Federated Learning with Differential Privacy: Algorithms and Performance Analysis

Kang Wei, *Student Member, IEEE*, Jun Li, *Senior Member, IEEE*, Ming Ding, *Senior Member, IEEE*, Chuan Ma, Howard H. Yang, *Member, IEEE*, Farhad Farokhi, *Senior Member, IEEE*, Shi Jin, *Senior Member, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, H. Vincent Poor, *Fellow, IEEE*

***Abstract*—Federated learning (FL), as a type of distributed machine learning, is capable of significantly preserving clients' private data from being exposed to adversaries. Nevertheless, private information can still be divulged by analyzing uploaded parameters from clients, e.g., weights trained in deep neural networks. In this paper, to effectively prevent information leakage, we propose a novel framework based on the concept of differential privacy (DP), in which artificial noises are added to parameters at the clients' side before aggregating, namely, noising before model aggregation FL (NbAFL). First, we prove that the NbAFL can satisfy DP under distinct protection levels by properly adapting different variances of artificial noises. Then we develop a theoretical convergence bound of the loss function of the trained FL model in the NbAFL. Specifically, the theoretical bound reveals the following three key properties: 1) There is a tradeoff between a convergence performance and privacy protection levels, i.e., better convergence performance leads to a lower protection level; 2) Given a fixed privacy protection level, increasing the number N of overall clients participating in FL can improve the convergence performance; and 3) There is an optimal number aggregation times (communication rounds) in terms of convergence performance for a given protection level. Furthermore, we propose a K-client random scheduling strategy, where K ($1 \leq K < N$) clients are randomly selected from the N overall clients to participate in each aggregation. We also develop a corresponding convergence bound for the loss function in this case and the K-client random scheduling strategy also retains the above three properties. Moreover, we find that there is an optimal K that achieves the best convergence performance at a fixed privacy level. Evaluations demonstrate that our theoretical results are consistent with simulations, thereby facilitating the design of various privacy-preserving FL algorithms with different tradeoff requirements on convergence performance and privacy levels.**

***Index Terms*—Federated learning, differential privacy, convergence performance, information leakage, client selection**

I. Introduction

With AlphaGo's success, it is expected that big data-driven artificial intelligence (AI) will soon be applied in many aspects of our daily life, including medical care, agriculture, transportation systems, etc. At the same time, the rapid growth of Internet-of-Things (IoT) applications calls for data mining and learning securely and reliably in distributed systems [1]–[3]. When integrating AI in a variety of IoT applications, distributed machine learning (ML) is preferred for many data processing tasks by defining parametrized functions from inputs to outputs as compositions of basic building blocks [4], [5]. Federated learning (FL), as a recent advance in distributed ML, was proposed, in which data are acquired and processed locally at the client side, and then the updated ML parameters are transmitted to a central server for aggregation [6]–[8]. The goal of FL is to fit a model generated by an empirical risk minimization (ERM) objective. However, FL also poses several key challenges, such as private information leakage, expensive communication costs between servers and clients, and device variability [9]–[15].

Generally, distributed stochastic gradient descent (SGD) is adopted in FL for training ML models. In [16], [17], bounds for FL convergence performance were developed based on distributed SGD, with a one-step local update before global aggregation. The work in [18] considered partially global aggregation, where after each local update step, parameter aggregation is performed over a non-empty subset of the clients set. In order to analyze the convergence, federated proximal (FedProx) was proposed [19] by adding regularization on each local loss function. The work in [20] obtained a convergence bound for SGD based FL that incorporates non-independent-and-identically-distributed (non-*i.i.d.*) data distributions among clients.

At the same time, with the ever increasing awareness of data security of personal information, privacy preservation has become a worldwide and significant issue, especially for big data applications and distributed learning systems. One prominent advantage of FL is that it enables local training

This work is supported in part by the National Key Research and Development Program under Grant 2018YFB1004800, the National Natural Science Foundation of China (Grant No. 61872184 and Grant No. 61727802), the SUTD Growth Plan Grant for AI, the U.S.National Science Foundation under Grant CCF-1908308, and the Princeton Center for Statistics and Machine Learning under a Data X Grant. Corresponding author: Jun Li and Chuan Ma (e-mail:{jun.li, chuan.ma}@njust.edu.cn).

Jun Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, 210094, CHINA. He is also with the School of Computer Science and Robotics, National Research Tomsk Polytechnic University, Tomsk, 634050, RUSSIA. Email: jun.li@njust.edu.cn.

Chuan Ma and Kang Wei are with School of Electrical and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail:{kang.wei, chuan.ma}@njust.edu.cn).

Ming Ding is with Data61, CSIRO, Sydney, NSW 2015, Australia (e-mail:ming.ding@data61.csiro.au).

Farhad Farokhi is with the Department of Electrical and Electronic Engineering, the University of Melbourne, Melbourne, VIC 3010, Australia. When working on this paper, he was also affiliated with CSIROs Data61, Australia (e-mail:ffarokhi@unimelb.edu.au).

Howard H. Yang and Tony Q. S. Quek are with the Information System Technology and Design Pillar, Singapore University of Technology and Design, Singapore (e-mail:{howard yang, tonyquek}@sutd.edu.sg).

Shi Jin is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jinshi@seu.edu.cn).

H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

without personal data exchange between the server and clients, thereby protecting clients' data from being eavesdropped upon by hidden adversaries. Nevertheless, private information can still be divulged to some extent analyzing the differences of parameters trained and uploaded by the clients, e.g., weights trained in neural networks [21]–[23].

A natural approach to preventing information leakage is to add artificial noise, known as differential privacy (DP) techniques [24], [25]. Existing works on DP based learning algorithms include local DP (LDP) [26]–[28], DP based distributed SGD [29], [30] and DP meta learning [31]. In LDP, each client perturbs its information locally and only sends a randomized version to a server, thereby protecting both the clients and server against private information leakage. The work in [27] proposed solutions to building up an LDP-compliant SGD, which powers a variety of important ML tasks. The work in [28] considered the distributed estimation at the server over uploaded data from clients while providing protections on these data with LDP. The work in [32] introduced an algorithm for user-level differentially private training of large neural networks, in particular a complex sequence model for next-word prediction. The work in [33] developed a chain abstraction model on tensors to efficiently override operations (or encode new ones) such as sending/sharing a tensor between workers, and then provided the elements to implement recently proposed DP and multiparty computation protocols using this framework. The work in [29] improved the computational efficiency of DP based SGD by tracking detailed information about the privacy loss, and obtained accurate estimates of the overall privacy loss. The work in [30] proposed novel DP based SGD algorithms and analyzed their performance bounds which were shown to be related to privacy levels and the sizes of datasets. Also, the work in [31] focused on the class of gradient-based parameter-transfer methods and developed a DP based meta learning algorithm that not only satisfies the privacy requirement but also retains provable learning performance in convex settings.

More specifically, DP based FL approaches are usually devoted to capturing the tradeoff between privacy and convergence performance in the training process. The work in [34] proposed an FL algorithm with the consideration on preserving clients' privacy. This algorithm can achieve good training performance at a given privacy level, especially when there is a sufficiently large number of participating clients. The work in [35] presented an alternative approach that utilizes both DP and secure multiparty computation (SMC) to prevent differential attacks. However, the above two works on DP-based FL design have not taken into account privacy protection during the parameter uploading stage, i.e., the clients' private information can be potentially intercepted by hidden adversaries when uploading the training results to the server. Moreover, these two works only showed empirical results using simulations, but lacked theoretical analysis on the FL system, such as the tradeoffs between privacy, convergence performance, and convergence rate. To the authors' knowledge, a theoretical analysis on convergence behavior of FL with privacy-preserving noise perturbations has not yet been considered in existing studies, which will be the major focus of this work. Compared with conventional works, such as [34], [35], which focus mainly on simulation results, our theoretical performance analysis is more efficient to find the optimal parameters, e.g., the number of chosen clients K and the number of maximum aggregation times T, to achieve the minimum loss function.

In this paper, to effectively prevent information leakage, we propose a novel framework based on the concept of DP, in which each client perturbs its trained parameters locally by purposely adding noise before uploading them to the server for aggregation, namely, noising before model aggregation FL (NbAFL). To the best of the authors' knowledge, this is the first piece of work of its kind that theoretically analyzes the convergence properties of differentially private FL algorithms.

The main contributions of this paper are summarized as follows:

- We prove that the proposed NbAFL scheme satisfies the requirement of DP in terms of global data under a certain noise perturbation level with Gaussian noises by properly adapting their variances.
- We develop a convergence bound on the loss function of the trained FL model in the NbAFL with artificial Gaussian noises. Our developed bound reveals the following three key properties: 1) There is a tradeoff between the convergence performance and privacy protection levels, i.e., a better convergence performance leads to a lower protection level; 2) Increasing the number N of overall clients participating in FL can improve the convergence performance, given a fixed privacy protection level; and 3) There is an optimal number of maximum aggregation times in terms of convergence performance for a given protection level.
- We propose a K-client random scheduling strategy, where K ($1 \leq K < N$) clients are randomly selected from the N overall clients to participate in each aggregation. We also develop a corresponding convergence bound on the loss function in this case. From our analysis, the K-client random scheduling strategy retains the above three properties. Also, we find that there exists an optimal value of K that achieves the best convergence performance at a fixed privacy level.
- We conduct extensive simulations based on real-world datasets to validate the properties of our theoretical bound in NbAFL. Evaluations demonstrate that our theoretical results are consistent with simulations. Therefore, our analytical results are helpful for the design on privacy-preserving FL architectures with different tradeoff requirements on convergence performance and privacy levels.

The remainder of this paper is organized as follows. In Section II, we introduce background on FL, DP and a conventional DP-based FL algorithm. In Section III, we detail the proposed NbAFL and analyze the privacy performance based on DP. In Section IV, we analyze the convergence bound of NbAFL and reveal the relationship between privacy levels, convergence performance, the number of clients, and the number of global aggregations. In Section V, we propose the K-client random

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.2988575, IEEE Transactions on Information Forensics and Security

scheduling scheme and develop the convergence bound. We show the analytical results and simulations in Section VI. We conclude the paper in Section VII. A summary of basic concepts and notations is provided in Tab. I.

Table I: Summary of Main Notation

$\mathcal{M}$	A randomized mechanism for DP
$\mathcal{D}, \mathcal{D}'$	Adjacent databases
ϵ, δ	The parameters related to DP
$\mathcal{C}_i$	The i-th client
$\mathcal{D}_i$	The database held by the owner $\mathcal{C}_i$
$\mathcal{D}$	The database held by all the clients
$\lvert\cdot\rvert$	The cardinality of a set
N	Total number of all clients
K	The number of chosen clients ($1 < K < N$)
t	The index of the t-th aggregation
T	The number of aggregation times
$\mathbf{w}$	The vector of model parameters
$F(\mathbf{w})$	Global loss function
$F_i(\mathbf{w})$	Local loss function from the i-th client
μ	A presetting constant of the proximal term
$\mathbf{w}_i^{(t)}$	Local uploading parameters of the i-th client
$\mathbf{w}^{(0)}$	Initial parameters of the global model
$\mathbf{w}^{(t)}$	Global parameters generated from all local parameters at the t-th aggregation
$\mathbf{v}^{(t)}$	Global parameters generated from K clients' parameters at the t-th aggregation
$\mathbf{w}^*$	True optimal model parameters that minimize $F(\mathbf{w})$

II. Preliminaries

In this section, we will present preliminaries and related background knowledge on FL and DP. Also, we introduce the threat model that will be discussed in our following analysis.

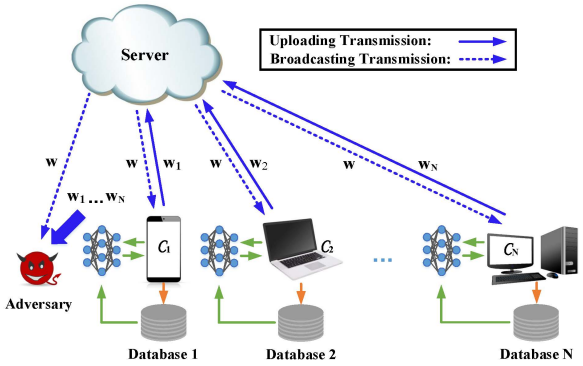


Figure 1: A FL training model with hidden adversaries who can eavesdrop trained parameters from both the clients and the server.

A. Federated Learning

Let us consider a general FL system consisting of one server and N clients, as depicted in Fig. 1. Let $\mathcal{D}_i$ denote the local database held by the client $\mathcal{C}_i$, where $i \in \{1, 2, \ldots, N\}$. At the server, the goal is to learn a model over data that resides at the N associated clients. An active client, participating in the local training, needs to find a vector $\mathbf{w}$ of an AI model to minimize a certain loss function. Formally, the server aggregates the weights received from the N clients as

$$\mathbf{w} = \sum_{i=1}^{N} p_i \mathbf{w}_i, \tag{1}$$

where $\mathbf{w}_i$ is the parameter vector trained at the i-th client, $\mathbf{w}$ is the parameter vector after aggregating at the server, N is the number of clients, $p_i = \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \geq 0$ with $\sum_{i=1}^{N} p_i = 1$, and $|\mathcal{D}| = \sum_{i=1}^{N} |\mathcal{D}_i|$ is the total size of all data samples. Such an optimization problem can be formulated as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} p_i F_i(\mathbf{w}, \mathcal{D}_i), \tag{2}$$

where $F_i(\cdot)$ is the local loss function of the i-th client. Generally, the local loss function $F_i(\cdot)$ is given by local empirical risks. The training process of such a FL system usually contains the following four steps:

- **Step 1:** *Local training:* All active clients locally compute training gradients or parameters and send locally trained ML parameters to the server;
- **Step 2:** *Model aggregating:* The server performs secure aggregation over the uploaded parameters from N clients without learning local information;
- **Step 3:** *Parameters broadcasting:* The server broadcasts the aggregated parameters to the N clients;
- **Step 4:** *Model updating:* All clients update their respective models with the aggregated parameters and test the performance of the updated models.

In the FL process, the N clients with the same data structure collaboratively learn a ML model with the help of a cloud server. After a sufficient number of local training and update exchanges between the server and its associated clients, the solution to the optimization problem (2) is able to converge to that of the global optimal learning model.

B. Threat Model

The server in this paper is assumed to be honest. However, there are external adversaries targeting at clients' private information. Although the individual dataset $\mathcal{D}_i$ of the i-th client is kept locally in FL, the intermediate parameter $\mathbf{w}_i$ needs to be shared with the server, which may reveal the clients' private information as demonstrated by model inversion attacks. For example, authors in [36] demonstrated a model-inversion attack that recovers images from a facial recognition system. In addition, the privacy leakage can also happen in the broadcasting (through downlink channels) phase by analyzing the global parameter $\mathbf{w}$.

We also assume that uplink channels are more secure than downlink broadcasting channels, since clients can be assigned to different channels (e.g., time slots, frequency bands) dynamically in each uploading time, while downlink channels are broadcasting. Hence, we assume that there are at most L ($L \leq T$) exposures of uploaded parameters from each client in the uplink[1] and T exposures of aggregated parameters in the downlink, where T is the number of aggregation times.

C. Differential Privacy

(ϵ, δ)-DP provides a strong criterion for privacy preservation of distributed data processing systems. Here, $\epsilon > 0$ is the

[1]Here we assume that the adversary cannot know where the parameters come from.

distinguishable bound of all outputs on neighboring datasets $\mathcal{D}_i, \mathcal{D}_i'$ in a database, and δ represents the event that the ratio of the probabilities for two adjacent datasets $\mathcal{D}_i, \mathcal{D}_i'$ cannot be bounded by e^ϵ after adding a privacy preserving mechanism. With an arbitrarily given δ, a privacy preserving mechanism with a larger ϵ gives a clearer distinguishability of neighboring datasets and hence a higher risk of privacy violation. Now, we will formally define DP as follows.

Definition 1: *((ϵ, δ)-DP [24]): A randomized mechanism $\mathcal{M}: \mathcal{X} \to \mathcal{R}$ with domain $\mathcal{X}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-DP, if for all measurable sets $\mathcal{S} \subseteq \mathcal{R}$ and for any two adjacent databases $\mathcal{D}_i, \mathcal{D}_i' \in \mathcal{X}$,*

$$\Pr[\mathcal{M}(\mathcal{D}_i) \in \mathcal{S}] \leq e^\epsilon \Pr[\mathcal{M}(\mathcal{D}_i') \in \mathcal{S}] + \delta. \tag{3}$$

For numerical data, a Gaussian mechanism defined in [24] can be used to guarantee (ϵ, δ)-DP. According to [24], we present the following DP mechanism by adding artificial Gaussian noises.

In order to ensure that the given noise distribution $n \sim \mathcal{N}(0, \sigma^2)$ preserves (ϵ, δ)-DP, where $\mathcal{N}$ represents the Gaussian distribution, we choose noise scale $\sigma \geq c\Delta s/\epsilon$ and the constant $c \geq \sqrt{2\ln(1.25/\delta)}$ for $\epsilon \in (0, 1)$. In this result, n is the value of an additive noise sample for a data in the dateset, Δs is the sensitivity of the function s given by $\Delta s = \max_{\mathcal{D}_i, \mathcal{D}_i'} \|s(\mathcal{D}_i) - s(\mathcal{D}_i')\|$, and s is a real-valued function.

Considering the above DP mechanism, choosing an appropriate level of noise remains a significant research problem, which will affect the privacy guarantee of clients and the convergence rate of the FL process.

III. Federated Learning with Differential Privacy

In this section, we first introduce the concept of global DP and analyze the DP performance in the context of FL. Then we propose the NbAFL scheme that can satisfy the DP requirement by adding proper noisy perturbations at both the clients and the server.

A. Global Differential Privacy

Here, we define a global (ϵ, δ)-DP requirement for both uplink and downlink channels. From the uplink perspective, using a clipping technique, we can ensure that $\|\mathbf{w}_i\| \leq C$, where $\mathbf{w}_i$ denotes training parameters from the i-th client without perturbation and C is a clipping threshold for bounding $\mathbf{w}_i$. We assume that the batch size in the local training is equal to the number of training samples and then define local training process in the i-th client by

$$\begin{aligned} s_{\mathrm{U}}^{\mathcal{D}_i} \triangleq \mathbf{w}_i &= \arg\min_{\mathbf{w}} F_i(\mathbf{w}, \mathcal{D}_i) \\ &= \frac{1}{|\mathcal{D}_i|} \sum_{j=1}^{|\mathcal{D}_i|} \arg\min_{\mathbf{w}} F_i(\mathbf{w}, \mathcal{D}_{i,j}), \end{aligned} \tag{4}$$

where $\mathcal{D}_i$ is the i-th client's database and $\mathcal{D}_{i,j}$ is the j-th sample in $\mathcal{D}_i$. Thus, the sensitivity of $s_{\mathrm{U}}^{\mathcal{D}_i}$ can be expressed as

$$\begin{aligned} \Delta s_{\mathrm{U}}^{\mathcal{D}_i} &= \max_{\mathcal{D}_i, \mathcal{D}_i'} \|s_{\mathrm{U}}^{\mathcal{D}_i} - s_{\mathrm{U}}^{\mathcal{D}_i'}\| \\ &= \max_{\mathcal{D}_i, \mathcal{D}_i'} \left\| \frac{1}{|\mathcal{D}_i|} \sum_{j=1}^{|\mathcal{D}_i|} \arg\min_{\mathbf{w}} F_i(\mathbf{w}, \mathcal{D}_{i,j}) \right. \\ &\quad \left. - \frac{1}{|\mathcal{D}_i'|} \sum_{j=1}^{|\mathcal{D}_i'|} \arg\min_{\mathbf{w}} F_i(\mathbf{w}, \mathcal{D}_{i,j}') \right\| = \frac{2C}{|\mathcal{D}_i|}, \end{aligned} \tag{5}$$

where $\mathcal{D}_i'$ is an adjacent dataset to $\mathcal{D}_i$ which has the same size but only differ by one sample, and $\mathcal{D}_{i,j}'$ is the j-th sample in $\mathcal{D}_i'$. From the above result, a global sensitivity in the uplink channel can be defined by

$$\Delta s_{\mathrm{U}} \triangleq \max\left\{\Delta s_{\mathrm{U}}^{\mathcal{D}_i}\right\}, \quad \forall i. \tag{6}$$

To achieve a small global sensitivity, the ideal condition is that all the clients use sufficient local datasets for training. Hence, we define the minimum size of the local datasets by m and then obtain $\Delta s_{\mathrm{U}} = \frac{2C}{m}$. To ensure (ϵ, δ)-DP for each client in the uplink in one exposure, we set the noise scale, represented by the standard deviation of the additive Gaussian noise, as $\sigma_{\mathrm{U}} = c\Delta s_{\mathrm{U}}/\epsilon$. Considering L exposures of local parameters, we need to set $\sigma_{\mathrm{U}} = cL\Delta s_{\mathrm{U}}/\epsilon$ due to the linear relation between ϵ and σ_{U} in the Gaussian mechanism.

From the downlink perspective, the aggregation operation for $\mathcal{D}_i$ can be expressed as

$$s_{\mathrm{D}}^{\mathcal{D}_i} \triangleq \mathbf{w} = p_1\mathbf{w}_1 + \ldots + p_i\mathbf{w}_i + \ldots + p_N\mathbf{w}_N, \tag{7}$$

where $1 \leq i \leq N$ and $\mathbf{w}$ is the aggregated parameters at the server to be broadcast to the clients. Regarding the sensitivity of $s_{\mathrm{D}}^{\mathcal{D}_i}$, i.e., $\Delta s_{\mathrm{D}}^{\mathcal{D}_i}$, we have the following lemma.

Lemma 1 (Sensitivity for the aggregation operation): *In FL training process, the sensitivity for $\mathcal{D}_i$ after the aggregation operation $s_{\mathrm{D}}^{\mathcal{D}_i}$ is given by*

$$\Delta s_{\mathrm{D}}^{\mathcal{D}_i} = \frac{2Cp_i}{m}. \tag{8}$$

Proof: See Appendix A. ∎

Remark 1: *From the above lemma, to achieve a small global sensitivity in the downlink channel which is defined by*

$$\Delta s_{\mathrm{D}} \triangleq \max\left\{\Delta s_{\mathrm{D}}^{\mathcal{D}_i}\right\} = \max\left\{\frac{2Cp_i}{m}\right\}, \quad \forall i, \tag{9}$$

the ideal condition is that all the clients should use the same size of local datasets for training, i.e., $p_i = 1/N$.

From the above remark, when setting $p_i = 1/N$, $\forall i$, we can obtain the optimal value of the sensitivity Δs_{D}. So here we should add noise at the client side first and then decide whether or not to add noises at server to satisfy the (ϵ, δ)-DP criterion in the downlink channel.

Theorem 1 (DP guarantee for downlink channels): *To ensure (ϵ, δ)-DP in the downlink channels with T aggregations,*

Algorithm 1: Noising before Aggregation FL

Data: T, $\mathbf{w}^{(0)}$, μ, ϵ and δ

Initialization: $t = 1$ and $\mathbf{w}_i^{(0)} = \mathbf{w}^{(0)}$, $\forall i$

while $t \leq T$ **do**

  Local training process:

  while $\mathcal{C}_i \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$ **do**

    Update the local parameters $\mathbf{w}_i^{(t)}$ as

    $$\mathbf{w}_i^{(t)} = \arg\min_{\mathbf{w}_i} \left(F_i(\mathbf{w}_i) + \tfrac{\mu}{2}\|\mathbf{w}_i - \mathbf{w}^{(t-1)}\|^2\right)$$

    Clip the local parameters

    $$\mathbf{w}_i^{(t)} = \mathbf{w}_i^{(t)} / \max\left(1, \tfrac{\|\mathbf{w}_i^{(t)}\|}{C}\right)$$

    Add noise and upload parameters

    $$\widetilde{\mathbf{w}}_i^{(t)} = \mathbf{w}_i^{(t)} + \mathbf{n}_i^{(t)}$$

  Model aggregating process:

  Update the global parameters $\mathbf{w}^{(t)}$ as

  $$\mathbf{w}^{(t)} = \sum_{i=1}^{N} p_i \widetilde{\mathbf{w}}_i^{(t)}$$

  The server broadcasts global noised parameters

  $$\widetilde{\mathbf{w}}^{(t)} = \mathbf{w}^{(t)} + \mathbf{n}_{\mathrm{D}}^{(t)}$$

  Local testing process:

  while $\mathcal{C}_i \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$ **do**

    Test the aggregating parameters $\widetilde{\mathbf{w}}^{(t)}$ using local dataset

  $t \leftarrow t + 1$

Result: $\widetilde{\mathbf{w}}^{(T)}$

the standard deviation of Gaussian noises $\mathbf{n}_{\mathrm{D}}$ that are added to the aggregated parameter $\boldsymbol{w}$ by the server can be given as

$$\sigma_{\mathrm{D}} = \begin{cases} \frac{2cC\sqrt{T^2 - L^2 N}}{mN\epsilon} & T > L\sqrt{N}, \\ 0 & T \leq L\sqrt{N}. \end{cases} \tag{10}$$

Proof: See Appendix B. ∎

Theorem 1 shows that to satisfy a (ϵ, δ)-DP requirement for the downlink channels, additional noises $\mathbf{n}_{\mathrm{D}}$ need to be added by the server. With a certain L, the standard deviation of additional noises is depending on the relationship between the number of aggregation times T and the number of clients N. The intuition is that a larger T can lead to a higher chance of information leakage, while a larger number of clients is helpful for hiding their private information. This theorem also provides the variance value of the noises that should be added to the aggregated parameters. Based on the above results, we propose the following NbAFL algorithm.

B. Proposed NbAFL

Algorithm 1 outlines our NbAFL for training an effective model with a global (ϵ, δ)-DP requirement. We denote by μ the presetting constant of the proximal term and by $\mathbf{w}^{(0)}$ the initiate global parameter. At the beginning of this algorithm, the server broadcasts the required privacy level parameters (ϵ, δ) are set and the initiate global parameter $\mathbf{w}^{(0)}$ are sent to clients. In the t-th aggregation, N active clients respectively train the parameters by using local databases with preset termination conditions. After completing the local training, the i-th client, $\forall i$, will add noises to the trained parameters $\mathbf{w}_i^{(t)}$, and upload the noised parameters $\widetilde{\mathbf{w}}_i^{(t)}$ to the server for aggregation.

Then the server update the global parameters $\mathbf{w}^{(t)}$ by aggregating the local parameters integrated with different weights. Additive noises $\mathbf{n}_{\mathrm{D}}^{(t)}$ are added to this $\mathbf{w}^{(t)}$ according to **Theorem 1** before being broadcast to the clients. Based on the received global parameters $\widetilde{\mathbf{w}}^{(t)}$, each client will estimate the accuracy by using local testing databases and start the next round of training process based on these received parameters. The FL process completes after the aggregation time reaches a preset number T and the algorithm returns $\widetilde{\mathbf{w}}^{(T)}$.

Now, let us focus on the privacy preservation performance of the NbAFL. First, the set of all local parameters are received by the server. Owing to the local perturbations in the NbAFL, it will be difficult for malicious adversaries to infer the information at the i-client from its uploaded parameters $\widetilde{\mathbf{w}}_i$. After the model aggregation, the aggregated parameters $\mathbf{w}$ will be sent back to clients via broadcast channels. This poses threats on clients's privacy as potential adversaries may reveal sensitive information about individual clients from $\mathbf{w}$. In this case, additive noises may be posed to $\mathbf{w}$ based on **Theorem 1**.

IV. Convergence Analysis on NbAFL

In this section, we are ready to analyze the convergence performance of the proposed NbAFL. First, we analyze the expected increment of adjacent aggregations in the loss function with Gaussian noises. Then, we focus on deriving the convergence property under the global (ϵ, δ)-DP requirement.

For the convenience of the analysis, we make the following assumptions on the loss function and network parameters.

Assumption 1: *We make assumptions on the global loss function $F(\cdot)$ defined by $F(\cdot) \triangleq \sum_{i=1}^{N} p_i F_i(\cdot)$, and the i-th local loss function $F_i(\cdot)$ as follows:*

1) *$F_i(\mathbf{w})$ is convex;*
2) *$F_i(\mathbf{w})$ satisfies the Polyak-Lojasiewicz condition with the positive parameter l, which implies that $F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{1}{2l}\|\nabla F(\mathbf{w})\|^2$, where $\mathbf{w}^*$ is the optimal result;*
3) *$F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*) = \Theta$;*
4) *$F_i(\mathbf{w})$ is β-Lipschitz, i.e., $\|F_i(\mathbf{w}) - F_i(\mathbf{w}')\| \leq \beta\|\mathbf{w} - \mathbf{w}'\|$, for any $\mathbf{w}$, $\mathbf{w}'$;*
5) *$F_i(\mathbf{w})$ is ρ-Lipschitz smooth, i.e., $\|\nabla F_i(\mathbf{w}) - \nabla F_i(\mathbf{w}')\| \leq \rho\|\mathbf{w} - \mathbf{w}'\|$, for any $\mathbf{w}$, $\mathbf{w}'$, where ρ is a constant determined by the practical loss function;*
6) *For any i and $\mathbf{w}$, $\|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\| \leq \varepsilon_i$, where ε_i is the divergence metric.*

Similar to the gradient divergence, the divergence metric ε_i is the metric to capture the divergence between the gradients of the local loss functions and that of the aggregated loss function, which is essential for analyzing SGD. The divergence is related to how the data is distributed at different nodes. Using **Assumption 1** and assume $\|\nabla F(\mathbf{w})\|$ to be uniformly away from zero, we then have the following lemma.

Lemma 2 (B-dissimilarity of various clients): *For a given ML parameter* $\mathbf{w}$*, there exists* B *satisfying*

$$\mathbb{E}\left\{\|\nabla F_i(\mathbf{w})\|^2\right\} \leq \|\nabla F(\mathbf{w})\|^2 B^2, \quad \forall i. \tag{11}$$

Proof: See Appendix C. ∎

Lemma 2 comes from the assumption of the divergence metric and demonstrates the statistical heterogeneity of all clients. As mentioned earlier, the values of ρ and B are determined by the specific global loss function $F(\mathbf{w})$ in practice and the training parameters $\mathbf{w}$. With the above preparation, we are now ready to analyze the convergence property of NbAFL. First, we present the following lemma to derive an expected increment bound on the loss function during each iteration of parameters with artificial noises.

Lemma 3 (Expected increment in the loss function): *After receiving updates, from the* t*-th to the* $(t+1)$*-th aggregation, the expected difference in the loss function can be upper-bounded by*

$$\begin{aligned}\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t+1)}) - F(\widetilde{\mathbf{w}}^{(t)})\} &\leq \lambda_2 \mathbb{E}\{\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\|^2\} \\ &+ \lambda_1 \mathbb{E}\{\|\mathbf{n}^{(t+1)}\| \|\nabla F(\widetilde{\mathbf{w}}^{(t)})\|\} + \lambda_0 \mathbb{E}\{\|\mathbf{n}^{(t+1)}\|^2\}, \end{aligned} \tag{12}$$

where $\lambda_0 = \frac{\rho}{2}$, $\lambda_1 = \frac{1}{\mu} + \frac{\rho B}{\mu}$, $\lambda_2 = -\frac{1}{\mu} + \frac{\rho B}{\mu^2} + \frac{\rho B^2}{2\mu^2}$ *and* $\mathbf{n}^{(t)}$ *are the equivalent noises imposed on the parameters after the* t*-th aggregation, given by* $\mathbf{n}^{(t)} = \sum_{i=1}^{N} p_i \mathbf{n}_i^{(t)} + \mathbf{n}_{\mathrm{D}}^{(t)}$.

Proof: See Appendix D. ∎

In this lemma, the value of an additive noise sample n in vector $\mathbf{n}^{(t)}$ satisfies the following Gaussian distribution $n \sim \mathcal{N}(0, \sigma_{\mathrm{A}}^2)$. Also, we can obtain $\sigma_{\mathrm{A}} = \sqrt{\sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{U}}^2/N}$ from Section III. From the right hand side (RHS) of the above inequality, we can see that it is crucial to select a proper proximal term μ to achieve a low upper-bound. It is clear that artificial noises with a large σ_{A} may improve the DP performance in terms privacy protection. However, from the RHS of (12), a large σ_{A} may enlarge the expected difference of the loss function between two consecutive aggregations, leading to a deterioration of convergence performance.

Furthermore, to satisfy the global (ϵ, δ)-DP, by using **Theorem 1**, we have

$$\sigma_{\mathrm{A}} = \begin{cases} \frac{cT\Delta s_{\mathrm{D}}}{\epsilon} & T > L\sqrt{N}, \\ \frac{cL\Delta s_{\mathrm{U}}}{\sqrt{N}\epsilon} & T \leq L\sqrt{N}. \end{cases} \tag{13}$$

Next, we will analyze the convergence property of NbAFL with the (ϵ, δ)-DP requirement.

Theorem 2 (Convergence upper bound of the NbAFL): *With required protection level* ϵ*, the convergence upper bound of* **Algorithm 1** *after* T *aggregations is given by*

$$\begin{aligned}&\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(T)}) - F(\mathbf{w}^*)\} \leq \\ &\qquad P^T\Theta + \left(\frac{\kappa_1 T}{\epsilon} + \frac{\kappa_0 T^2}{\epsilon^2}\right)\left(1 - P^T\right), \end{aligned} \tag{14}$$

where $P = 1 + 2l\lambda_2$, $\kappa_1 = \frac{\lambda_1 \beta c C}{m(1-P)}\sqrt{\frac{2}{N\pi}}$ *and* $\kappa_0 = \frac{\lambda_0 c^2 C^2}{m^2(1-P)N}$.

Proof: See Appendix D. ∎

Theorem 2 reveals an important relationship between privacy and utility by taking into account the protection level ϵ and the number of aggregation times T. As the number of aggregation times T increases, the first term of the upper bound decreases but the second term increases. Furthermore, By viewing T as a continuous variable and by writing the RHS of (14) as $h(T)$, we have

$$\begin{aligned}\frac{d^2h(T)}{d^2T} &= \left(\Theta - \frac{\kappa_1 T}{\epsilon} - \frac{\kappa_0 T^2}{\epsilon^2}\right) P^T \ln^2 P \\ &\quad - 2\left(\frac{\kappa_1}{\epsilon} + \frac{2\kappa_0 T}{\epsilon^2}\right) P^T \ln P + \frac{2\kappa_0}{\epsilon^2}\left(1 - P^T\right). \end{aligned} \tag{15}$$

It can be seen that the second term and third term of on the RHS of (15) are always positive. When N and ϵ are set to be large enough, we can see that κ_1 and κ_0 are small, and thus the first term can also be positive. In this case, we have $d^2h(T)/d^2T > 0$ and the upper bound is convex for T.

Remark 2: *As can be seen from this theorem, the expected gap between the achieved loss function* $F(\widetilde{\mathbf{w}}^{(T)})$ *and the minimum one* $F(\mathbf{w}^*)$ *is a decreasing function of* ϵ*. By increasing* ϵ*, i.e., relaxing the privacy protection level, the performance of NbAFL algorithm will improve. This is reasonable because the variance of artificial noises decreases, thereby improving the convergence performance.*

Remark 3: *The number of clients* N *will also affect its iterative convergence performance, i.e., a larger* N *would achieve a better convergence performance. This is because a lager* N *leads to a lower variance of the artificial noises.*

Remark 4: *There is an optimal number of maximum aggregation times* T *in terms of convergence performance for given* ϵ *and* N*. In more detail, a larger* T *may lead to a higher variance of artificial noises, and thus pose a negative impact on convergence performance. On the other hand, more iterations can generally boost the convergence performance if noises are not large enough. In this sense, there is a tradeoff on choosing a proper* T*.*

V. K-Client Random Scheduling Policy

In this section, we consider the case where only $K (K < N)$ clients are selected to participate in the aggregation process, namely K-client random scheduling.

We now discuss how to add artificial noises in the K-client random scheduling to satisfy a global (ϵ, δ)-DP. It is obvious that in the uplink channels, each of the K scheduled clients should add noises with scale $\sigma_{\mathrm{U}} = cL\Delta s_{\mathrm{U}}/\epsilon$ for achieving (ϵ, δ)-DP. This is equivalent to the noise scale in the all-clients selection case in **Section III**, since each client only considers its own privacy for uplink channels in both cases. However, the derivation of the noise scale in the downlink will be different for the K-client random scheduling. As an extension of **Theorem 1**, we present the following lemma in the case of K-client random scheduling on how to obtain σ_{D}.

Lemma 4 (DP guarantee for K-client random scheduling): *In the NbAFL algorithm with* K*-client random scheduling, to satisfy a global* (ϵ, δ)*-DP, and the standard deviation* σ_{D}

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.2988575, IEEE Transactions on Information Forensics and Security

of additive Gaussian noises for downlink channels should be set as

$$\sigma_{\mathrm{D}} = \begin{cases} \frac{2cC\sqrt{\frac{T^2}{b^2}-L^2K}}{mK\epsilon} & T > \frac{\epsilon}{\gamma}, \\ 0 & T \leq \frac{\epsilon}{\gamma}, \end{cases} \tag{16}$$

where $b = -\frac{T}{\epsilon}\ln\left(1-\frac{N}{K}+\frac{N}{K}e^{\frac{-\epsilon}{T}}\right)$ *and* $\gamma = -\ln\left(1-\frac{K}{N}+\frac{K}{N}e^{\frac{-\epsilon}{L\sqrt{K}}}\right)$.

Proof: See Appendix F. ∎

Lemma 4 recalculates σ_{D} by considering the number of chosen clients K. Generally, the number of clients N is fixed, we thus focus on the effect of K. Based on the DP analysis in **Lemma 4**, we can obtain the following theorem.

Theorem 3 (Convergence under K-client random scheduling): *With required protection level ϵ and the number of chosen clients K, for any $\Theta > 0$, the convergence upper bound after T aggregation times is given by*

$$\begin{aligned} \mathbb{E}\{F(\widetilde{\mathbf{v}}^T) - F(\mathbf{w}^*)\} &\leq Q^T\Theta \\ &+ \frac{1-Q^T}{1-Q}\left(\frac{cC\alpha_1\beta}{-mK\ln\left(1-\frac{N}{K}+\frac{N}{K}e^{-\frac{\epsilon}{T}}\right)}\sqrt{\frac{2}{\pi}}\right. \\ &\left. + \frac{c^2C^2\alpha_0}{m^2K^2\ln^2\left(1-\frac{N}{K}+\frac{N}{K}e^{-\frac{\epsilon}{T}}\right)}\right). \end{aligned} \tag{17}$$

where $Q = 1+\frac{2l}{\mu^2}\left(\frac{\rho B^2}{2}+\rho B+\frac{\rho B^2}{K}+\frac{2\rho B^2}{\sqrt{K}}+\frac{\mu B}{\sqrt{K}}-\mu\right)$, $\alpha_0 = \frac{2\rho K}{N}+\rho$, $\alpha_1 = 1+\frac{2\rho B}{\mu}+\frac{2\rho B\sqrt{K}}{\mu N}$, *and* $\widetilde{\mathbf{v}}^{(T)} = \sum_{i=1}^{K} p_i\left(\mathbf{w}_i^{(T)}+\mathbf{n}_i^{(T)}\right)+\mathbf{n}_{\mathrm{D}}^{(T)}$.

Proof: See Appendix G. ∎

The above theorem provides the convergence upper bound between $F(\widetilde{\mathbf{v}}^T)$ and $F(\mathbf{w}^*)$ under K-random scheduling. Using K-client random scheduling, we can obtain an important relationship between privacy and utility by taking into account the protection level ϵ, the number of aggregation times T and the number of chosen clients K.

Remark 5: *From the bound derived in* **Theorem 3***, we conclude that there is an optimal K in between 0 and N that achieves the optimal convergence performance. That is, by finding a proper K, the K-client random scheduling policy is superior to the one that all N clients participate in the FL aggregations.*

VI. Simulation Results

In this section, we evaluate the proposed NbAFL by using multi-layer perception (MLP) and real-world federated datasets. In order to characterize the convergence property of NbAFL, we conduct experiments by varying the protection levels of ϵ, the number of clients N, the number of maximum aggregation times T and the number of chosen clients K.

We conduct experiments on the standard MNIST dataset for handwritten digit recognition consisting of 60000 training examples and 10000 testing examples [37]. Each example is a 28×28 size gray-level image. Our baseline model uses a a MLP network with a single hidden layer containing 256 hidden units. In this feed-forward neural network, we use a ReLU units and softmax of 10 classes (corresponding to the 10 digits). For the optimizer of networks, we set the learning rate to 0.002. Then, we evaluate this MLP for the multi-class classification task with the standard MNIST dataset, namely, recognizing from 0 to 9, where each client has 100 training samples locally. This setting is in line with the ideal condition in **Remark 1**.

We can note that parameter clipping C is a popular ingredient of SGD and ML for non-privacy reasons. A proper value of clipping threshold C should be considered for the DP based FL framework. In the following experiments (except subsection B), we utilize the method in [29] and choose C by taking the median of the norms of the unclipped parameters over the course of training. The values of ρ, β, l and B are determined by the specific loss function, and we will use estimated values in our simulations [20].

A. Performance Evaluation on Protection Levels

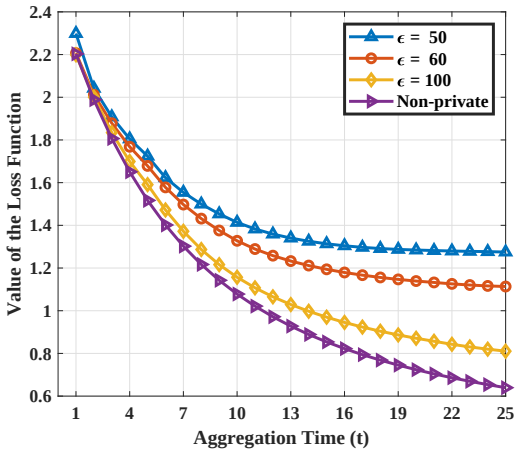


Figure 2: The comparison of training loss with various protection levels for 50 clients using $\epsilon = 50$, $\epsilon = 60$ and $\epsilon = 100$, respectively.

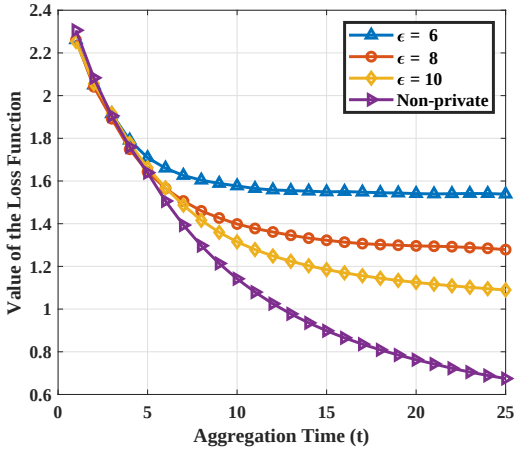


Figure 3: The comparison of training loss with various protection level for 50 clients using $\epsilon = 6$, $\epsilon = 8$ and $\epsilon = 10$, respectively.

In Figs. 2, we choose various protection levels $\epsilon = 50$, $\epsilon = 60$ and $\epsilon = 100$ to show the results of the loss function in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.2988575, IEEE Transactions on Information Forensics and Security

NbAFL. Furthermore, we also include a non-private approach to compare with our NbAFL. In this experiment, we set $N = 50$, $T = 25$ and $\delta = 0.01$, and compute the values of the loss function as a function of the aggregation times t. As shown in Fig. 2, values of the loss function in NbAFL are decreasing as we relax the privacy guarantees (increasing ϵ). Such observation results are in line with **Remark 2**. We also choose high protection levels $\epsilon = 6$, $\epsilon = 8$ and $\epsilon = 10$ for this experiment, where each client has 512 training samples locally. We set $N = 50$, $T = 25$ and $\delta = 0.01$. From Fig. 3, we can draw a similar conclusion as in **Remark 2** that values of the loss function in NbAFL are decreasing as we relax the privacy guarantees.

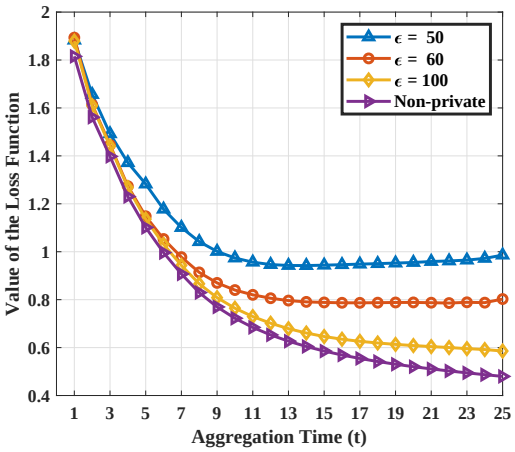


Figure 4: The comparison of training loss with various privacy levels for 50 clients using $\epsilon = 50$, $\epsilon = 60$ and $\epsilon = 100$, respectively.

Considering the K-client random scheduling, in Fig. 4, we investigate the performances with various protection levels $\epsilon = 50$, $\epsilon = 60$ and $\epsilon = 100$. For simulation parameters, we set $N = 50$, $K = 20$, $T = 25$, and $\delta = 0.01$. As shown in Figs. 4, the convergence performance under the K-client random scheduling is improved with an increasing ϵ.

B. Impact of the Clipping Threshold C

In Fig. 5, we choose various clipping thresholds $C = 10, 15, 20$ and 25 to show the results of the loss function for 50 clients using $\epsilon = 60$ in NbAFL. As shown in Fig. 5, when $C = 20$, the convergence performance of NbAFL can obtain the best value. We can note that limiting the parameter norm has two opposing effects. On the one hand, if the clipping threshold C is too small, clipping destroys the intended gradient direction of parameters. On the other hand, increasing the norm bound C forces us to add more noise to the parameters because of its effect on the sensitivity.

C. Impact of the Number of Clients N

Figs. 6 compares the convergence performance of NbAFL under required protection level $\epsilon = 60$ and $\delta = 10^{-2}$ as a function of clients' number, N. In this experiment, we set $N = 50$, $N = 60$, $N = 80$ and $N = 100$. We notice that the performance among different numbers of clients is

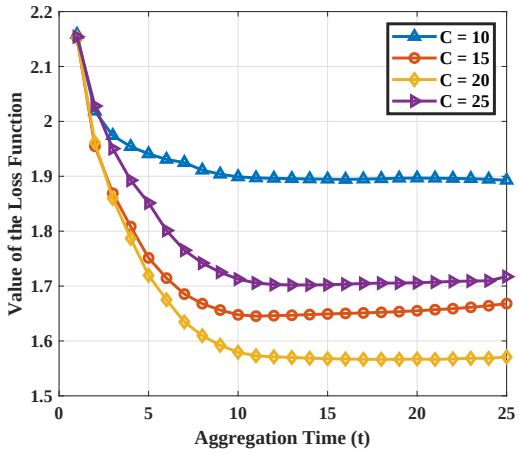


Figure 5: The comparison of training loss with various clipping thresholds for 50 clients using $\epsilon = 60$.

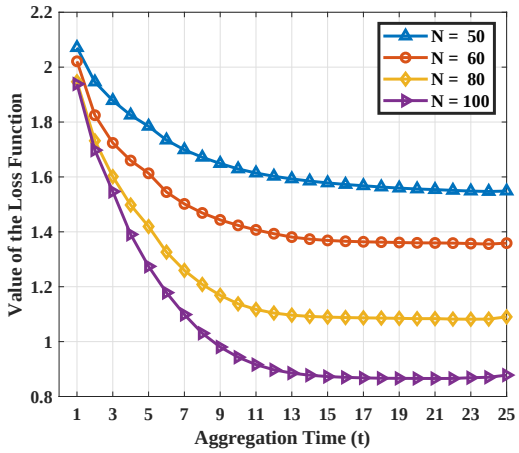


Figure 6: The value of the loss function with various numbers of clients under $\epsilon = 60$ under NbAFL Algorithm with 50 clients.

governed by **Remark 3**. This is because that more clients not only provide larger global datasets for training, but also bring down the standard deviation of additive noises due to the aggregation.

D. Impact of the Number of Maximum Aggregation Times T

In Fig. 7, we show the experimental results of training loss as a function of maximum aggregation times with various privacy levels $\epsilon = 50, 60, 80$ and 100 under NbAFL algorithm. This observation is in line with **Remark 4**, and the reason comes from the fact that a lower privacy level decreases the standard deviation of additive noises and the server can obtain better quality ML model parameters from the clients. Fig. 7 also implies that an optimal number of maximum aggregation times increases almost with respect to the increasing ϵ.

In Fig. 9, we plot the values of the loss function in the normalized NbAFL using solid lines and the K-random scheduling based NbAFL using dotted lines with various numbers of maximum aggregation times. This figure shows that the value of loss function is a convex function of maximum aggregation times for a given protection level under NbAFL

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.2988575, IEEE Transactions on Information Forensics and Security

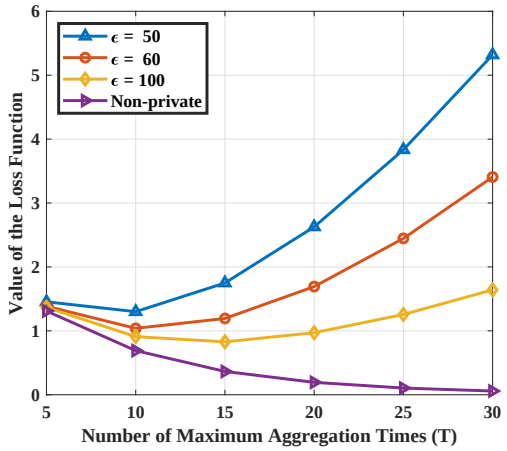


Figure 7: The convergence upper bounds with various privacy levels $\epsilon = 50$, 60 and 100 under 50-clients' NbAFL algorithm.

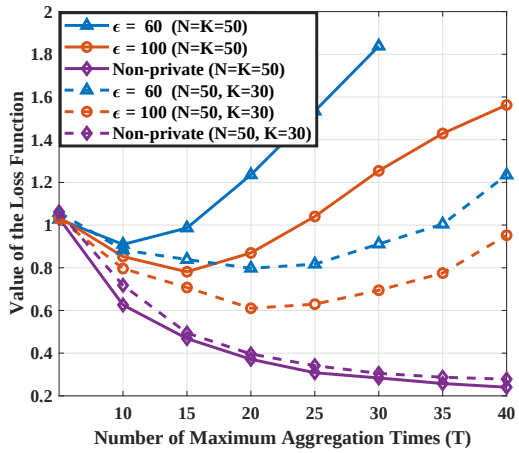


Figure 9: The value of the loss function with various privacy levels $\epsilon = 60$ and $\epsilon = 100$ under NbAFL Algorithm with 50 clients.

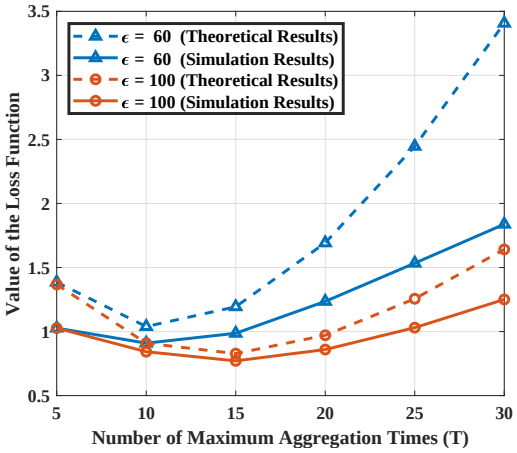


Figure 8: The comparison of the loss function between experimental and theoretical results with the various aggregation times under NbAFL Algorithm with 50 clients.

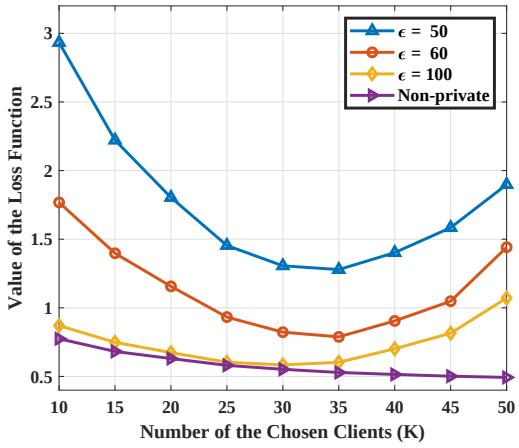


Figure 10: The value of the loss function with various numbers of chosen clients under $\epsilon = 50, 60, 100$ under NbAFL Algorithm and non-private approach with 50 clients.

algorithm, which validates **Remark 4**. From Fig. 9, we can also see that for a given ϵ, K-client random scheduling based NbAFL algorithm has a better convergence performance than the normalized NbAFL algorithm for a larger T. This is because that K-client random scheduling can bring down the variance of artificial noises with little performance loss.

E. Impact of the Number of Chosen Clients K

In Fig. 10, we plot values of the loss function with various numbers of chosen clients K under the random scheduling policy in NbAFL. The number of clients is $N = 50$, and K clients are randomly chosen to participate in training and aggregation in each iteration. In this experiment, we set $\epsilon = 50$, $\epsilon = 60$, $\epsilon = 100$ and $\delta = 0.01$. Meanwhile, we also exhibit the performance of the non-private approach with various numbers of chosen clients K. Note that an optimal K which further improves the convergence performance exists for various protection levels, due to a trade-off between enhance privacy protection and involving larger global training datasets in each model updating round. This observation is in line with **Remark 5**. The figure shows that in NbAFL, for a given protection level ϵ, the K-client random scheduling can obtain a better tradeoff than the normal selection policy.

VII. Conclusions

In this paper, we have focused on information leakage in SGD based FL. We have first defined a global (ϵ, δ)-DP requirement for both uplink and downlink channels, and developed variances of artificial noises at clients and server sides. Then, we have proposed a novel framework based on the concept of global (ϵ, δ)-DP, named NbAFL. We have developed theoretically a convergence bound on the loss function of the trained FL model in the NbAFL. From theoretical convergence bounds, we have obtained the following results: 1) There is a tradeoff between the convergence performance and privacy protection levels, i.e., better convergence performance leads to a lower protection level; 2) Increasing the number N of overall clients participating in FL can improve the convergence performance, given a fixed privacy protection level; and 3) There is an optimal number of maximum aggregation times in terms of convergence performance for a given protection level. Furthermore, we have proposed a

K-client random scheduling strategy and also developed a corresponding convergence bound on the loss function in this case. In addition to the above three properties. we find that there exists an optimal value of K that achieves the best convergence performance at a fixed privacy level. Extensive simulation results confirm the correctness of our analysis. Therefore, our analytical results are helpful for the design on privacy-preserving FL architectures with different tradeoff requirements on convergence performance and privacy levels.

We can note that the size and the distribution of data both greatly affect the quality of the FL training. As a future work, it is of great interest to analytically evaluate the convergence performance of NbAFL with varying size and distribution of data at client sides.

Appendix A
Proof of Lemma 1

From the downlink perspective, for all $\mathcal{D}_i$ and $\mathcal{D}_i'$ which differ in a signal entry, the sensitivity can be expressed as

$$\Delta s_{\mathrm{D}}^{\mathcal{D}_i} = \max_{\mathcal{D}_i, \mathcal{D}_i'} \| s_{\mathrm{D}}^{\mathcal{D}_i} - s_{\mathrm{D}}^{\mathcal{D}_i'} \|. \tag{18}$$

Based on (4) and (7), we have

$$s_{\mathrm{D}}^{\mathcal{D}_i} = p_1 \mathbf{w}_1(\mathcal{D}_1) + \ldots + p_i \mathbf{w}_i(\mathcal{D}_i) + \ldots + p_N \mathbf{w}_N(\mathcal{D}_N) \tag{19}$$

and

$$s_{\mathrm{D}}^{\mathcal{D}_i'} = p_1 \mathbf{w}_1(\mathcal{D}_1) + \ldots + p_i \mathbf{w}_i(\mathcal{D}_i') + \ldots + p_N \mathbf{w}_N(\mathcal{D}_N), \tag{20}$$

Furthermore, the sensitivity can be given as

$$\begin{aligned} \Delta s_{\mathrm{D}}^{\mathcal{D}_i} &= \max_{\mathcal{D}_i, \mathcal{D}_i'} \| p_i \mathbf{w}_i(\mathcal{D}_i) - p_i \mathbf{w}_i(\mathcal{D}_i') \| \\ & p_i \max_{\mathcal{D}_i, \mathcal{D}_i'} \| \mathbf{w}_i(\mathcal{D}_i) - \mathbf{w}_i(\mathcal{D}_i') \| = p_i \Delta s_{\mathrm{U}}^{\mathcal{D}_i} \leq \frac{2 C p_i}{m}. \end{aligned} \tag{21}$$

Hence, we know $\Delta s_{\mathrm{D}}^{\mathcal{D}_i} = \frac{2Cp_i}{m}$. This completes the proof. □

Appendix B
Proof of Theorem 1

To ensure a global (ϵ, δ)-DP in the uplink channels, the standard deviation of additive noises in client sides can be set to $\sigma_{\mathrm{U}} = cL\Delta s_{\mathrm{U}}/\epsilon$ due to the linear relation between ϵ and σ_{U} with Gaussian mechanism, where $\Delta s_{\mathrm{U}} = \frac{2C}{m}$ is the sensitivity for the aggregation operation and m is the data size of each client. We then set the sample in the i-th local noise vector to a same distribution $n_i \sim \varphi(n)$ (i.i.d for all i) because each client is coincident with the same global (ϵ, δ)-DP. The aggregation process with artificial noises added by clients can be expressed as

$$\widetilde{\mathbf{w}} = \sum_{i=1}^{N} p_i \left(\mathbf{w}_i + \mathbf{n}_i \right) = \sum_{i=1}^{N} p_i \mathbf{w}_i + \sum_{i=1}^{N} p_i \mathbf{n}_i. \tag{22}$$

The distribution $\phi_N(n)$ of $\sum_{i=1}^{N} p_i n_i$ can be expressed as

$$\phi_N(n) = \bigotimes_{i=1}^{N} \varphi_i \left(n \right), \tag{23}$$

where $p_i n_i \sim \varphi_i \left(n \right)$, and $\bigotimes$ is convolutional operation.

When we use Gaussian mechanism for n_i with noise scale σ_{U}, the distribution of $p_i n_i$ is also Gaussian distribution. To obtain a small sensitivity Δs_{D}, we set $p_i = 1/N$. Furthermore, the noise scale $\sigma_{\mathrm{U}}/\sqrt{N}$ of the Gaussian distribution $\phi_N(n)$ can be calculated. To ensure a global (ϵ, δ)-DP in downlink channels, we know the standard deviation of additive noises can be set to $\sigma_{\mathrm{A}} = cT\Delta s_{\mathrm{D}}/\epsilon$, where $\Delta s_{\mathrm{D}} = 2C/mN$. Hence, we can obtain the standard deviation of additive noises at the server as

$$\sigma_{\mathrm{D}} = \sqrt{\sigma_{\mathrm{A}}^2 - \frac{\sigma_{\mathrm{U}}^2}{N}} = \begin{cases} \frac{2cC\sqrt{T^2 - L^2 N}}{mN\epsilon} & T > L\sqrt{N}, \\ 0 & T \leq L\sqrt{N}. \end{cases} \tag{24}$$

Hence, **Theorem 1** has been proved. □

Appendix C
Proof of Lemma 2

Due to **Assumption 1**, we have

$$\mathbb{E}\left\{ \|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \right\} \leq \mathbb{E}\{\varepsilon_i^2\} \tag{25}$$

and

$$\begin{aligned} &\mathbb{E}\left\{ \|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \right\} \\ &= \mathbb{E}\left\{ \|\nabla F_i(\mathbf{w})\|^2 \right\} - 2\mathbb{E}\left\{ \nabla F_i(\mathbf{w})^\top \right\} \nabla F(\mathbf{w}) \\ &+ \|\nabla F(\mathbf{w})\|^2 = \mathbb{E}\left\{ \|\nabla F_i(\mathbf{w})\|^2 \right\} - \|\nabla F(\mathbf{w})\|^2. \end{aligned} \tag{26}$$

Considering (25), (26) and $\nabla F(\mathbf{w}) = \mathbb{E}\{\nabla F_i(\mathbf{w})\}$, we have

$$\begin{aligned} \mathbb{E}\left\{ \|\nabla F_i(\mathbf{w})\|^2 \right\} &\leq \|\nabla F(\mathbf{w})\|^2 + \mathbb{E}\{\varepsilon_i^2\} \\ &= \|\nabla F(\mathbf{w})\|^2 B(\mathbf{w})^2. \end{aligned} \tag{27}$$

Note that when $\|\nabla F(\mathbf{w})\|^2 \neq 0$, there exists

$$B(\mathbf{w}) = \sqrt{1 + \frac{\mathbb{E}\{\varepsilon_i^2\}}{\|\nabla F(\mathbf{w})\|^2}} \geq 1, \tag{28}$$

which satisfies the equation. We can notice that a smaller value of $B(\mathbf{w})$ implies that the local loss functions are more locally similar. When all the local loss functions are the same, then $B(\mathbf{w}) = 1$, for all $\mathbf{w}$. Therefore, we can have

$$\mathbb{E}\left\{ \|\nabla F_i(\mathbf{w})\|^2 \right\} \leq \|\nabla F(\mathbf{w})\|^2 B^2, \quad \forall i, \tag{29}$$

where B is the upper bound of $B(\mathbf{w})$. This completes the proof. □

Appendix D
Proof of Lemma 3

Considering the aggregation process with artificial noises added by clients and the server in the $(t+1)$-th aggregation, we have

$$\widetilde{\mathbf{w}}^{(t+1)} = \sum_{i=1}^{N} p_i \mathbf{w}_i^{(t+1)} + \mathbf{n}^{(t+1)}, \tag{30}$$

where

$$\mathbf{n}^{(t)} = \sum_{i=1}^{N} p_i \mathbf{n}_i^{(t)} + \mathbf{n}_{\mathrm{D}}^{(t)}. \tag{31}$$

Because $F_i(\cdot)$ is ρ-Lipschitz smooth, we know

$$F_i(\widetilde{\mathbf{w}}^{(t+1)}) \leq F_i(\widetilde{\mathbf{w}}^{(t)}) + \nabla F_i(\widetilde{\mathbf{w}}^{(t)})^\top(\widetilde{\mathbf{w}}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}) + \frac{\rho}{2}\|\widetilde{\mathbf{w}}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|^2, \quad (32)$$

for all $\widetilde{\mathbf{w}}^{(t+1)}$, $\widetilde{\mathbf{w}}^{(t)}$. Combining $F(\widetilde{\mathbf{w}}^{(t)}) = \mathbb{E}\{F_i(\widetilde{\mathbf{w}}^{(t)})\}$ and $\nabla F(\widetilde{\mathbf{w}}^{(t)}) = \mathbb{E}\{\nabla F_i(\widetilde{\mathbf{w}}^{(t)})\}$, we have

$$\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t+1)}) - F(\widetilde{\mathbf{w}}^{(t)})\} \leq \mathbb{E}\{\nabla F(\widetilde{\mathbf{w}}^{(t)})^\top(\widetilde{\mathbf{w}}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)})\} + \frac{\rho}{2}\mathbb{E}\{\|\widetilde{\mathbf{w}}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|^2\}. \quad (33)$$

We define

$$J(\mathbf{w}_i^{(t+1)}; \widetilde{\mathbf{w}}^{(t)}) \triangleq F_i(\mathbf{w}_i^{(t+1)}) + \frac{\mu}{2}\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|^2. \quad (34)$$

Then, we know

$$\nabla J(\mathbf{w}_i^{(t+1)}; \widetilde{\mathbf{w}}^{(t)}) = \nabla F_i(\mathbf{w}_i^{(t+1)}) + \mu\left(\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\right) \quad (35)$$

and

$$\widetilde{\mathbf{w}}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)} = \sum_{i=1}^{N}\left(\mathbf{w}_i^{(t+1)} + \mathbf{n}_i^{(t+1)}\right) + \mathbf{n}_{\mathrm{D}}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}$$
$$= \frac{1}{\mu}\mathbb{E}\{\nabla J(\mathbf{w}_i^{(t+1)}; \widetilde{\mathbf{w}}^{(t)}) - \nabla F_i(\mathbf{w}_i^{(t+1)})\} + \mathbf{n}^{(t+1)}. \quad (36)$$

Because $F_i(\cdot)$ is ρ-Lipschitz smooth, we can obtain

$$\mathbb{E}\{\nabla F_i(\mathbf{w}_i^{(t+1)})\} \leq \mathbb{E}\{\nabla F_i(\widetilde{\mathbf{w}}^{(t)}) + \rho\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|\} = \nabla F(\widetilde{\mathbf{w}}^{(t)}) + \rho\mathbb{E}\{\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|\}. \quad (37)$$

Now, let us bound $\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|$. We know

$$\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\| \leq \|\mathbf{w}_i^{(t+1)} - \widehat{\mathbf{w}}_i^{(t+1)}\| + \|\widehat{\mathbf{w}}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|, \quad (38)$$

where $\widehat{\mathbf{w}}_i^{(t+1)} = \arg\min_{\mathbf{w}} J_i(\mathbf{w}; \widetilde{\mathbf{w}}^{(t)})$. Let us define $\overline{\mu} = \mu + l > 0$, then we know $J_i(\mathbf{w}; \widetilde{\mathbf{w}}^{(t)})$ is $\overline{\mu}$-convexity. Based on this, we can obtain

$$\|\widehat{\mathbf{w}}_i^{(t+1)} - \mathbf{w}_i^{(t+1)}\| \leq \frac{\theta}{\overline{\mu}}\|\nabla F_i(\widetilde{\mathbf{w}}^{(t)})\| \quad (39)$$

and

$$\|\widehat{\mathbf{w}}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\| \leq \frac{1}{\overline{\mu}}\|\nabla F_i(\widetilde{\mathbf{w}}^{(t)})\|, \quad (40)$$

where θ denotes a θ solution of $\min_{\mathbf{w}} J_i(\mathbf{w}; \widetilde{\mathbf{w}}^{(t)})$, which is defined in [19]. Now, we can use the inequality (39) and (40) to obtain

$$\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\| \leq \frac{1+\theta}{\overline{\mu}}\|\nabla F_i(\widetilde{\mathbf{w}}^{(t)})\|. \quad (41)$$

Therefore,

$$\|\widetilde{\mathbf{w}}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\| \leq \|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\| + \|\mathbf{n}^{(t+1)}\| \leq \mathbb{E}\{\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|\} + \|\mathbf{n}^{(t+1)}\| \leq \frac{1+\theta}{\overline{\mu}}\mathbb{E}\{\|\nabla F_i(\widetilde{\mathbf{w}}^{(t)})\|\} + \|\mathbf{n}^{(t+1)}\| \leq \frac{B(1+\theta)}{\overline{\mu}}\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\| + \|\mathbf{n}^{(t+1)}\|. \quad (42)$$

Using (37) and (38), we know

$$\|\mathbb{E}\{\nabla F_i(\mathbf{w}_i^{(t+1)})\} - \nabla F(\widetilde{\mathbf{w}}^{(t)}) - \mathbb{E}\{\nabla J(\mathbf{w}_i^{(t+1)}; \widetilde{\mathbf{w}}^{(t)})\}\| \leq \rho\mathbb{E}\{\|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{w}}^{(t)}\|\} + \mathbb{E}\{\|\nabla J(\mathbf{w}_i^{(t+1)}; \widetilde{\mathbf{w}}^{(t)})\|\} \leq \frac{\rho B(1+\theta)}{\overline{\mu}}\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\| + B\theta\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\|. \quad (43)$$

Substituting (37), (42) and (43) into (33), we know

$$\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t+1)}) - F(\widetilde{\mathbf{w}}^{(t)})\} \leq \mathbb{E}\left\{\nabla F(\widetilde{\mathbf{w}}^{(t)})^\top\left(-\frac{1}{\mu}\nabla F(\widetilde{\mathbf{w}}^{(t)}) + \frac{1}{\mu}\mathbf{n}^{(t+1)} + \left(\frac{\rho B(1+\theta)}{\mu\overline{\mu}} + \frac{B\theta}{\mu}\right)\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\|\right)\right\} + \frac{\rho}{2}\mathbb{E}\left\{\left[\frac{B(1+\theta)}{\overline{\mu}}\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\| + \|\mathbf{n}^{(t+1)}\|\right]^2\right\}. \quad (44)$$

Then, using triangle inequation, we can obtain

$$\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t+1)}) - F(\widetilde{\mathbf{w}}^{(t)})\} \leq \lambda_2\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\|^2 + \lambda_1\mathbb{E}\{\|\mathbf{n}^{(t+1)}\|\}\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\| + \lambda_0\mathbb{E}\{\|\mathbf{n}^{(t+1)}\|^2\}. \quad (45)$$

where

$$\lambda_2 = -\frac{1}{\mu} + \frac{B}{\mu}\left[\frac{\rho(1+\theta)}{\overline{\mu}} + \theta\right] + \frac{\rho B^2(1+\theta)^2}{2\overline{\mu}^2}, \quad (46)$$

$$\lambda_1 = \frac{1}{\mu} + \frac{\rho B(1+\theta)}{\overline{\mu}} \text{ and } \lambda_0 = \frac{\rho}{2}. \quad (47)$$

In this convex case, where $\overline{\mu} = \mu$, if $\theta = 0$, all subproblems are solved accurately. We know $\lambda_2 = -\frac{1}{\mu} + \frac{\rho B}{\mu^2} + \frac{\rho B^2}{2\mu^2}$, $\lambda_1 = \frac{1}{\mu} + \frac{\rho B}{\mu}$ and $\lambda_0 = \frac{\rho}{2}$. This completes the proof. □

APPENDIX E
PROOF OF THEOREM 2

We assume that F satisfies the Polyak-Lojasiewicz inequality [38] with positive parameter l, which implies that

$$\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t)}) - F(\mathbf{w}^*)\} \leq \frac{1}{2l}\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\|^2. \quad (48)$$

Moreover, subtract $\mathbb{E}\{F(\mathbf{w}^*)\}$ in both sides of (45), we know

$$\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t+1)}) - F(\mathbf{w}^*)\} \leq \mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t)}) - F(\mathbf{w}^*)\} + \lambda_2\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\|^2 + \lambda_1\mathbb{E}\{\|\mathbf{n}^{(t+1)}\|\}\|\nabla F(\widetilde{\mathbf{w}}^{(t)})\| + \lambda_0\mathbb{E}\{\|\mathbf{n}^{(t+1)}\|^2\}. \quad (49)$$

Considering $\|\nabla F(\mathbf{w}^{(t)})\| \leq \beta$ and (48), we have

$$\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t+1)}) - F(\mathbf{w}^*)\} \leq (1+2l\lambda_2)\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(t)}) - F(\mathbf{w}^*)\} + \lambda_1\beta\mathbb{E}\{\|\mathbf{n}^{(t+1)}\|\} + \lambda_0\mathbb{E}\{\|\mathbf{n}^{(t+1)}\|^2\}, \quad (50)$$

where $F(\mathbf{w}^*)$ is the loss function corresponding to the optimal parameters $\mathbf{w}^*$. Considering the same and independent distribution of additive noises, we define $\mathbb{E}\{\|\mathbf{n}^{(t)}\|\} = \mathbb{E}\{\|\mathbf{n}\|\}$

and $\mathbb{E}\{\|\mathbf{n}^{(t)}\|^2\} = \mathbb{E}\{\|\mathbf{n}\|^2\}$, for $0 \le t \le T$. Applying (50) recursively, we have

$$\begin{aligned}\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(T)}) - F(\mathbf{w}^*)\} &\le (1+2l\lambda_2)^T \mathbb{E}\{F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\} \\ &+ \left(\lambda_1\beta\mathbb{E}\{\|\mathbf{n}\|\} + \lambda_0\mathbb{E}\{\|\mathbf{n}\|^2\}\right) \sum_{t=0}^{T-1} (1+2l\lambda_2)^t \\ &= (1+2l\lambda_2)^T \mathbb{E}\{F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\} \\ &+ \left(\lambda_1\beta\mathbb{E}\{\|\mathbf{n}\|\} + \lambda_0\mathbb{E}\{\|\mathbf{n}\|^2\}\right) \frac{(1+2l\lambda_2)^T - 1}{2l\lambda_2}. \quad (51)\end{aligned}$$

If $T \le L\sqrt{N}$ and then $\sigma_{\mathrm{D}} = 0$, this case is special. Hence, we will consider the condition that $T > L\sqrt{N}$. Based on (13), we have $\sigma_{\mathrm{A}} = \Delta s_{\mathrm{D}} T c/\epsilon$. Hence, we can obtain

$$\mathbb{E}\{\|\mathbf{n}\|\} = \frac{\Delta s_{\mathrm{D}} T c}{\epsilon}\sqrt{\frac{2N}{\pi}} \text{ and } \mathbb{E}\{\|\mathbf{n}\|^2\} = \frac{\Delta s_{\mathrm{D}}^2 T^2 c^2 N}{\epsilon^2}. \quad (52)$$

Substituting (52) into (51), setting $\Delta s_{\mathrm{D}} = 1/mN$ and $F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*) = \Theta$, we have

$$\begin{aligned}\mathbb{E}\{F(\widetilde{\mathbf{w}}^{(T)}) - F(\mathbf{w}^*)\} &\le (1+2l\lambda_2)^T \Theta \\ &+ \left(\frac{\lambda_1 T\beta c}{\epsilon}\sqrt{\frac{2}{N\pi}} + \frac{\lambda_0 T^2 c^2}{\epsilon^2 N}\right) \frac{(1+2l\lambda_2)^T - 1}{2l\lambda_2} \quad (53) \\ &= P^T\Theta + \left(\frac{\kappa_1 T}{\epsilon} + \frac{\kappa_0 T^2}{\epsilon^2}\right)\left(1 - P^T\right),\end{aligned}$$

where $P = 1 + 2l\lambda_2$, $\kappa_1 = \frac{\lambda_1\beta c}{m(P-1)}\sqrt{\frac{2}{N\pi}}$ and $\kappa_0 = \frac{\lambda_0 c^2}{m^2(P-1)N}$. This completes the proof. □

Appendix F
Proof of Lemma 4

We define the sampling parameter $q \triangleq K/N$ to represent the probability of being selected by the server for each client in an aggregation. Let $\mathcal{M}_{1:T}$ denote $(\mathcal{M}_1, \ldots, \mathcal{M}_T)$ and similarly let $o_{1:T}$ denote a sequence of outcomes $(o_1, \ldots, o_T)$. Considering a global (ϵ, δ)-DP in the downlinks channels, we use σ_{A} to represent the standard deviation of aggregated Gaussian noises. With neighboring datasets $\mathcal{D}_i$ and $\mathcal{D}_i'$, we are looking at

$$\begin{aligned}&\left|\ln\frac{\Pr[\mathcal{M}_{1:T}(\mathcal{D}'_{i,1:T}) = o_{1:T}]}{\Pr[\mathcal{M}_{1:T}(\mathcal{D}_{i,1:T}) = o_{1:T}]}\right| \\ &= \left|\sum_{i=1}^{T} \ln\frac{(1-q)e^{-\frac{\|n\|^2}{2\sigma_{\mathrm{A}}^2}} + qe^{-\frac{\|n+\Delta s_{\mathrm{D}}\|^2}{2\sigma_{\mathrm{A}}^2}}}{e^{-\frac{\|n\|^2}{2\sigma_{\mathrm{A}}^2}}}\right| \quad (54) \\ &= \left|\ln\prod_{i=1}^{T}\left(1 - q + qe^{-\frac{2n\Delta s_{\mathrm{D}} + \Delta s_{\mathrm{D}}^2}{2\sigma_{\mathrm{A}}^2}}\right)\right|.\end{aligned}$$

This quantity is bounded by ϵ, we require

$$\left|\ln\frac{\Pr[\mathcal{M}_{1:T}(\mathcal{D}'_{i,1:T}) = o_{1:T}]}{\Pr[\mathcal{M}_{1:T}(\mathcal{D}_{i,1:T}) = o_{1:T}]}\right| \le \epsilon. \quad (55)$$

Considering the independence of adding noises, we know

$$T\ln\left(1 - q + qe^{-\frac{2n\Delta s_{\mathrm{D}} + \|\Delta s_{\mathrm{D}}\|^2}{2\sigma_{\mathrm{A}}^2}}\right) \ge -\epsilon. \quad (56)$$

We can obtain the result

$$n \le -\frac{\sigma_{\mathrm{A}}^2}{\Delta s_{\mathrm{D}}}\ln\left(\frac{\exp(-\frac{\epsilon}{T})}{q} - \frac{1}{q} + 1\right) - \frac{\Delta s_{\mathrm{D}}}{2}. \quad (57)$$

We set

$$b = -\frac{T}{\epsilon}\ln\left(\frac{\exp(-\epsilon/T) - 1}{q} + 1\right). \quad (58)$$

Hence,

$$\ln\left(\frac{\exp(-\epsilon/T) - 1}{q} + 1\right) = -\frac{b\epsilon}{T}. \quad (59)$$

Note that ϵ and T should satisfy

$$\epsilon < -T\ln(1-q) \text{ or } T > \frac{-\epsilon}{\ln(1-q)}. \quad (60)$$

Then,

$$n \le \frac{\sigma_{\mathrm{A}}^2 b\epsilon}{T\Delta s_{\mathrm{D}}} - \frac{\Delta s_{\mathrm{D}}}{2}. \quad (61)$$

Using the tail bound $\Pr[n > \eta] \le \frac{\sigma_{\mathrm{A}}}{\sqrt{2\pi}}\frac{1}{\eta}e^{-\eta^2/2\sigma_{\mathrm{A}}^2}$, we can obtain

$$\ln\left(\frac{\eta}{\sigma_{\mathrm{A}}}\right) + \frac{\eta^2}{2\sigma_{\mathrm{A}}^2} > \ln\left(\sqrt{\frac{2}{\pi}}\frac{1}{\delta}\right). \quad (62)$$

Let us set $\sigma_{\mathrm{A}} = c\Delta s_{\mathrm{D}} T/b\epsilon$, if $b\epsilon/T \in (0,1)$, the inequation (62) can be solved as

$$c^2 \ge 2\ln\left(\frac{1.25}{\delta}\right). \quad (63)$$

Meanwhile, ϵ and T should satisfy

$$\epsilon < -T\ln\left(1 - q + \frac{q}{e}\right) \text{ or } T > \frac{-\epsilon}{\ln\left(1 - q + \frac{q}{e}\right)}. \quad (64)$$

If $b\epsilon/T > 1$, we can also obtain $\sigma_{\mathrm{A}} = c\Delta s_{\mathrm{D}} T/b\epsilon$ by adjusting the value of c. The standard deviation of requiring noises is given as

$$\sigma_{\mathrm{A}} \ge \frac{c\Delta s_{\mathrm{D}} T}{b\epsilon}. \quad (65)$$

Hence, if Gaussian noises are added at the client sides, we can obtain the additive noise scale in the server as

$$\begin{aligned}\sigma_{\mathrm{D}} &= \sqrt{\left(\frac{c\Delta s_{\mathrm{D}} T}{b\epsilon}\right)^2 - \frac{c^2 L^2 \Delta s_{\mathrm{U}}^2}{K\epsilon^2}} \\ &= \begin{cases}\frac{2cC\sqrt{\frac{T^2}{b^2} - L^2 K}}{mK\epsilon} & T > bL\sqrt{K}, \\ 0 & T \le bL\sqrt{K}.\end{cases} \quad (66)\end{aligned}$$

Furthermore, considering (60), we can obtain

$$\sigma_{\mathrm{D}} = \begin{cases}\frac{2cC\sqrt{\frac{T^2}{b^2} - L^2 K}}{mK\epsilon} & T > \frac{\epsilon}{\gamma}, \\ 0 & T \le \frac{\epsilon}{\gamma},\end{cases} \quad (67)$$

where

$$\gamma = -\ln\left(1 - q + qe^{\frac{-\epsilon}{L\sqrt{K}}}\right). \quad (68)$$

This completes the proof. □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.2988575, IEEE Transactions on Information Forensics and Security

APPENDIX G
PROOF OF THEOREM 3

Here we define

$$\mathbf{v}^{(t)} = \sum_{i=1}^{K} p_i \mathbf{w}_i^{(t)}, \tag{69}$$

$$\widetilde{\mathbf{v}}^{(t)} = \sum_{i=1}^{K} p_i \left(\mathbf{w}_i^{(t)} + \mathbf{n}_i^{(t)} \right) + \mathbf{n}_{\mathrm{D}}^{(t)} \tag{70}$$

and

$$\mathbf{n}^{(t+1)} = \sum_{i=1}^{K} p_i \mathbf{n}_i^{(t+1)} + \mathbf{n}_{\mathrm{D}}^{(t)}. \tag{71}$$

which considers the aggregated parameters under K-random scheduling. Because $F_i(\cdot)$ and $F(\cdot)$ are β-Lipschitz, we obtain that

$$\mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t+1)})\} - F(\mathbf{w}^{(t+1)}) \leq \beta \|\widetilde{\mathbf{v}}^{(t+1)} - \mathbf{w}^{(t+1)}\|. \tag{72}$$

Because β is the Lipchitz continuity constant of function F, we have

$$\begin{aligned} \beta \leq \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\| + \rho \Big(& \|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t)}\| \\ & + \|\mathbf{v}^{(t+1)} - \widetilde{\mathbf{v}}^{(t)}\| \Big). \end{aligned} \tag{73}$$

From (42), we know

$$\|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t)}\| \leq \frac{B(1+\theta)}{\overline{\mu}} \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\|. \tag{74}$$

Then, we have

$$\begin{aligned} \mathbb{E}\{\|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t+1)}\|^2\} &= \|\mathbf{w}^{(t+1)}\|^2 \\ &- 2[\mathbf{w}^{(t+1)}]^{\top} \mathbb{E}\{\widetilde{\mathbf{v}}^{(t+1)}\} + \mathbb{E}\{\|\widetilde{\mathbf{v}}^{(t+1)}\|^2\}. \end{aligned} \tag{75}$$

Furthermore, we can obtain

$$\begin{aligned} \mathbb{E}\{\widetilde{\mathbf{v}}^{(t+1)}\} &= \frac{1}{\binom{N}{K}} \frac{\binom{N}{K}}{N} K \sum_{i=1}^{N} p_i \mathbf{w}_i^{(t+1)} + \mathbf{n}^{(t+1)} \\ &= \mathbb{E}\{\mathbf{w}_i^{(t+1)}\} + \mathbf{n}^{(t+1)} = \mathbf{w}^{(t+1)} + \mathbf{n}^{(t+1)} \end{aligned} \tag{76}$$

and

$$\begin{aligned} \mathbb{E}\{\|\widetilde{\mathbf{v}}^{(t+1)}\|^2\} &= \mathbb{E}\left\{ \left\| \sum_{i=1}^{K} p_i \mathbf{w}_i^{(t+1)} \right\|^2 \right\} \\ &+ \mathbb{E}\left\{ \left\| \sum_{i=1}^{K} p_i \mathbf{n}_i^{(t+1)} \right\|^2 \right\} + 2\mathbb{E}\left\{ \left[\sum_{i=1}^{K} p_i \mathbf{w}_i^{(t+1)} \right]^{\top} \mathbf{n}^{(t+1)} \right\}. \end{aligned} \tag{77}$$

Note that we set $p_i = D_i / \sum_{i=1}^{K} D_i = 1/K$ in K-client random scheduling in order to a small sensitivity Δs_{D}. We have

$$\begin{aligned} & \mathbb{E}\left\{ \left\| \sum_{i=1}^{K} p_i \mathbf{w}_i^{(t+1)} \right\|^2 \right\} \\ & \qquad \leq \frac{1}{K^2} \sum_{i=1}^{K} \|\mathbf{w}_i^{(t+1)}\|^2 + \frac{K-1}{K} \|\mathbf{w}^{(t+1)}\|^2 \end{aligned} \tag{78}$$

and

$$\begin{aligned} \mathbb{E}\{\|\widetilde{\mathbf{v}}^{(t+1)}\|^2\} \leq \frac{1}{K^2} \sum_{i=1}^{K} \|\mathbf{w}_i^{(t+1)}\|^2 + \frac{K-1}{K} \|\mathbf{w}^{(t+1)}\|^2 \\ + \|\mathbf{n}^{(t+1)}\|^2 + 2[\mathbf{w}^{(t+1)}]^{\top} \mathbf{n}^{(t+1)}. \end{aligned} \tag{79}$$

Combining (75) and (79), we can obtain

$$\begin{aligned} & \mathbb{E}\{\|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t+1)}\|^2\} \\ & \qquad \leq \frac{1}{K^2} \sum_{i=1}^{K} \|\mathbf{w}_i^{(t+1)} - \widetilde{\mathbf{v}}^{(t)}\|^2 + \|\mathbf{n}^{(t+1)}\|^2. \end{aligned} \tag{80}$$

Using (41), we know

$$\begin{aligned} \mathbb{E}\{\|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t+1)}\|^2\} \leq \|\mathbf{n}^{(t+1)}\|^2 + \\ \frac{B^2(1+\theta)^2}{K\overline{\mu}^2} \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\|^2. \end{aligned} \tag{81}$$

Moreover,

$$\begin{aligned} \mathbb{E}\{\|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t+1)}\|\} \leq \|\mathbf{n}^{(t+1)}\| \\ + \frac{B(1+\theta)}{\overline{\mu}\sqrt{K}} \|\nabla F(\widetilde{\mathbf{v}}^{t})\|. \end{aligned} \tag{82}$$

Substituting (45), (73) and (82) into (72), setting $\theta = 0$ and $\overline{\mu} = \mu$, we can obtain

$$\begin{aligned} & \mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t+1)})\} - F(\widetilde{\mathbf{v}}^{(t)}) \leq F(\mathbf{w}^{(t+1)}) - F(\widetilde{\mathbf{v}}^{(t)}) \\ & \Big(\|\nabla F(\widetilde{\mathbf{v}}^{(t)})\| + 2\rho \|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t)}\| \Big) \mathbb{E}\|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t+1)}\| \\ & \quad + \rho \mathbb{E}\{\|\mathbf{w}^{(t+1)} - \widetilde{\mathbf{v}}^{(t+1)}\|^2\} = \alpha_2 \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\|^2 \\ & \quad + \alpha_1 \|\mathbf{n}^{(t+1)}\| \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\| + \alpha_0 \|\mathbf{n}^{(t+1)}\|^2, \end{aligned} \tag{83}$$

where

$$\alpha_2 = \frac{1}{\mu^2} \left(\frac{\rho B^2}{2} + \rho B + \frac{\rho B^2}{K} + \frac{2\rho B^2}{\sqrt{K}} + \frac{\mu B}{\sqrt{K}} - \mu \right), \tag{84}$$

$$\alpha_1 = 1 + \frac{2\rho B}{\mu} + \frac{2\rho B\sqrt{K}}{\mu N} \text{ and } \alpha_0 = \frac{2\rho K}{N} + \rho. \tag{85}$$

In this case, we take expectation $\mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t+1)}) - F(\widetilde{\mathbf{v}}^{(t)})\}$ as follows,

$$\begin{aligned} & \mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t+1)}) - F(\widetilde{\mathbf{v}}^{(t)})\} \leq \alpha_2 \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\|^2 \\ & \quad + \alpha_1 \mathbb{E}\{\|\mathbf{n}^{(t+1)}\|\} \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\| + \alpha_0 \mathbb{E}\{\|\mathbf{n}^{(t+1)}\|^2\}. \end{aligned} \tag{86}$$

For $\Theta > 0$ and $f(\mathbf{v}^{(0)}) - f(\mathbf{w}^{*}) = \Theta$, we can obtain

$$\begin{aligned} & \mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t+1)}) - F(\mathbf{w}^{*})\} \\ & \qquad \leq \mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t)}) - F(\mathbf{w}^{*})\} + \alpha_2 \|\nabla F(\widetilde{\mathbf{v}}^{(t)})\|^2 \\ & \qquad + \alpha_1 \beta \mathbb{E}\{\|\mathbf{n}^{(t+1)}\|\} + \alpha_0 \mathbb{E}\{\|\mathbf{n}^{(t+1)}\|^2\}. \end{aligned} \tag{87}$$

If we select the penalty parameter μ to make $\alpha_2 < 0$ and using (48), we know

$$\begin{aligned} \mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t+1)}) - F(\mathbf{w}^{*})\} \leq (1 + 2l\alpha_2) \mathbb{E}\{F(\widetilde{\mathbf{v}}^{(t)}) - F(\mathbf{w}^{*})\} \\ + \alpha_1 \beta \mathbb{E}\{\|\mathbf{n}^{(t+1)}\|\} + \alpha_0 \mathbb{E}\{\|\mathbf{n}^{(t+1)}\|^2\}. \end{aligned} \tag{88}$$

Considering independence of additive noises and applying (88) recursively, we have

$$\begin{aligned}\mathbb{E}\{F(\widetilde{\mathbf{v}}^{(T)})-F(\mathbf{w}^*)\} &\leq (1+2l\alpha_2)^T\mathbb{E}\{F(\mathbf{v}^{(0)})-F(\mathbf{w}^*)\}\\ &+\frac{1-(1+2l\alpha_2)^T}{2l\alpha_2}\left(\alpha_1\beta\mathbb{E}\{\|\mathbf{n}\|\}+\alpha_0\mathbb{E}\{\|\mathbf{n}\|^2\}\right)\\ &=Q^T\Theta+\frac{1-Q^T}{1-Q}\left(\alpha_1\beta\mathbb{E}\{\|\mathbf{n}\|\}+\alpha_0\mathbb{E}\{\|\mathbf{n}\|^2\}\right), \quad (89)\end{aligned}$$

where $Q=1+2l\alpha_2$. Substituting (65) into (89), we can obtain

$$\mathbb{E}\{\|\mathbf{n}\|\}=\frac{\Delta s_{\mathrm{D}}Tc}{b\epsilon}\sqrt{\frac{2N}{\pi}},\mathbb{E}\{\|\mathbf{n}\|^2\}=\frac{\Delta s_{\mathrm{D}}^2T^2c^2N}{b^2\epsilon^2} \quad (90)$$

and

$$\begin{aligned}\mathbb{E}\{F(\widetilde{\mathbf{v}}^{T})-F(\mathbf{w}^*)\} &\leq Q^T\Theta\\ &+\frac{1-Q^T}{1-Q}\left(\frac{cC\alpha_1\beta}{-mK\ln\left(1-\frac{N}{K}+\frac{N}{K}e^{-\frac{\epsilon}{T}}\right)}\sqrt{\frac{2}{\pi}}\right. \quad (91)\\ &\left.+\frac{c^2C^2\alpha_0}{m^2K^2\ln^2\left(1-\frac{N}{K}+\frac{N}{K}e^{-\frac{\epsilon}{T}}\right)}\right).\end{aligned}$$

This completes the proof. □

REFERENCES

[1] J. Li, S. Chu, F. Shu, J. Wu, and D. N. K. Jayakody, "Contract-Based Small-Cell Caching for Data Disseminations in Ultra-Dense Cellular Networks," *IEEE Trans. Mobile Comput.*, vol. 18, no. 5, pp. 1042–1053, May 2019.

[2] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, and B. Vucetic, "High-Reliability and Low-Latency Wireless Communication for Internet of Things: Challenges, Fundamentals, and Enabling Technologies," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7946–7970, Oct. 2019.

[3] H. Lee, S. H. Lee, and T. Q. S. Quek, "Deep Learning for Distributed Optimization: Applications to Wireless Resource Management," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2251–2266, Oct. 2019.

[4] W. Sun, J. Liu, and Y. Yue, "AI-Enhanced Offloading in Edge Computing: When Machine Learning Meets Industrial IoT," *IEEE Netw.*, vol. 33, no. 5, pp. 68–74, Sep. 2019.

[5] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, Jun. 2018.

[6] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated Learning of Deep Networks using Model Averaging," *arXiv*, 2016. [Online]. Available: http://arxiv.org/abs/1602.05629

[7] J. Konecný *et al.*, "Federated Learning: Strategies for Improving Communication Efficiency," *arXiv*, 2016. [Online]. Available: http://arxiv.org/abs/1610.05492

[8] U. Mohammad and S. Sorour, "Adaptive Task Allocation for Asynchronous Federated Mobile Edge Learning," *arXiv*, 2019. [Online]. Available: http://arxiv.org/abs/1905.01656

[9] X. Wang *et al.*, "In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[10] Y. Qiang, L. Yang, C. Tianjian, and T. Yongxin, "Federated Machine Learning: Concept and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 12:1–12:19, Jan. 2019.

[11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1908.07873

[12] N. H. Tran, W. Bao, A. Zomaya, N. Minh N.H., and C. S. Hong, "Federated Learning over Wireless Networks: Optimization Model Design and Analysis," in *Proc. IEEE INFOCOM*, Apr. 2019, pp. 1387–1395.

[13] H. H. Yang, Z. Liu, T. Q. S. Quek, and H. V. Poor, "Scheduling Policies for Federated Learning in Wireless Networks," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 317–333, 2020.

[14] M. Hao, H. Li, G. Xu, S. Liu, and H. Yang, "Towards Efficient and Privacy-Preserving Federated Deep Learning," in *Proc. IEEE ICC*, Paris, France, May 2019, pp. 1–6.

[15] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019. [Online]. Available: https://arxiv.org/abs/1910.14648

[16] A. Alekh and D. J. C, "Distributed Delayed Stochastic Optimization," in *Proc. IEEE CDC*, Maui, HI, USA, Dec. 2012.

[17] L. Xiangru, H. Yijun, L. Yuncheng, and L. Ji, "Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization," in *Proc. ACM NIPS*, Montreal, Canada, Dec. 2015, pp. 2737–2745.

[18] X. Lian *et al.*, "Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent," in *Proc. ACM NIPS*, Long Beach, California, USA, Dec. 2017, pp. 5336–5346.

[19] T. Li *et al.*, "On the Convergence of Federated Optimization in Heterogeneous Networks," *arXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1812.06127

[20] S. Wang *et al.*, "Adaptive Federated Learning in Resource Constrained Edge Computing Systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[21] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in *Proc. ACM CCS*, Denver, Colorado, USA, Oct. 2015, pp. 1310–1321.

[22] Z. Wang *et al.*, "Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning," in *Proc. IEEE INFOCOM*, Paris, France, Apr. 2019, pp. 2512–2520.

[23] C. Ma, J. Li, M. Ding, H. Hao Yang, F. Shu, T. Q. S. Quek, and H. V. Poor, "On Safeguarding Privacy and Security in the Framework of Federated Learning," *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1909.06512

[24] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[25] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SuLQ Framework," in *Proc. ACM PODS*, Baltimore, Maryland, Jun. 2005, pp. 128–138.

[26] Úlfar Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," in *Proc. ACM CCS*, Scottsdale, Arizona, USA, Nov. 2014, pp. 1054–1067.

[27] N. Wang *et al.*, "Collecting and Analyzing Multidimensional Data with Local Differential Privacy," in *Proc. IEEE ICDE*, Macao, China, Apr. 2019, pp. 638–649.

[28] S. Wang, L. Huang, Y. Nie, X. Zhang, P. Wang, H. Xu, and W. Yang, "Local Differential Private Data Aggregation for Discrete Distribution Estimation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 9, pp. 2046–2059, Sep. 2019.

[29] A. Martin *et al.*, "Deep Learning with Differential Privacy," in *Proc. ACM CCS*, Vienna, Austria, Oct. 2016, pp. 308–318.

[30] N. Wu, F. Farokhi, D. Smith, and M. A. Kâafar, "The Value of Collaboration in Convex Machine Learning with Differential Privacy," *arXiv*, 2019. [Online]. Available: http://arxiv.org/abs/1906.09679

[31] J. Li, M. Khodak, S. Caldas, and A. Talwalkar, "Differentially Private Meta-Learning," *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1909.05830

[32] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning Differentially Private Language Models Without Losing Accuracy," *arXiv*, Feb. 2018. [Online]. Available: http://arxiv.org/abs/1710.06963

[33] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A Generic Framework for Privacy Preserving Deep Learning," *arXiv*, Nov. 2018. [Online]. Available: http://arxiv.org/abs/1811.04017

[34] R. C. Geyer, T. Klein, and M. Nabi, "Differentially Private Federated Learning: A Client Level Perspective," *arXiv*, 2017. [Online]. Available: http://arxiv.org/abs/1712.07557

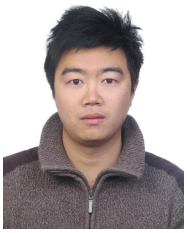
[35] S. Truex *et al.*, "A Hybrid Approach to Privacy-Preserving Federated Learning," *arXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1812.03224

[36] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. ACM CCS*, New York, NY, USA, 2015, pp. 1322–1333.

[37] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[38] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed. Springer Publishing Company, Incorporated, 2014.

Kang Wei received the B.Sc. degree in information engineering from Xidian University, Xian, China, in 2014. From September 2017 to August 2018, he was a M.Sc. candidate, and from September 2018 to now he is a Ph.D. candidate at the school of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His current research interests include data privacy and security, differential privacy, AI and machine learning, information theory, and channel coding theory in NAND flash memory.



Jun Li (M'09-SM'16) received Ph. D degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. From January 2009 to June 2009, he worked in the Department of Research and Innovation, Alcatel Lucent Shanghai Bell as a Research Scientist. From June 2009 to April 2012, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia. From April 2012 to June 2015, he is a Research Fellow at the School of Electrical Engineering, the University of Sydney, Australia. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He was a visiting professor at Princeton University from 2018 to 2019. His research interests include network information theory, game theory, distributed intelligence, multiple agent reinforcement learning, and their applications in ultra-dense wireless networks, mobile edge computing, network privacy and security, and industrial Internet of things. He has co-authored more than 200 papers in IEEE journals and conferences, and holds 1 US patents and more than 10 Chinese patents in these areas. He was serving as an editor of IEEE Communication Letters and TPC member for several flagship IEEE conferences. He received Exemplary Reviewer of IEEE Transactions on Communications in 2018, and best paper award from IEEE International Conference on 5G for Future Wireless Networks in 2017.

Ming Ding (M'12-SM'17) received the B.S. and M.S. degrees (with first class Hons.) in electronics engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, and the Doctor of Philosophy (Ph.D.) degree in signal and information processing from SJTU, in 2004, 2007, and 2011, respectively. From April 2007 to September 2014, he worked at Sharp Laboratories of China in Shanghai, China as a Researcher/Senior Researcher/Principal Researcher. He also served as the Algorithm Design Director and Programming Director for a system-level simulator of future telecommunication networks in Sharp Laboratories of China for more than 7 years. Currently, he is a senior research scientist at Data61, CSIRO, in Sydney, NSW, Australia. His research interests include information technology, data privacy and security, machine Learning and AI, etc. He has authored over 100 papers in IEEE journals and conferences, all in recognized venues, and around 20 3GPP standardization contributions, as well as a Springer book "Multi-point Cooperative Communication Systems: Theory and Applications". Also, he holds 21 US patents and co-invented another 100+ patents on 4G/5G technologies in CN, JP, KR, EU, etc. Currently, he is an editor of IEEE Transactions on Wireless Communications and IEEE Wireless Communications Letters. Besides, he is or has been Guest Editor/Co-Chair/Co-Tutor/TPC member of several IEEE top-tier journals/conferences, e.g., the IEEE Journal on Selected Areas in Communications, the IEEE Communications Magazine, and the IEEE Globecom Workshops, etc. He was the lead speaker of the industrial presentation on unmanned aerial vehicles in IEEE Globecom 2017, which was awarded as the Most Attended Industry Program in the conference. Also, he was awarded in 2017 as the Exemplary Reviewer for IEEE Transactions on Wireless Communications.

Chuan Ma received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013 and Ph.D. degree from the University of Sydney, Australia, in 2018. He is now working as a lecturer at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He has published more than 10 journal and conference papers, including a best paper in WCNC 2018. His research interests include stochastic geometry, wireless caching networks and machine learning, and now focuses on the big data analysis and privacy preservation.

Howard H. Yang (S'13-M'17) received the B.Sc. degree in Communication Engineering from Harbin Institute of Technology (HIT), China, in 2012, and the M.Sc. degree in Electronic Engineering from Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2013. He earned the Ph.D. degree in Electronic Engineering from Singapore University of Technology and Design (SUTD), Singapore, in 2017. From August 2015 to March 2016, he was a visiting student in the WNCG under supervisor of Prof. Jeffrey G. Andrews at the University of Texas at Austin. Dr. Yang is now a Postdoctoral Research Fellow with Singapore University of Technology and Design in the Wireless Networks and Decision Systems (WNDS) group led by Prof. Tony Q. S. Quek.

He has held a visiting research appointment at Princeton University from September 2018 to April 2019. His research interests cover various aspects of wireless communications, networking, and signal processing, currently focusing on the modeling of modern wireless networks, high dimensional statistics, graph signal processing, and machine learning. He received the IEEE WCSP 10-Year Anniversary Excellent Paper Award in 2019 and the IEEE WCSP Best Paper Award in 2014.

Farhad Farokhi is a Lecturer (=Assistant Professor) at the Department of Electrical and Electronic Engineering at the University of Melbourne. Prior to that, he was a Research Scientist at the Information Security and Privacy Group at CSIRO's Data61, a Research Fellow at the University of Melbourne, and Postdoctoral Fellow at KTH Royal Institute of Technology. In 2014, he received his PhD degree from KTH Royal Institute of Technology. During his PhD studies, he was a visiting researcher at the University of California at Berkeley and the University of Illinois at Urbana-Champaign. Farhad has been the recipient of the VESKI Victoria Fellowship from the Victorian State Government as well as the McKenzie Fellowship and the 2015 Early Career Researcher Award from the University of Melbourne. He was a finalist in the 2014 European Embedded Control Institute (EECI) PhD Award. He has been part of numerous projects on data privacy and cyber-security funded by the Defence Science and Technology Group (DSTG), the Department of the Prime Minister and Cabinet (PMC), the Department of Environment and Energy (DEE), and CSIRO in Australia.

Shi Jin (S'06-M'07-SM'17) received the B.S. degree in communications engineering from Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from the Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with the faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory. He serves as an Associate Editor for the IEEE Transactions on Wireless Communications, and IEEE Communications Letters, and IET Communications. Dr. Jin and his co-authors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and a 2010 Young Author Best Paper Award by the IEEE Signal Processing Society.

Tony Q.S. Quek (S'98-M'08-SM'12-F'18) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. Currently, he is the Cheng Tsang Man Chair Professor with Singapore University of Technology and Design (SUTD). He also serves as the Head of ISTD Pillar, Sector Lead of the SUTD AI Program, and the Deputy Director of the SUTD-ZJU IDEA. His current research topics include wireless communications and networking, network intelligence, internet-of-things, URLLC, and big data processing.

Dr. Quek has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is currently serving as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the Chair of IEEE VTS Technical Committee on Deep Learning for Wireless Communications as well as an elected member of the IEEE Signal Processing Society SPCOM Technical Committee. He was an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, and an Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards – Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, and the 2016-2019 Clarivate Analytics Highly Cited Researcher. He is a Distinguished Lecturer of the IEEE Communications Society and a Fellow of IEEE.

H. Vincent Poor (S'72-M'77-SM'82-F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering. From 2006 until 2016, he served as Dean of Princetons School of Engineering and Applied Science. He has also held visiting appointments at several other institutions, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, signal processing and machine learning, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the forthcoming book Advanced Data Analytics for Power Systems (Cambridge University Press, 2020).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences, the Royal Society and other national and international academies. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, a D.Sc. honoris causa from Syracuse University, awarded in 2017, and a D.Eng. honoris causa from the University of Waterloo, awarded in 2019.