II

COMPARISON OF PRIVACY LEAKAGE PER ITERATION FOR SINGLE-LAYER NEURAL NETWORK WITH $\sigma_{k,t}^2 = 0.1$. $\epsilon_{\ell,\max}$ AND $\epsilon_{c,\max}$ DENOTE THE MAXIMUM LOCAL AND CENTRAL LEAKAGES ACROSS ITERATIONS, RESPECTIVELY

	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%

(a) $L = 1, T = 400$.

	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%

(b) $L = 0.1, T = 2500$.

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 [40], 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 $\delta' = 2e^{-2\mu^2|\mathcal{K}_t|/K} + 10^{-5}$ 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.

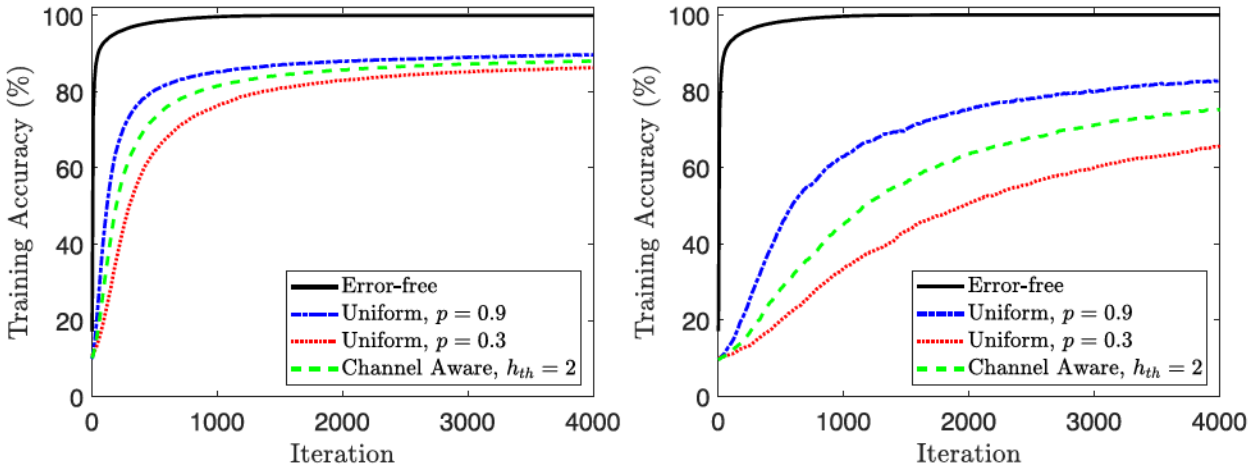
We train two models: (a) a single-layer neural network (NN) (with no hidden layer) and (b) a two-layer NN (with one hidden layer), using MNIST dataset, which consists of $60,000$ training and $10,000$ testing samples. The loss function we used is cross-entropy, and ADAM optimizer for training with a 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. For the experiments, we set the alignment constant $\gamma_t = 1$. Thus, empirically, the scaling factor is computed as follows,

$$\alpha_{k,t} = \min\left[\frac{1}{h_{k,t}}, \frac{\sqrt{P_k}}{\sqrt{\|\mathrm{g}_k(\mathrm{w}_t)\|^2 + d\sigma_{k,t}^2}}\right]. \tag{24}$$

In Fig. 5 and 6, 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(a), 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 (14), i.e.,

$$\epsilon_{c,t} \leq \log\left[1 + \frac{p}{1-\delta'}\left(e^{\frac{c}{\sqrt{K(p-\beta)}}} - 1\right)\right], \tag{25}$$

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 is enhanced. Hence, users are encouraged to participate more when c belongs to this range. In general, c depends on



(a) $L = 1$. **(b)** $L = 0.2$.

Fig. 6. The impact of the sampling probability on the training accuracy for two-layer neural network trained on MNIST dataset with $\sigma_{k,t}^2 = 0.8$.

TABLE III

COMPARISON OF PRIVACY LEAKAGE PER ITERATION FOR TWO-LAYER NEURAL NETWORK WITH $\sigma_{k,t}^2 = 0.8$

	Channel Aware	Uniform	
	$h_{\text{th}} = 2$	$p = 0.3$	$p = 0.9$
$\epsilon_{\ell,\max}$	1.390	2.084	0.8953
$\epsilon_{c,\max}$	1.991	1.653	1.487
Avg. $\lvert\mathcal{K}\rvert$	96	60	180
Testing Acc.	88.72%	87.10%	90.28%

(a) $L = 1$.

	Channel Aware	Uniform	
	$h_{\text{th}} = 2$	$p = 0.3$	$p = 0.9$
$\epsilon_{\ell,\max}$	0.2795	0.4169	0.1791
$\epsilon_{c,\max}$	0.2620	0.1505	0.1633
Avg. $\lvert\mathcal{K}\rvert$	96	60	180
Testing Acc.	75.89%	66.33%	83.68%

(b) $L = 0.2$.

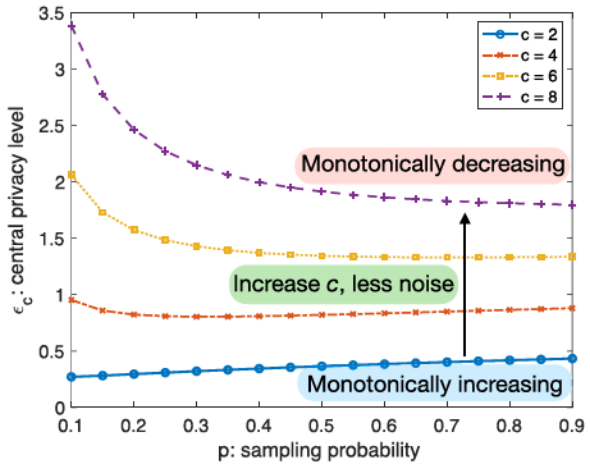


Fig. 7. Central DPs as a function of p for different values of c with $K = 20$.

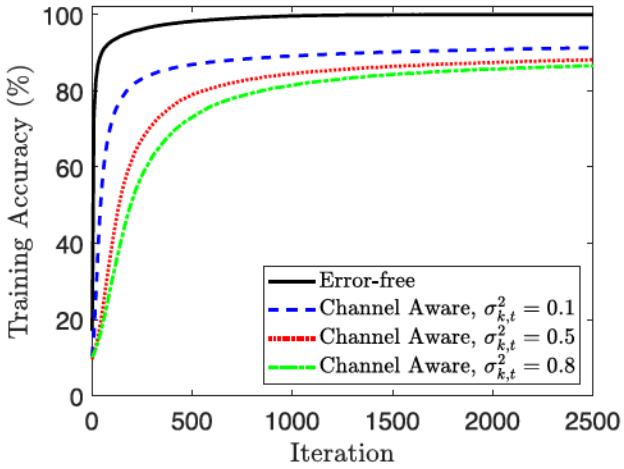


Fig. 8. The impact of the perturbation noise on training accuracy for two-layer NN.

$\sigma_{k,t}, L, \delta_\ell$, and c for Fig. 5a and Table II(a) 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 increases as p increases from Table II(b). 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. This can also be seen in Fig. 7 that the first term dominates when $c = 2$ and the second term dominates when $c = 4, 6, 8$. Similar trends can be found in Table III.

From Table II, we can also 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 low LDP leakage with significant less average number of participants compare to uniform sampling with $p = 0.9$.

V. CONCLUSION & FUTURE DIRECTIONS

In this work, we showed the privacy benefits of user sampling and wireless aggregation for federated learning. More specifically, we showed that for certain settings (when c

is relatively small), the benefit of user sampling outweighs the advantage of wireless aggregation, therefore, creating tension between central DP, local DP and convergence rate. To minimize central DP, user sampling is essential, and we can tradeoff local DP and convergence rate for central DP by sampling less. However, for other settings (when c is relatively large), the privacy amplification from wireless aggregation outweighs the disadvantage of additional leakage from sampling more, making the tension between central DP, local DP and convergence rate disappear. Hence, user sampling is, in fact, discouraged to minimize central DP. The resulting leakage for central DP was shown to scale as $O(1/K^{3/4})$, improving upon prior results on this topic. We also showed that knowing only the statistics of the number of participants at each iteration is at least as good as knowing the exact number of participants and hence eliminating the need for coordination between the PS and users.

There are several interesting future directions which we briefly discuss next. An immediate direction would be to study other variations of FL such as FedAvg, where each user performs multiple local model updates followed by model exchange with the PS. We believe that recent results (without privacy) such as [1], [41], [42], together with the techniques developed in this paper would be useful in such a generalization. Another interesting direction would be to design data and channel dependent sampling mechanisms, where the sampling probabilities at each user can depend on both the local gradients/losses as well as the local channel quality of each user.

APPENDIX A
GAUSSIAN MECHANISM FOR LDP

In this paper, we assume that each user's local perturbation noise is drawn from Gaussian distribution. This well-known technique is known as Gaussian mechanism and can provide rigorous privacy guarantees for LDP.

Definition 3 (Gaussian Mechanism [21]): Suppose a user wants to release a function $f(X)$ of an input X subject to $(\epsilon_\ell, \delta_\ell)$-LDP. The Gaussian release mechanism is defined as $M(X) \triangleq f(X) + \mathcal{N}(0, \sigma^2\mathbf{I})$. If the sensitivity of the function is bounded by Δ_f, i.e., $\|f(x) - f(x')\|_2 \leq \Delta_f$, $\forall x$, then for any $\delta_\ell \in (0, 1]$, Gaussian mechanism satisfies $(\epsilon_\ell, \delta_\ell)$-LDP, where $\epsilon_\ell = \frac{\Delta_f}{\sigma}\sqrt{2\log\left(\frac{1.25}{\delta_\ell}\right)}$.

APPENDIX B
PROOF OF THEOREM 1

In this section, we prove the privacy amplification due to non-uniform sampling of the users. For the per-iteration analysis, we drop the iteration index t for brevity. Let Y denote the output seen at the PS through MAC and Y_{-k} denote the output when user k does not participate. Recall that DP guarantees that any post-processing done on the received signal does not leak more information about the input. Therefore, it is sufficient to show the following,

$$\Pr(Y \in \mathcal{S}) \leq e^{\epsilon_c} \Pr(Y_{-k} \in \mathcal{S}) + \delta_c, \quad \forall k, \tag{26}$$

and obtain ϵ_c. The challenge of this proof is the random participation of users and that the local noises get aggregated over the wireless channel. In this case, let $\mathcal{K}$ denote the random set of users that participate in an iteration, and let $R = |\mathcal{K}|$ denote the random variable representing the number of participants. One can readily check that R is a summation of K Bernoulli random variables and has mean $\mu_R = \sum_{k=1}^{K} p_k$, where p_k is the sampling probability of user k. The number of participants $R = |\mathcal{K}|$ determines the amplification of local DP via wireless aggregation, and in turn, determines the central DP. To take all possible $\mathcal{K}$ into account for the analysis, we condition the lefthand side of (26) with the event that $\mathcal{K}$ deviates from the mean, i.e., $|R - \mu_R| \geq \beta K$ for any $\beta > 0$, and bound it using Hoeffding's inequality and local DP guarantee. To apply local DP guarantee, we need additional conditioning on the event $\mathcal{E}_k$ that denotes the event where user k participates in the training, i.e., $k \in \mathcal{K}$. Note that $p_k = \Pr(\mathcal{E}_k), \forall k$ and the conditional probabilities $\bar{p}_k = \Pr(\mathcal{E}_k || R - \mu_R| < \beta K), \forall k$ can be readily bounded by p_k's using total probability theorem and Hoeffding's inequality, i.e., one can show that $\bar{p}_k \leq p_k/(1-\delta')$. For any $k \in [K]$, we have the following inequalities:

$$\begin{aligned}
&\Pr(Y \in \mathcal{S}) \\
&= \Pr(|R - \mu_R| \geq \beta K) \Pr(Y \in \mathcal{S} || R - \mu_R| \geq \beta K) \\
&\quad + \Pr(|R - \mu_R| < \beta K) \Pr(Y \in \mathcal{S} || R - \mu_R| < \beta K) \\
&\leq \delta' \cdot 1 + \Pr(|R - \mu_R| < \beta K) \Pr(Y \in \mathcal{S} || R - \mu_R| < \beta K),
\end{aligned} \tag{27}$$

where the inequality follows from the fact that any probability is upper bounded by 1 and from the Lemma below:

Lemma 3: (Hoeffding's Inequality for Binomial Random Variable) For a binomial random variable X with K trials and mean μ_X, the probability that X deviates from the mean by more than βK can be bounded as,

$$\Pr(|X - \mu_X| \geq \beta K) \leq 2e^{-2\beta^2 K} \triangleq \delta', \tag{28}$$

for any $\beta > 0$, and any $\delta' \in [0, 1)$.

To further upper bound (27), we use the following Lemma.

Lemma 4: Let $\bar{p}_k = \Pr(\mathcal{E}_k || R - \mu_R| < \beta K)$ and c be some constant that depends on the privacy mechanism, specifically for the Gaussian mechanism we have $c \triangleq \frac{2L}{\sigma_{\min}}\sqrt{2\log\frac{1.25}{\delta_\ell}}$, where L is the Lipschitz constant. The following inequality is true when the local mechanism satisfies $\left(\frac{c}{\sqrt{\mu_R - \beta K}}, \delta_\ell\right)$-LDP:

$$\begin{aligned}
&\Pr(Y \in \mathcal{S} || R - \mu_R| < \beta K) \\
&\leq \bar{p}_k \delta_\ell + \left[\bar{p}_k \left(e^{\frac{c}{\sqrt{\mu_R - \beta K}}} - 1\right) + 1\right] \\
&\quad \times \Pr(Y_{-k} \in \mathcal{S} || R - \mu_R| < \beta K)
\end{aligned}$$

Using Lemma 4, we can bound (27) as follows:

$$\begin{aligned}
&\Pr(Y \in \mathcal{S}) \\
&\leq \delta' + \Pr(|R - \mu_R| < \beta K) \Big[\bar{p}_k \delta_\ell \\
&\quad + \left[\bar{p}_k \left(e^{\frac{c}{\sqrt{\mu_R - \beta K}}} - 1\right) + 1\right] \Pr(Y_{-k} \in \mathcal{S} || R - \mu_R| < \beta K)\Big] \\
&\overset{(a)}{\leq} \delta' + \bar{p}_k \delta_\ell + \Pr(|R - \mu_R| < \beta K) \\
&\quad \times \left[\bar{p}_k \left(e^{\frac{c}{\sqrt{\mu_R - \beta K}}} - 1\right) + 1\right] \frac{\Pr(Y_{-k} \in \mathcal{S})}{\Pr(|R - \mu_R| < \beta K)}
\end{aligned}$$

$$\overset{(b)}{\leq} \delta' + \frac{p_k}{1-\delta'}\delta_\ell + \left[\frac{p_k}{1-\delta'}\left(e^{\frac{c}{\sqrt{\mu_R-\beta K}}}-1\right)+1\right]\Pr(Y_{-k}\in\mathcal{S}) \quad (29)$$

where (a) follows from total probability theorem and the fact that $\Pr(|R-\mu_R|<\beta K)\bar{p}_k\delta_\ell \leq \bar{p}_k\delta_\ell$; and (b) follows from inequality $\bar{p}_k \leq p_k/(1-\delta')$ mentioned at the beginning of the proof. We can obtain a bound for each user k in a similar fashion. By selecting the bound that gives us the largest privacy parameters, we recover the result of Theorem 1. We next prove Lemma 4.

Proof of Lemma 4: With the $\mathcal{E}_k$ defined above, let $\mathcal{E}_k^c$ denote its complementary event. Then, using total probability theorem, we have

$$\begin{aligned}
&\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K)\\
&= \bar{p}_k\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k)\\
&\quad + (1-\bar{p}_k)\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k^c)\\
&\overset{(a)}{=} \bar{p}_k\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k)\\
&\quad + (1-\bar{p}_k)\Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K), \quad (30)
\end{aligned}$$

where we can show that (a) is true as follows,

$$\begin{aligned}
&\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k^c)\\
&= \sum_{\substack{A_{-k}\subseteq[K],\\ ||A_{-k}|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A_{-k}||R-\mu_R|<\beta K,\mathcal{E}_k^c)\\
&\quad\times\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k^c,\mathcal{K}=A_{-k})\Big]\\
&\overset{(a)}{=} \sum_{\substack{A_{-k}\subseteq[K],\\ ||A_{-k}|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A_{-k}||R-\mu_R|<\beta K)\\
&\quad\times\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k^c,\mathcal{K}=A_{-k})\Big]\\
&\overset{(b)}{=} \sum_{\substack{A_{-k}\subseteq[K],\\ ||A_{-k}|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A_{-k}||R-\mu_R|<\beta K)\\
&\quad\times\Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{K}=A_{-k})\Big]\\
&= \Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K) \quad (31)
\end{aligned}$$

where (a) holds since user k is not in the set A_{-k}, therefore, conditioning on the event $\mathcal{E}_k^c$ does not change the probability; and (b) follows due to similar argument. Next, we upper bound $\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k)$ as follows:

$$\begin{aligned}
&\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k)\\
&= \sum_{\substack{A\subseteq[K]:k\in A,\\ ||A|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A||R-\mu_R|<\beta K,\mathcal{E}_k)\\
&\quad\times\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k,\mathcal{K}=A)\Big], \quad (32)
\end{aligned}$$

Note that, in wireless setting, when each user k applies a mechanism that satisfies $(\epsilon_\ell,\delta_\ell)$-LDP, it implies $(c/\sqrt{|A|},\delta_\ell)$-DP [27] (using quasi-convexity property of DP [43]), we have,

$$\begin{aligned}
&\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k,\mathcal{K}=A)\\
&\leq e^{\frac{c}{\sqrt{|A|}}}\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k^c,\mathcal{K}=A_{-k})+\delta_\ell. \quad (33)
\end{aligned}$$

Plugging (33) into (32), we obtain the following:

$$\begin{aligned}
&\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k)\\
&\leq \sum_{\substack{A\subseteq[K]:k\in A,\\ ||A|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A||R-\mu_R|<\beta K,\mathcal{E}_k)\\
&\quad\times\Big[e^{\frac{c}{\sqrt{|A|}}}\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k^c,\mathcal{K}=A_{-k})+\delta_\ell\Big]\Big]\\
&\overset{(a)}{=} \delta_\ell + \sum_{\substack{A\subseteq[K]:k\in A,\\ ||A|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A_{-k}||R-\mu_R|<\beta K)e^{\frac{c}{\sqrt{|A|}}}\\
&\quad\times\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{E}_k^c,\mathcal{K}=A_{-k})\Big]\\
&\overset{(b)}{=} \delta_\ell + \sum_{\substack{A\subseteq[K]:k\in A,\\ ||A|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A_{-k}||R-\mu_R|<\beta K)e^{\frac{c}{\sqrt{|A|}}}\\
&\quad\times\Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{K}=A_{-k})\Big]\\
&\overset{(c)}{\leq} e^{\frac{c}{\sqrt{\mu_R-\beta K}}} \sum_{\substack{A\subseteq[K]:k\in A,\\ ||A|-\mu_R|<\beta K}} \Big[\Pr(\mathcal{K}=A_{-k}||R-\mu_R|<\beta K)\\
&\quad\times\Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K,\mathcal{K}=A_{-k})\Big]+\delta_\ell\\
&= e^{\frac{c}{\sqrt{\mu_R-\beta K}}}\Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K)+\delta_\ell \quad (34)
\end{aligned}$$

where (a) and (b) follows the similar argument as the one used in (31), distributive property of multiplication and the fact that the probability multiplied with δ_ℓ sums up to one. From the condition on the cardinality of the set R, we know that $R=|A|$ and $\mu_R-\beta K<|A|<\mu_R+\beta K$. Therefore, (c) follows from using the lower bound on $|A|$. Then, by combining (30), (31) and (34), we have

$$\begin{aligned}
&\Pr(Y\in\mathcal{S}||R-\mu_R|<\beta K)\\
&\leq \bar{p}_k e^{\frac{c}{\sqrt{\mu_R-\beta K}}}\Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K)+\bar{p}_k\delta_\ell\\
&\quad + (1-\bar{p}_k)\Pr(Y_{-k}\in\mathcal{S}||R-\mu_R|<\beta K). \quad (35)
\end{aligned}$$

Rearranging the above inequality, we recover the result of Lemma 4. ■

Appendix C
Proof of Lemma 1

In this section, we find the optimal sampling probability p_t^* that minimizes the central privacy level $\epsilon_{c,t}$ for the wireless FedSGD scheme. For the per-iteration analysis, we drop the iteration index for brevity. We minimize ϵ_c as follows:

$$\begin{aligned}
\epsilon_c &= \log\left[\frac{p}{1-\delta'}\left(e^{\frac{c}{\sqrt{Kp-\beta K}}}-1\right)+1\right]\\
&\leq \frac{p}{1-\delta'}\left(e^{\frac{c}{\sqrt{Kp-\beta K}}}-1\right). \quad (36)
\end{aligned}$$

We assume that p takes the form of $\tilde{k}/K$, i.e., $p=\frac{\tilde{k}}{K}$, where we assume that $p>\beta$, so that the exponent term is non negative. Then,

$$\begin{aligned}
\epsilon_c &\leq \frac{\tilde{k}}{K(1-\delta')}\left(e^{\frac{c}{\sqrt{\tilde{k}-\beta K}}}-1\right)\\
&\overset{(a)}{\leq} \frac{\tilde{k}}{K(1-\delta')}\times\frac{c}{\sqrt{\tilde{k}-\beta K}} \triangleq \tilde{\epsilon}_c, \quad (37)
\end{aligned}$$

where in step (a) for large K, we assume that the exponent is less than 1, and we use the fact that $e^x - 1 \leq x, \forall x \leq 1$. Taking the derivative of the right-hand side w.r.t. $\tilde{k}$ and setting it to zero yields the following:

$$\frac{\partial \tilde{\epsilon}_c}{\partial \tilde{k}} = \frac{c}{K(1-\delta')}\left[-\frac{\tilde{k}}{2}(\tilde{k}-\beta K)^{\frac{-3}{2}} + (\tilde{k}-\beta K)^{\frac{-1}{2}}\right] = 0$$
$$\Rightarrow \tilde{k} = 2\beta K.$$

We then check the second derivative of the right-hand side and obtain,

$$\frac{\partial^2 \tilde{\epsilon}_c}{\partial \tilde{k}^2} \propto \frac{-1}{2} \times \left[-\frac{3}{2}\tilde{k}(\tilde{k}-\beta K)^{-5/2} + 2(\tilde{k}-\beta K)^{-3/2}\right].$$

It can be readily shown that $\frac{\partial^2 \tilde{\epsilon}_c}{\partial \tilde{k}^2} \leq 0$ when $\tilde{k} \geq 2\beta K$. To this end, the optimal sampling probability that minimize ϵ_c is $p^* = 2\beta$. Using Lemma 3, we know that $\beta = \frac{1}{\sqrt{K}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\delta'}\right)}$. By plugging p^* and β into (36), we get:

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

This completes the proof of Lemma 1.

Appendix D
Proof of Lemma 2

The final received signal at the PS from (10) can be expressed as: $\mathrm{y}_t = \sum_{k\in\mathcal{K}_t} h_{k,t}\alpha_{k,t}\mathrm{g}_k(\mathrm{w}_t) + \mathrm{z}_t$ and the variance of the effective Gaussian noise z_t is

$$\sigma^2 = \sigma_{z_t}^2 = \sum_{k\in\mathcal{K}_t} h_{k,t}^2\alpha_{k,t}^2\sigma_{k,t}^2 + N_0 \overset{(a)}{=} \gamma_t^2 \sum_{k\in\mathcal{K}_t}\sigma_{k,t}^2 + N_0,$$

where step (a) follows from the alignment condition in (6). In order to invoke the result of the Gaussian mechanism (Appendix A), we next obtain a bound on the sensitivity for user k. To bound the local sensitivity of user k, we fix the gradients of the remaining $\mathcal{K}_t\backslash k$ users. The local sensitivity of user k can then be bounded as

$$\begin{aligned}\Delta_{k,t} &= \max_{\mathcal{D}_k,\mathcal{D}_k'} ||\mathrm{y}_t - \mathrm{y}_t'||_2 \\ &= \max_{\mathcal{D}_k,\mathcal{D}_k'} ||h_{k,t}\alpha_{k,t}(\mathrm{g}_k(\mathrm{w}_t) - \mathrm{g'}_k(\mathrm{w}_t))||_2 \\ &\leq h_{k,t}\alpha_{k,t}\max_{\mathcal{D}_k,\mathcal{D}_k'} ||\mathrm{g}_k(\mathrm{w}_t)||_2 + ||\mathrm{g'}_k(\mathrm{w}_t)||_2 \\ &\overset{(a)}{\leq} 2h_{k,t}\alpha_{k,t}L \overset{(b)}{=} 2\gamma_t L, \end{aligned} \tag{38}$$

where (a) follows from the fact that $\|\mathrm{g}_k(\mathrm{w}_t)\|_2 \leq L, \forall k$; and (b) follows from the channel inversion transmission scheme. We next show the guarantee on the local DP of user k when user k is a participant. Following similar steps used for proving (27), it can be shown that,

$$\begin{aligned}\Pr(Y_x^{(k)} \in \mathcal{S}|\mathcal{E}_k) &\leq \delta' + \delta_\ell + e^{\epsilon_{\ell,t}^{(k)}}\Pr(Y_{x'}^{(k)} \in \mathcal{S}|\mathcal{E}_k) \\ &\overset{(a)}{\leq} \delta' + \delta_\ell + e^{\frac{c}{\sqrt{1+\mu_R-\beta K}}}\Pr(Y_{x'}^{(k)} \in \mathcal{S}|\mathcal{E}_k), \end{aligned} \tag{39}$$

where $\mu_R = \sum_{i=1,i\neq k}^{K} p_i$ and $\epsilon_{\ell,t}^{(k)}$ is the effective local privacy level of user k due to sampling and wireless aggregation. Step (a) follows from applying the Gaussian mechanism, i.e.,

$$\begin{aligned}\epsilon_{\ell,t}^{(k)} &\leq \frac{2\gamma_t L}{\sqrt{\sum_{k=1}^{1+\mu_R-\beta K}\gamma_t^2\sigma_{k,t}^2 + N_0}}\sqrt{2\log\frac{1.25}{\delta_\ell}} \\ &\leq \frac{2\gamma_t L}{\sqrt{(1+\mu_R-\beta K)\times\gamma_t^2\min_{k,t}\sigma_{k,t}^2}}\sqrt{2\log\frac{1.25}{\delta_\ell}} \\ &= \frac{1}{\sqrt{1+\mu_R-\beta K}} \times \frac{2L}{\sigma_{\min}}\sqrt{2\log\frac{1.25}{\delta_\ell}}. \end{aligned}$$

Note that (39) is conditioning on the event when user k participates. We next use the total probability theorem and obtain the following set of steps:

$$\begin{aligned}&\Pr(Y_x^{(k)} \in \mathcal{S}) \\ &= p_k\Pr(Y_x^{(k)} \in \mathcal{S}|\mathcal{E}_k) + (1-p_k)\Pr(Y_x^{(k)} \in \mathcal{S}|\mathcal{E}_k^c) \\ &\overset{(a)}{\leq} p_k e^{\epsilon_\ell}\Pr(Y_{x'}^{(k)} \in \mathcal{S}|\mathcal{E}_k) + p_k(\delta_\ell + \delta') \\ &\quad + (1-p_k)e^{\epsilon_\ell}\Pr(Y_{x'}^{(k)} \in \mathcal{S}|\mathcal{E}_k^c) \\ &= e^{\epsilon_\ell}\Pr(Y_{x'}^{(k)} \in \mathcal{S}) + p_k(\delta_\ell + \delta'), \end{aligned}$$

where step (a) follows from (39) and the fact that when user k is not participating, we have

$$\begin{aligned}\Pr(Y_x^{(k)} \in \mathcal{S}|\mathcal{E}_k^c) &= e^0\Pr(Y_{x'}^{(k)} \in \mathcal{S}|\mathcal{E}_k^c) \\ &\leq e^{\epsilon_\ell}\Pr(Y_{x'}^{(k)} \in \mathcal{S}|\mathcal{E}_k^c), \quad \forall x, x'. \end{aligned}$$

We arrive at the proof of Lemma 2.

Appendix E
Proofs of Theorem 2 and Theorem 3

When the data is i.i.d., we can invoke a slightly modified version of the result of [44] on convergence of SGD for μ-smooth and λ-strongly convex loss, which states

$$\mathbb{E}\left[F(\mathrm{w}_T)\right] - F(\mathrm{w}^*) \leq \frac{2\mu}{\lambda^2 T}\left(\sum_{t=1}^{T} G_t^2/T\right), \tag{40}$$

where G_t^2 is the upper bound on the second moment of the gradient estimate, i.e., $\mathbb{E}\left[\|\hat{\mathrm{g}}_t\|_2^2\right] \leq G_t^2$.

A. $|\mathcal{K}_t|$ Is Unknown at the PS

To prove the convergence rate of the proposed algorithm, we recall that the gradient estimate at the PS in (8) needs to satisfy: (a) Unbiasedness, i.e., $\mathbb{E}\left[\hat{\mathrm{g}}_t\right] = \mathrm{g}_t$, since the total additive noise is zero mean; and (b) Bounded second moment, $\mathbb{E}\left[\|\hat{\mathrm{g}}_t\|_2^2\right] \leq G_t^2$, which we prove as follows. Recall that the estimated gradient at the PS is

$$\begin{aligned}\hat{\mathrm{g}}_t &= \frac{1}{\mu_{|\mathcal{K}_t|}}\sum_{k\in\mathcal{K}_t}\mathrm{g}_k(\mathrm{w}_t) + \frac{1}{\gamma_t\mu_{|\mathcal{K}_t|}}\mathrm{z}_t \\ &= \frac{1}{\mu_{|\mathcal{K}_t|}}\sum_{k\in\mathcal{K}_t}\frac{1}{b}\sum_{i\in\mathcal{B}_k}\nabla f_k(\mathrm{w}_t;(\mathrm{u}_i^{(k)}, v_i^{(k)})) + \frac{1}{\gamma_t\mu_{|\mathcal{K}_t|}}\mathrm{z}_t. \end{aligned}$$

By taking the expectation over the randomness of SGD, user sampling and noise, we have

$$\mathbb{E}\left[\hat{\mathbf{g}}_t\right] = \frac{1}{\mu_{|\mathcal{K}_t|}b}\mathbb{E}\left[\sum_{k\in\mathcal{K}_t}\sum_{i\in\mathcal{B}_k}\nabla f_k(\mathbf{w}_t;(\mathbf{u}_i^{(k)},v_i^{(k)}))\right] = \frac{1}{\mu_{|\mathcal{K}_t|}b}\mathbb{E}\left[|\mathcal{K}_t|\right]b\mathbf{g}_t = \mathbf{g}_t.$$

Therefore, the estimated gradient is unbiased. We next obtain the bound on the second moment of the estimated gradient. We have

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\mathbf{g}}_t\|_2^2\right] \\
&= \mathbb{E}\left[\|\frac{1}{\gamma_t\mu_{|\mathcal{K}_t|}}\sum_{k\in\mathcal{K}_t}\mathbf{g}_k(\mathbf{w}_t)+\frac{\mathbf{z}_t}{\gamma_t\mu_{|\mathcal{K}_t|}}\|_2^2\right] \\
&\overset{(a)}{=} \frac{1}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}\left(\mathbb{E}\left[\|\sum_{k\in\mathcal{K}_t}\mathbf{g}_k(\mathbf{w}_t)\|_2^2\right]+\mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right]\right) \\
&= \frac{1}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}\Bigg(\mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right] \\
&\quad +\mathbb{E}\left[\sum_{k\in\mathcal{K}_t}\|\mathbf{g}_k(\mathbf{w}_t)\|_2^2+\sum_{k\in\mathcal{K}_t}\sum_{k'\in\mathcal{K}_t}\mathbf{g}_k(\mathbf{w}_t)^T\mathbf{g}_{k'}(\mathbf{w}_t)\right]\Bigg) \\
&\overset{(b)}{\leq} \frac{1}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}\Bigg(\mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right]+\mathbb{E}\Bigg[\sum_{k\in\mathcal{K}_t}\|\mathbf{g}_k(\mathbf{w}_t)\|_2^2 \\
&\quad +\sum_{k\in\mathcal{K}_t}\sum_{k'\in\mathcal{K}_t}\|\mathbf{g}_k(\mathbf{w}_t)\|_2\|\mathbf{g}_{k'}(\mathbf{w}_t)\|_2\Bigg]\Bigg) \\
&\overset{(c)}{\leq} \frac{1}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}\mathbb{E}\left[|\mathcal{K}_t|L^2+|\mathcal{K}_t|(|\mathcal{K}_t|-1)\,L^2\right]+\frac{\mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right]}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2} \\
&= \frac{L^2\mathbb{E}\left[|\mathcal{K}_t|^2\right]}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}+\frac{\mathbb{E}\left[\|\mathbf{z}_t\|_2^2\right]}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2} \\
&\leq \frac{L^2(\mu_{|\mathcal{K}_t|}^2+\sigma_{|\mathcal{K}_t|}^2)}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}+\frac{d}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}\left[\max_k\sigma_{k,t}^2\mathbb{E}\left[|\mathcal{K}_t|\right]+N_0\right] \\
&= \frac{L^2(\mu_{|\mathcal{K}_t|}^2+\sigma_{|\mathcal{K}_t|}^2)}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}+\frac{d}{\gamma_t^2\mu_{|\mathcal{K}_t|}^2}\left[\max_k\sigma_{k,t}^2\mu_{|\mathcal{K}_t|}+N_0\right] \\
&\triangleq G_t^2, \qquad (41)
\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}\left[\mathbf{g}_t^T\mathbf{z}_t\right]=0$, (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 $\forall k$. Plugging G_t^2 from (41) in (40), we arrive at the proof of Theorem 2.

B. $|\mathcal{K}_t|$ Is Known at the PS

We then move to the case when $|\mathcal{K}_t|$ is known at the PS. Recall that the estimated gradient at the PS for the known $|\mathcal{K}_t|$ case is

$$\hat{\mathbf{g}}_t = \frac{1}{\zeta_t|\mathcal{K}_t|}\sum_{k\in\mathcal{K}_t}\mathbf{g}_k(\mathbf{w}_t)+\frac{1}{\gamma_t\zeta_t|\mathcal{K}_t|}\mathbf{z}_t = \frac{1}{\zeta_t|\mathcal{K}_t|}\sum_{k\in\mathcal{K}_t}\frac{1}{b}\sum_{i\in\mathcal{B}_k}\nabla f_k(\mathbf{w}_t;(\mathbf{u}_i^{(k)},v_i^{(k)}))+\frac{1}{\gamma_t\zeta_t|\mathcal{K}_t|}\mathbf{z}_t,$$

where ζ_t is used for maintaining unbiasedness of the estimated gradient and will be specified later. By taking the expectation over the randomness of SGD, user sampling and additive noise, we have

$$\begin{aligned}
&\mathbb{E}\left[\hat{\mathbf{g}}_t\right] \\
&\overset{(a)}{=} \Pr(|\mathcal{K}_t|=0)\mathbb{E}\left[\hat{\mathbf{g}}_t\,|\,|\mathcal{K}_t|=0\right] \\
&\quad +\sum_{k'=1}^{K}\Pr(|\mathcal{K}_t|=k')\mathbb{E}\left[\hat{\mathbf{g}}_t\,|\,|\mathcal{K}_t|=k'\right] \\
&\overset{(b)}{=} \mathbf{0}+\sum_{k'=1}^{K}\Pr(|\mathcal{K}_t|=k') \\
&\quad \times\mathbb{E}\left[\frac{1}{\zeta_t|\mathcal{K}_t|}\sum_{k\in\mathcal{K}_t}\frac{1}{b}\sum_{i\in\mathcal{B}_k}\nabla f_k(\mathbf{w}_t;(\mathbf{u}_i^{(k)},v_i^{(k)}))\Big||\mathcal{K}_t|=k'\right] \\
&\overset{(c)}{=} \frac{\mathbf{g}_t}{\zeta_t}\sum_{k'=1}^{K}\Pr(|\mathcal{K}_t|=k')=\frac{\mathbf{g}_t}{\zeta_t}\Big(1-\prod_{k'=1}^{K}(1-p_{k',t})\Big), \qquad (42)
\end{aligned}$$

where (a) follows from total probability theorem; (b) follows from the fact that when $|\mathcal{K}_t|=0$, $\hat{\mathbf{g}}_t=\mathbf{0}$; and (c) follows from the i.i.d. assumption so that the conditional expectation is $\mathbf{g}_t$. In order get unbiased estimate for $\mathbf{g}_t$, ζ_t is chosen as $\zeta_t = 1-\prod_{k=1}^{K}(1-p_{k,t})$. To bound the second moment, the proof follows similar steps as the unknown $|\mathcal{K}_t|$ case, and is omitted due to space limitation.

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. Artif. Intell. Statist. (AISTATS)*, 2017, pp. 1273–1282.

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 2017, pp. 3–18.

[3] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Membership inference attacks against generative models," *Proc. Privacy Enhancing Technol.*, 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 *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 691–706.

[5] A. Triastcyn and B. Faltings, "Federated learning with Bayesian differential privacy," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 2587–2596.

[6] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "CpSGD: Communication-efficient and differentially-private distributed SGD," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7564–7575.

[7] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2018, *arXiv:1812.06127*. [Online]. Available: http://arxiv.org/abs/1812.06127

[8] J. Chen and R. Luss, "Stochastic gradient descent with biased but consistent gradient estimators," 2018, *arXiv:1807.11880*. [Online]. Available: http://arxiv.org/abs/1807.11880

[9] M. M. Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, Mar. 2020.

[10] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

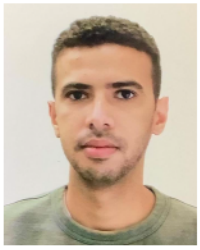
[11] W.-T. Chang and R. Tandon, "MAC aware quantization for distributed gradient descent," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[12] 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, Jan. 2020.

[13] 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, Mar. 2020.

[14] M. M. Amiri and D. Gunduz, "Over-the-air machine learning at the wireless edge," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPWAC)*, Cannes, France, Jul. 2019, pp. 1–5.

[15] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient radio resource allocation for federated edge learning," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.
[16] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 2897–2911, 2020.
[17] S. Wang *et al.*, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 1205–1221, Jun. 2019.
[18] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning ACROSS heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8866–8870.
[19] L. U. Khan *et al.*, "Federated learning for edge networks: Resource optimization and incentive mechanism," *IEEE Commun. Mag.*, vol. 58, no. 10, pp. 88–93, Oct. 2020.
[20] G. Zhu, Y. Du, D. Gunduz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, Mar. 2021.
[21] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
[22] M. Joseph, A. Roth, J. Ullman, and B. Waggoner, "Local differential privacy for evolving data," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 2375–2384.
[23] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," 2017, *arXiv:1712.07557*. [Online]. Available: http://arxiv.org/abs/1712.07557
[24] O. Choudhury *et al.*, "Differential privacy-enabled federated learning for sensitive health data," 2019, *arXiv:1910.02578*. [Online]. Available: http://arxiv.org/abs/1910.02578
[25] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 6277–6287.
[26] B. Balle, P. Kairouz, B. McMahan, O. D. Thakkar, and A. Thakurta, "Privacy amplification via random check-ins," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1–26.
[27] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2604–2609.
[28] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2021.
[29] A. Sonee and S. Rini, "Efficient federated learning over multiple access channel with differential privacy constraints," 2020, *arXiv:2005.07776*. [Online]. Available: http://arxiv.org/abs/2005.07776
[30] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloq. Automata, Lang. Program. (ICALP) II*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds., 2006, pp. 1–12, doi: 10.1007/11787006_1.
[31] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 58–77.
[32] B. Hasircioglu and D. Gunduz, "Private wireless federated learning with anonymous over-the-air computation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5195–5199.
[33] U. Erlingsson *et al.*, "Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation," 2020, *arXiv:2001.03618*. [Online]. Available: http://arxiv.org/abs/2001.03618
[34] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1376–1385.
[35] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Proc. IEEE 51st Annu. Symp. Found. Comput. Sci. (FOCS)*, Washington, DC, USA, Oct. 2010, pp. 51–60.
[36] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 263–275.
[37] T. Luo, M. Pan, P. Tholoniat, A. Cidon, R. Geambasu, and M. Lécuyer, "Privacy budget scheduling," in *Proc. 15th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2021, pp. 55–74.
[38] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.
[39] A. Ajalloeian and S. U. Stich, "On the convergence of SGD with biased gradients," 2020, *arXiv:2008.00051*. [Online]. Available: http://arxiv.org/abs/2008.00051
[40] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[41] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," 2019, *arXiv:1907.02189*. [Online]. Available: http://arxiv.org/abs/1907.02189
[42] W. Chen, S. Horvath, and P. Richtarik, "Optimal client sampling for federated learning," 2020, *arXiv:2010.13723*. [Online]. Available: http://arxiv.org/abs/2010.13723
[43] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proc. 13th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2019, pp. 2468–2479.
[44] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2012, pp. 1571–1578.



Mohamed Seif Eldin Mohamed (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2014, and the M.Sc. degree in wireless technologies from Nile University, Giza, Egypt, in 2016. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, The University of Arizona. He joined as a Graduate Research Assistant with The University of Arizona in 2017. His research interests include information theory, machine learning, and wireless communications.

Wei-Ting Chang (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from West Virginia University in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Arizona. His current research interests are in information and coding theory with applications to wireless communications, secure coded distributed computing, communication efficient distributed machine learning, and security and privacy in machine learning.

Ravi Tandon (Senior Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology Kanpur (IIT Kanpur) in 2004 and the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park (UMCP) in 2010.

From 2010 to 2012, he was a Post-Doctoral Research Associate at Princeton University. He is currently an Associate Professor with the Department of ECE, The University of Arizona. Prior to joining The University of Arizona in Fall 2015, he was a Research Assistant Professor at Virginia Tech with positions at the Bradley Department of ECE, Hume Center for National Security and Technology, and the Discovery Analytics Center, Department of Computer Science. His current research interests include information theory and its applications to wireless networks, signal processing, communications, security and privacy, machine learning, and data mining.

Dr. Tandon was a recipient of the Best Paper Award at IEEE GLOBECOM 2011, the NSF CAREER Award in 2017, and the 2018 Keysight Early Career Professor Award. He serves as an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE TRANSACTIONS ON COMMUNICATIONS.