Personalized Federated Learning With Differential Privacy

Rui Hu, *Student Member, IEEE*, Yuanxiong Guo, *Senior Member, IEEE*, Hongning Li, *Member, IEEE*, Qingqi Pei, *Senior Member, IEEE*, and Yanmin Gong, *Member, IEEE*

***Abstract*—To provide intelligent and personalized services on smart devices, machine learning techniques have been widely used to learn from data, identify patterns, and make automated decisions. Machine learning processes typically require a large amount of representative data that are often collected through crowdsourcing from end users. However, user data could be sensitive in nature, and training machine learning models on these data may expose sensitive information of users, violating their privacy. Moreover, to meet the increasing demand of personalized services, these learned models should capture their individual characteristics. This article proposes a privacy-preserving approach for learning effective personalized models on distributed user data while guaranteeing the differential privacy of user data. Practical issues in a distributed learning system such as user heterogeneity are considered in the proposed approach. In addition, the convergence property and privacy guarantee of the proposed approach are rigorously analyzed. The experimental results on realistic mobile sensing data demonstrate that the proposed approach is robust to user heterogeneity and offers a good tradeoff between accuracy and privacy.**

***Index Terms*—Data privacy, distributed algorithm, machine learning.**

I. Introduction

SMART devices equipped with sensing, communications, computing, and/or control capabilities, such as smartphones, wearable devices, and in-vehicle sensing devices, are becoming extremely popular nowadays. These devices generate, collect, store, and analyze an unprecedented amount of data as they interact with the physical world, which can provide intelligent and personalized services to people. For instance, smartwatches can record their users' physical activities and mental conditions for health monitoring at any time, and smart insoles can track the body temperature, motion, and heart rate of their users to help them stay injury free and run better.

For these smart devices to provide intelligent services, machine learning techniques need to be applied to learn powerful predictive models on the collected data. A common practice to learn predictive models from these crowdsourced data is to first collect data from all devices in a cloud server and then train a global model. However, it may be risky to store the privacy-sensitive data in a cloud server which may not be fully trustworthy. Moreover, as the data volume increases, the cost and latency of uploading all the raw data to a distant cloud server increase as well. On the other hand, a device may choose to learn a local model on its own data without sharing data with other devices. Such local models often perform poorly due to the limited training data size. Hence, how to benefit from data sharing without violating user privacy in learning predictive models from distributed data is a challenge. *Federated learning* [1] has been proposed recently as a promising approach to solve the challenge. In federated learning, all devices update the global model downloaded from the cloud server with their own data and only send the updates back to the server for aggregation. By sharing only the learned updates rather than the raw data, federated learning both achieves high communication efficiency and reduces privacy risks while obtaining effective predictive models.

Although promising, there remain several issues in applying federated learning to the real world. First, the model obtained through federated learning is a shared model that extracts the common knowledge of all participants without capturing personal inclinations [2]. For instance, when learning the sentiment of users on their personal messages, a single global model cannot capture such differences, since the same word from different users may convey different sentiments due to various personal opinions and language using habits. However, since people with close relationships are likely to have similar habits, it will be beneficial to allow the learning tasks of all users to learn from each other based on their relationships. Also known as *multitask learning*, this kind of method allows personalized models to be learned, which could both benefit from the collective data and keep personal characteristics. Second, in a federated learning system with lots of participants, the device heterogeneity has a large impact on learning efficiency. The network condition, data size, and computation capability of different devices are different and even time varying, which may result in the delay, dropout, or poor quality

Manuscript received December 15, 2019; revised March 20, 2020; accepted April 13, 2020. Date of publication April 30, 2020; date of current version October 9, 2020. The work of Rui Hu and Yanmin Gong was supported by the U.S. National Science Foundation under Grant CNS-1850623 and Grant CNS-2029685. The work of Yuanxiong Guo was supported by the U.S. National Science Foundation under Grant CNS-2029685. The work of Hongning Li is supported by the Key Program of NSFC-Tongyong Union Foundation under Grant U1636209. *(Corresponding author: Yanmin Gong.)*

Rui Hu and Yanmin Gong are with the Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: rui.hu@utsa.edu; yanmin.gong@utsa.edu).

Yuanxiong Guo is with the Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: yuanxiong.guo@utsa.edu).

Hongning Li and Qingqi Pei are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: hnli@xidian.edu.cn; qqpei@mail.xidian.edu.cn).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JIOT.2020.2991416

2327-4662 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

of the updates. Third, federated learning does not provide a rigorous privacy guarantee for participants. The aggregation server in federated learning is assumed to be fully trusted to coordinate the training. However, the server can easily violate the privacy of participants by observing their updates as shown in recent attacks [3], [4].

To address the aforementioned issues, we propose a novel federated learning scheme that provides an effective personalized model for each participant under the device heterogeneity while guaranteeing differential privacy of their data. In our proposed scheme, the personalized model of a participant is learned based on not only its own local data but also the shared updates computed from other participants' data. We provide differential privacy guarantee for shared updates by adding certain amounts of noise before releasing them. At the heart of our scheme is a new iterative algorithm that solves the multitask learning optimization problem in a distributed and privacy-preserving way. The iterative algorithm can learn optimal personalized models and the relationship between them simultaneously. Since the algorithm is an iterative process and would consume the privacy budget at each iteration, we further use moments accountant to characterize the end-to-end privacy loss after multiple iterations.

In summary, the main contributions of this article are as follows.

1) To the best of our knowledge, this is the first work that rigorously analyzes the personalized federated learning with a differential privacy in a heterogeneous IoT setting.
2) We propose a novel differentially private federated learning scheme for collaboratively training multiple personalized machine learning models from the data stored across smart devices in IoT.
3) We perform rigorous privacy analysis considering the heterogeneity of the IoT devices and convergence analysis for our proposed approach.
4) We conduct extensive evaluations based on real-world data sets, verify the effectiveness of the proposed approach, and observe the tradeoff among model accuracy and privacy empirically.

The remainder of this article is organized as follows. We first describe the preliminaries in Section II and problem setting in Section III. Then, we develop a distributed learning scheme to achieve personalized federated learning with differential privacy guarantee in Section IV. Next, we analyze the convergence rate of the proposed solution in Section V. The privacy analysis of the proposed solution is provided in Section VI, and numerical results are provided in Section VII. Finally, related literature is reviewed in Section VIII, and conclusions are made in Section IX.

II. Preliminaries

In this section, we first briefly describe the basics of moments accountant and their properties. The basic idea of moments accountant is to bound the privacy loss by bounding the log moment of the privacy loss. For neighboring databases $\mathcal{A}$ and $\mathcal{A}'$, the randomized mechanism $\mathcal{M}$, and auxiliary input au, the privacy loss incurred by observing o is defined as

$$z(o; \mathcal{M}, au, \mathcal{A}, \mathcal{A}') := \log\left(\frac{P_r[\mathcal{M}(au, \mathcal{A}) = o]}{P_r[\mathcal{M}(au, \mathcal{A}') = o]}\right).$$

Here, we take the privacy loss $z(o; \mathcal{M}, au, \mathcal{A}, \mathcal{A}')$ as a random variable because the mechanism $\mathcal{M}$ is randomized.

Accordingly, given the randomized mechanism $\mathcal{M}$ and any positive integer γ, the γth moment of the privacy loss $\mu_{\mathcal{M}}(\gamma; au, \mathcal{A}, \mathcal{A}')$ is defined as the log of the moment generating function evaluated at γ

$$\begin{aligned}&\mu_{\mathcal{M}}\big(\gamma; au, \mathcal{A}, \mathcal{A}'\big)\\ &\quad := \log \mathbb{E}_{o\sim\mathcal{M}(\mathcal{A})}\big[\exp\big(\gamma z\big(o; \mathcal{M}, au, \mathcal{A}, \mathcal{A}'\big)\big)\big].\end{aligned}$$

Then, the moments accountant is defined as the upper bound of $\mu_{\mathcal{M}}(\gamma; au, \mathcal{A}, \mathcal{A}')$ over all possible auxiliary information au and neighboring databases $\mathcal{A}$ and $\mathcal{A}'$ as

$$\mu_{\mathcal{M}}(\gamma) := \max_{au,\mathcal{A},\mathcal{A}'} \mu_{\mathcal{M}}\big(\gamma; au, \mathcal{A}, \mathcal{A}'\big).$$

Two important properties of moments accountant are provided as follows, which will be used to compute the moments accountant of an adaptive mechanism and convert the moments accountant into a differential privacy guarantee.

Theorem 1 (Composability) [5]: Suppose that an adaptive mechanism $\mathcal{M}_{1:k}$ consists of a sequence of randomized mechanisms $(\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k)$, where $\mathcal{M}_i$ takes the data set $\mathcal{A}$ and the output of $\mathcal{M}_{i-1}$ as its inputs. If the moments accountant of $\mathcal{M}_i$ is $\mu_{\mathcal{M}_i}(\gamma)$ $\forall i \in [k]$, then for any positive integer γ and any output $(o_1, \ldots, o_{k-1})$

$$\mu_{\mathcal{M}_{1:k}}(\gamma) = \sum_{i=1}^{k} \mu_{\mathcal{M}_i}(\gamma)$$

where o_i for $i < k$ is the output of mechanism $\mathcal{M}_i$ and $\mu_{\mathcal{M}_{1:k}}(\gamma)$ is conditioned on these $k-1$ outputs.

Theorem 2 (Tail Bound) [6]: For any $\epsilon > 0$, mechanism $\mathcal{M}$ is (ϵ, δ)-differentially private for

$$\delta = \min_{\gamma} \exp(\mu_{\mathcal{M}}(\gamma) - \gamma\epsilon).$$

III. Problem Setting

We consider a federated learning system as shown in Fig. 1.

In the system, a group of smart devices (also known as, users or participants) senses the physical world continuously and stores the collected data in their local databases. Each device has some embedded computing capabilities capable of training a local model. A cloud server will coordinate the collaboration among devices to improve their models from others' data.

Let m denote the total number of devices in the system, each needing to learn a personalized model. Each device t has a local training data set $\mathcal{A}_t = (\mathbf{X}_t, \mathbf{y}_t)$, where the ith column of the matrix $\mathbf{X}_t$ denotes a feature vector $\mathbf{x}_t^i \in \mathbb{R}^d$, and the ith element of the vector $\mathbf{y}_t$ denotes the corresponding label y_t^i that takes continuous value for regression problems and categorical value for classification problems. Let n_t be the total number of training samples in device t's database, and therefore, $\mathbf{X}_t \in \mathbb{R}^{d\times n_t}$. We assume that the feature vector $\|\mathbf{x}_t^i\|_2 \leq 1$ which can be enforced through normalization.

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

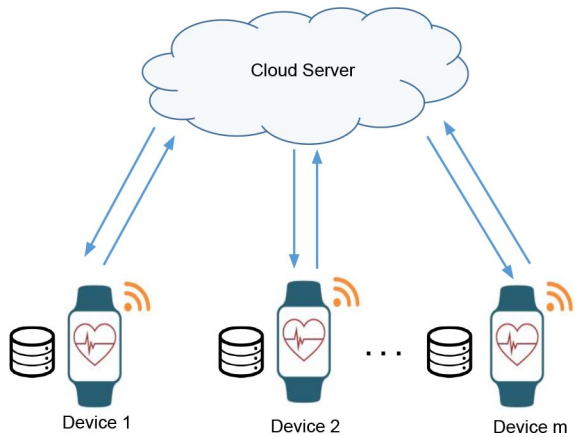


Fig. 1. Overall architecture of federated learning systems.

Denote by $\mathbf{w}_t \in \mathbb{R}^d$ the model parameters of device t and by $\mathbf{W} := [\mathbf{w}_1, \ldots, \mathbf{w}_m] \in \mathbb{R}^{d\times m}$ the collective model parameters of all devices. We use $n := \sum_{t=1}^{T} n_t$ to denote the total number of all data points and represent the overall feature data matrix as $\mathbf{X} := \mathrm{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_m) \in \mathbb{R}^{md\times n}$. Then the personalized federated learning can be formulated as the following multitask learning problem [7]:

$$\begin{aligned} \min_{\mathbf{W},\Omega} \ & \mathcal{P}(\mathbf{W}, \Omega) := \sum_{t=1}^{m}\sum_{i=1}^{n_t} \ell_t\big(\mathbf{w}_t^T\mathbf{x}_t^i, y_t^i\big) + \lambda\,\mathrm{tr}\Big(\mathbf{W}\Omega^{-1}\mathbf{W}^T\Big) \\ \text{s.t.} \ & \Omega \succeq 0, \quad \mathrm{tr}(\Omega) = 1 \end{aligned} \tag{1}$$

where $\Omega \in \mathbb{R}^{m\times m}$ is the task covariance matrix that models the relationship between different devices, $\lambda > 0$ is the regularization parameter, and $\ell_t(\cdot)$ is the convex loss function corresponding to device t's learning task. In the above optimization problem, the first term of the objective function measures the empirical loss of all training samples, and the second term measures the learning task relationship between devices. Note that $\mathcal{P}(\mathbf{W}, \Omega)$ is jointly convex with respect to $\mathbf{W}$ and Ω under our assumptions as proved in [8].

IV. Privacy-Preserving Distributed Framework

In this section, we propose a distributed framework to solve problem (1) with a rigorous privacy guarantee. We first describe the threat model and design goals and then propose a privacy-preserving algorithm to solve the problem.

A. Threat Model and Design Goals

We assume the information sent through the network is well protected during the transmission and the adversary here can be the "honest-but-curious" central server or users in the system. By observing the received updates, it is possible for the server or malicious users to recover the training data using the model inversion attack [3] or infer whether a sample is in the training data set with the membership inference attack [4]. The goal of our design is to ensure that the server or malicious users cannot learn much additional information of user samples from the received messages under any auxiliary information and attack. We design our privacy-preserving algorithm in the framework of differential privacy (DP) [9]. A differentially private algorithm provides a strong guarantee that the presence of an individual record in the data set will not significantly change the output of the algorithm. Specifically, we use the notion of (ϵ, δ)-DP, which is suitable for the iterative algorithm due to its composability property.

In this article, we achieve (ϵ, δ)-DP for each user using the Gaussian mechanism [9], which provides privacy guarantee through adding Gaussian noise to the uploaded local update. The size of the noise is calibrated by the update's *sensitivity* which captures how much a single individual's data changes the value of this update in the worst case. Given any function $f : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ with L_2-sensitivity s_f, the Gaussian mechanism on f is $\mathcal{M}(\mathcal{A}) := f(\mathcal{A}) + \mathcal{N}(0, s_f^2\sigma^2)$, where $\mathcal{N}(0, s_f^2\sigma^2)$ is a normal distribution with mean 0 and standard variance $s_f\sigma$. It has been proved in [9] that the Gaussian mechanism $\mathcal{M}$ achieves (ϵ, δ)-DP if $\sigma \geq \sqrt{2\log(1.25/\delta)}/\epsilon$ with $\epsilon \in (0, 1)$.

B. Proposed Privacy-Preserving Scheme

Although it is hard to optimize all the unknown variables of (1) simultaneously, (1) can be solved by an alternating optimization procedure [8] since the objective is separable with respect to $\mathbf{W}$ and Ω. Specifically, we alternatively update $\mathbf{W}$ with fixed Ω and then update Ω with fixed $\mathbf{W}$ until convergence. In what follows, we present the details of these two steps.

1) Optimize Ω With Fixed $\mathbf{W}$: When $\mathbf{W}$ is fixed, the corresponding subproblem becomes to minimize the following:

$$\begin{aligned} \min_{\Omega} \ & \mathcal{P}(\Omega) := \lambda\,\mathrm{tr}(\mathbf{W}\Omega^{-1}\mathbf{W}^T) \\ \text{s.t.} \ & \Omega \succeq 0, \mathrm{tr}(\Omega) = 1 \end{aligned} \tag{2}$$

which has an analytical solution Ω^* [8], i.e.,

$$\Omega^* := \frac{\big(\mathbf{W}^T\mathbf{W}\big)^{1/2}}{\mathrm{tr}\Big(\big(\mathbf{W}^T\mathbf{W}\big)^{1/2}\Big)}. \tag{3}$$

We can see that Ω^* can be computed from the latest $\mathbf{W}$ without requiring any user data, so this step can be performed efficiently on the server.

2) Optimize $\mathbf{W}$ With Fixed $\boldsymbol{\Omega}$: When Ω is fixed, the subproblem becomes

$$\min_{\mathbf{W}} \mathcal{P}(\mathbf{W}) := \sum_{t=1}^{m}\sum_{i=1}^{n_t} \ell_t\big(\mathbf{w}_t^T\mathbf{x}_t^i, y_t^i\big) + \lambda\,\mathrm{tr}\Big(\mathbf{W}\Omega^{-1}\mathbf{W}^T\Big). \tag{4}$$

Since the overall data set $\{\mathcal{A}_t\}_{t=1}^m$ is distributed across devices, a distributed and parallel algorithm without requiring expensive raw data transfer is highly desirable.

Toward that goal, we use the block dual coordinate descent considering the fact that the dual of $\mathcal{P}(\mathbf{W})$ has a better separability property. By taking the conjugate dual of $\mathcal{P}(\mathbf{W})$, we obtain the following dual problem:

$$\min_{\boldsymbol{\alpha}} \mathcal{D}(\boldsymbol{\alpha}) := \sum_{t=1}^{m}\sum_{i=1}^{n_t} \ell_t^*\big(-\alpha_t^i\big) + \frac{1}{4\lambda}\|\mathbf{X}\boldsymbol{\alpha}\|_{\tilde{\Omega}}^2 \tag{5}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^n$ is a column vector of all dual variables with the $(\sum_{\tau=1}^{t-1} n_\tau + i)$th element α_t^i corresponding to the training sample $(\mathbf{x}_t^i, y_t^i)$, ℓ_t^* is the conjugate function of ℓ_t, i.e., $\ell_t^*(-\alpha) =$

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Privacy-Preserving Algorithm

Input: Datasets $\{\mathcal{A}_t, t = 1, \ldots, m\}$, aggregation parameter $\xi \in (0, 1]$, correction parameter β, inner global iteration number K, and outer global iteration number H.

Initialize: $\boldsymbol{\alpha} \leftarrow \mathbf{0}$, $\mathbf{w} \leftarrow \mathbf{0}$, and $\boldsymbol{\Omega} \leftarrow (1/m)\mathbf{I}$.

1: **for** $h = 1$ to H **do**
2: **for** $k = 1$ to K **do**
3: **for all devices** $t = 1, \ldots, m$ **in parallel do**
4: $\Delta\boldsymbol{\alpha}_t \leftarrow \operatorname{argmin}_{\Delta\boldsymbol{\alpha}_t} \mathcal{G}_t^{\beta}(\Delta\boldsymbol{\alpha}_t; \mathbf{w}_t, \boldsymbol{\alpha}_t)$;
5: $\boldsymbol{\alpha}_t \leftarrow \boldsymbol{\alpha}_t + \xi\Delta\boldsymbol{\alpha}_t$;
6: $\Delta\boldsymbol{\alpha}_t \leftarrow \Delta\boldsymbol{\alpha}_t + \mathbf{b}_t$;
7: $\Delta\mathbf{u}_t \leftarrow \xi\mathbf{X}_t\Delta\boldsymbol{\alpha}_t + \mathbf{p}_t$;
8: return $\Delta\mathbf{u}_t$ to the server;
9: **end for**
10: update $\mathbf{w} \leftarrow \mathbf{w} + \frac{1}{2\lambda}\widetilde{\Omega}\Delta\mathbf{u}$ in the server;
11: send the updated block $\mathbf{w}_t$ back to device t;
12: **end for**
13: update $\boldsymbol{\Omega} \leftarrow \frac{(\mathbf{W}^T\mathbf{W})^{\frac{1}{2}}}{\operatorname{tr}((\mathbf{W}^T\mathbf{W})^{\frac{1}{2}})}$ with the most recent $\mathbf{W}$ and send the block $\widetilde{\Omega}_t$ back to device t;
14: **end for**

$\max_v\{-\alpha v - \ell(v)\}$, and $\widetilde{\Omega} := \Omega \otimes \mathbf{I}_{d\times d} \in \mathbb{R}^{md\times md}$. We assume that ℓ_t^* is convex and differentiable with $|\ell_t^{*\prime}(z)| \leq 1$ for all z.

Due to the convexity of (4), we have $\mathcal{P}(\mathbf{W}^\star) = \mathcal{D}(\boldsymbol{\alpha}^\star)$, and hence the optimal primal variables can be derived from the optimal dual variables as

$$\mathbf{w}(\boldsymbol{\alpha}) := \frac{1}{2\lambda}\widetilde{\Omega}\mathbf{X}\boldsymbol{\alpha} \tag{6}$$

where $\mathbf{w}(\boldsymbol{\alpha}) \in \mathbb{R}^{md}$ is a column vector formed by concatenating m blocks of primal variables, with the tth block vector $\mathbf{w}_t(\boldsymbol{\alpha})$ being the primal variable of device t.

In the following, we develop an iterative search algorithm to solve the dual (5). Specifically, given the current solution $\boldsymbol{\alpha}$ and $\mathbf{w}$, we define the following subproblem for device t at each iteration:

$$\min_{\Delta\boldsymbol{\alpha}_t} \mathcal{G}_t^{\beta}(\Delta\boldsymbol{\alpha}_t; \mathbf{w}_t, \boldsymbol{\alpha}_t) := \sum_{i=1}^{n_t} \ell_t^*\left(-\alpha_t^i - \Delta\alpha_t^i\right) + \mathbf{w}_t^T\mathbf{X}_t\boldsymbol{\Delta\alpha}_t + \frac{\beta}{4\lambda}\|\mathbf{X}_t\boldsymbol{\Delta\alpha}_t\|_{\widetilde{\Omega}_t}^2 \tag{7}$$

where $\boldsymbol{\alpha}_t \in \mathbb{R}^{n_t}$ is the tth block vector of $\boldsymbol{\alpha}$ representing the dual variables of device t, $\widetilde{\Omega}_t \in \mathbb{R}^{d\times d}$ refers to the tth diagonal block of $\widetilde{\Omega}$, and $\beta > 0$ is the correction parameter.

Note that in the traditional block dual coordinate descent, each local update minimizes the global objective based on all updated coordinates. However, in our approach, each local update minimizes the local objective based on the previous values of local coordinates, which can be executed in parallel and decrease the training efficiency. To compensate for such differences, the correction parameter β needs to be chosen carefully to ensure the sum of the local objectives of all devices approximately equal to the global objective $\mathcal{D}$.

Algorithm 1 outlines our privacy-preserving scheme using the Gaussian mechanism. Our algorithm contains two parts: 1) update $\mathbf{W}$ (lines 2–12) and 2) update $\boldsymbol{\Omega}$ (line 13). In part 1), each device first solves its own local (7) and uploads its local update $\Delta\mathbf{u}_t$ to the server. Then the server concatenates all $\Delta\mathbf{u}_t$ as $\Delta\mathbf{u}$ and updates the global parameters $\mathbf{W}$, which are sent back to the corresponding device. In part 2), the server updates $\boldsymbol{\Omega}$ using the most recent $\mathbf{W}$ and sends the block result $\widetilde{\Omega}_t$ to the corresponding device. This process iterates multiple rounds until convergence. The noise vectors $\mathbf{b}_t$ and $\mathbf{p}_t$ in lines 6 and 7 are drawn independently from the Gaussian distributions $\mathcal{N}(0, s_1^2\sigma_1^2)$ and $\mathcal{N}(0, s_2^2\sigma_2^2)$ and used to achieve (ϵ_1, δ_1)-DP and $(\epsilon_2.\delta_2)$-DP, respectively. Here, s_1 and s_2 represent the L_2-sensitivity of $\Delta\boldsymbol{\alpha}_t$ and $\Delta\mathbf{u}_t$, respectively. Note that H is the number of outer global iterations and K is the number of inner global iterations, which should be determined beforehand.

In this article, we consider a heterogeneous setting where devices have different network conditions, computation capabilities, and battery capacities. As described in Algorithm 1, devices need to compute and share their local updates for many iterations. During this process, devices may drop out if they run out of their resources (e.g., running out of power or getting disconnected from the network), which is named as node dropping. In this case, dropped devices are unable to share their computation results to the server. Here, we assume that a device will not always drop out, which means the device will recover and rejoin the training process if it was dropped out before. Besides, at each global iteration (either inner or outer global iteration), devices perform multiple local iterations in a given time period before sending their computation results to the server. However, as devices have different computation capabilities, they will perform different numbers of local iterations. We name this phenomenon as device variability. In the following, we provide a rigorous analysis of the convergence rate and the privacy loss of our algorithm, considering the node dropping and device variability.

V. Convergence Analysis

In this section, we analyze the convergence properties of our proposed Algorithm 1. Since (1) is jointly convex with respect to $\mathbf{W}$ and Ω, the alternating optimization is guaranteed to converge to the optimal solution. As it is easy to optimize $\boldsymbol{\Omega}$ given $\mathbf{W}$, we focus on analyzing the convergence of updating $\mathbf{W}$ in the rest of this section. Following the discussion of device heterogeneity, we first introduce an approximation parameter to quantify the quality of each update.

Definition 1 (Quality of Update): At each iteration k, we define the quality measurement of the solution calculated by device t to its subproblem as

$$\theta_t^k = \frac{\mathcal{G}_t^{\beta}\left(\Delta\boldsymbol{\alpha}_t^k; \mathbf{w}_t^k, \boldsymbol{\alpha}_t^k\right) - \mathcal{G}_t^{\beta}\left(\Delta\boldsymbol{\alpha}_t^\star; \mathbf{w}_t^k, \boldsymbol{\alpha}_t^k\right)}{\mathcal{G}_t^{\beta}\left(\mathbf{0}; \mathbf{w}_t^k, \boldsymbol{\alpha}_t^k\right) - \mathcal{G}_t^{\beta}\left(\Delta\boldsymbol{\alpha}_t^\star; \mathbf{w}_t^k, \boldsymbol{\alpha}_t^k\right)} \tag{8}$$

where $\theta_t^k \in [0, 1]$ and $\Delta\boldsymbol{\alpha}_t^\star$ is the exact minimizer of $\mathcal{G}_t^{\beta}(\Delta\boldsymbol{\alpha}_t^k; \mathbf{w}_t^k, \boldsymbol{\alpha}_t^k)$. $\theta_t^k = 0$ refers that the update is the exact solution, and $\theta_t^k = 1$ indicates that the update of model t makes no progress at iteration k.

According to Definition 1, if a device drops out at iteration k, then $\theta_t^k = 1$, otherwise $\theta_t^k \in [0, 1)$. Since a device will not always drop out, we have the probability of node dropping for any device $P(\theta_t^k = 1) < p_{\max}$ with $0 \leq p_{\max} < 1$. In

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

addition, due to the device variability, at iteration k, the local updates of devices will have different qualities. Generally, if a device has the computation capability to perform more local iterations, it will achieve higher quality of update. Here, in Assumption 1, we assume that the update of each device at each iteration will be more accurate than the previous one on average, which is a customary assumption for gradient descent-based algorithms [7].

Assumption 1: Let $\mathcal{I}_k = (\boldsymbol{\alpha}^k, \boldsymbol{\alpha}^{k-1}, \ldots, \boldsymbol{\alpha}^1)$ be a vector of the previous dual variables until iteration k and the expectation of θ_t^k under the previous values be $\Theta_t^k = \mathbb{E}(\theta_t^k|\mathcal{I}_k)$. We assume that $\mathbb{E}(\theta_t^k|\mathcal{I}_k, \theta_t^k < 1) \leq \Theta_{\max}$ with $0 \leq \Theta_{\max} < 1$.

Based on Assumption 1, we derive Theorems 3 and 4 which characterize the convergence of our privacy-preserving algorithm with respect to smooth and nonsmooth loss functions, respectively. Before that, we first introduce some lemmas used for both smooth and nonsmooth cases.

To make the local dual objective approximate to the global dual objective with respect to varying $\Delta\boldsymbol{\alpha}_t$, we choose β as follows.

Lemma 1: For any dual variable $\boldsymbol{\alpha} \in \mathbb{R}^n$ and the change of it $\Delta\boldsymbol{\alpha} = (\Delta\boldsymbol{\alpha}_1, \ldots, \Delta\boldsymbol{\alpha}_m)^T$, aggregation parameter $\xi \in (0, 1]$ and correction parameter β, when

$$\beta \geq \xi \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{\|\mathbf{X}\boldsymbol{\alpha}\|_{\widetilde{\Omega}}^2}{\sum_{t=1}^m \|\mathbf{X}_t\boldsymbol{\alpha}_t\|_{\widetilde{\Omega}_t}^2} \tag{9}$$

it holds that

$$\mathcal{D}(\boldsymbol{\alpha} + \xi\Delta\boldsymbol{\alpha}) \leq (1-\xi)\mathcal{D}(\boldsymbol{\alpha}) + \xi \sum_{t=1}^m \mathcal{G}_t^\beta(\Delta\boldsymbol{\alpha}_t; \mathbf{w}_t, \boldsymbol{\alpha}_t).$$

Proof: The proof is similar to the proof of [10, Lemma 3]. ■

Here, we choose $\beta = \xi m$ to ensure that the inequality (9) is always satisfied [10]. Next, we show that the expected objective gap is bounded in Lemmas 2 and 3.

Lemma 2: For the loss function ℓ_t with its conjugate function ℓ^* is convex and $|\ell_t^{*\prime}| \leq 1$ and any $s \in (0, 1]$

$$\begin{aligned}&\mathbb{E}\left[\mathcal{D}\left(\boldsymbol{\alpha}^k\right) - \mathcal{D}\left(\boldsymbol{\alpha}^k + \xi\sum_{t=1}^m \Delta\boldsymbol{\alpha}_t^k\right)|\mathcal{I}_k\right]\\ &\geq \xi\left(1-\overline{\Theta}\right)\left(sG\left(\boldsymbol{\alpha}^k\right) + \sum_{t=1}^m J_t\right)\end{aligned} \tag{10}$$

with

$$\begin{cases} G(\boldsymbol{\alpha}^k) := \sum_{t=1}^m \sum_{i=1}^{n_t} \ell_t^*(-\alpha_t^i) + \ell_t(\mathbf{w}_t(\boldsymbol{\alpha})^T\mathbf{x}_t^i, y_t^i) \\ \qquad\qquad + \mathbf{w}_t(\boldsymbol{\alpha})^T\mathbf{x}_t^i\alpha_t^i \\ J_t := \frac{\mu(1-s)s}{2}\left\|\mathbf{u}_t - \boldsymbol{\alpha}_t^k\right\|^2 - \frac{\beta s^2}{4\lambda}\left\|\mathbf{X}_t\left(\mathbf{u}_t - \boldsymbol{\alpha}_t^k\right)\right\|_{\widetilde{\Omega}_t}^2 \end{cases}$$

where $\mathbf{u}_t = \partial\ell_t(\mathbf{w}_t^T\mathbf{x}_t^i, y_t^i)$ and $\overline{\Theta} = p_{\max} + (1 - p_{\max})\Theta_{\max}$.

Proof: The proof is given in Appendix E. ■

Lemma 3: For the loss function ℓ_t with its conjugate function ℓ^* is convex and $|\ell_t^{*\prime}| \leq 1$ and any $s \in (0, 1]$

$$\mathbb{E}\left[\mathcal{D}\left(\boldsymbol{\alpha}^k\right) - \mathcal{D}\left(\boldsymbol{\alpha}^{k+1}\right)|\mathcal{I}_k\right] \geq (1-\xi)\xi\left(1-\overline{\Theta}\right)sG\left(\boldsymbol{\alpha}^k\right) + R \tag{11}$$

with

$$R := \begin{cases} 0, & \text{if } \ell_t \text{ is } (1/\mu)-\text{smooth} \\ -\frac{(1-\xi)}{\lambda}\xi\left(1-\overline{\Theta}\right)s^2L^2\beta\sum n_t^2, & \text{if } \ell_t \text{ is } L-\text{Lipschitz.}\end{cases}$$

Here, $G(\boldsymbol{\alpha}^k) = D(\boldsymbol{\alpha}^k) - (-P(\mathbf{W}(\boldsymbol{\alpha}^k)))$ and $\overline{\Theta} = p_{\max} + (1 - p_{\max})\Theta_{\max}$.

Proof: The proof is given in Appendix F. ■

A. Convergence Rate of Smooth Cases

Based on the above lemmas, we obtain the convergence rate of our privacy-preserving algorithm for smooth loss functions in Theorem 3.

Theorem 3 (Convergence Rate for Smooth Losses): Assume the loss function ℓ_t is $(1/\mu)$-smooth such that ℓ_t^* is μ-strongly convex and $|\ell_t^{*\prime}(z)| \leq 1$ for all z. Under Assumption 1, there exists a constant $s \in (0, 1)$ such that

$$K \geq \frac{1}{(1-\xi)\xi\left(1-\overline{\Theta}\right)s}\log\left(\frac{n}{\epsilon_G}\right) \tag{12}$$

will satisfy that $\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^K) - \mathcal{D}(\boldsymbol{\alpha}^*) \leq \epsilon_G$. Here, $\overline{\Theta} := p_{\max} + (1 - p_{\max})\Theta_{\max}$.

Proof: We first define the duality gap as $G(\boldsymbol{\alpha}) = D(\boldsymbol{\alpha}) - (-P(\mathbf{W}(\boldsymbol{\alpha})))$, and for each (7), we choose the correction parameter β and the aggregation parameter ξ according to Lemma 1. To prove Theorems 3 and 4, we first show that $\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^k) - \mathcal{D}(\boldsymbol{\alpha}^{k+1})|\mathcal{I}_k]$ has a lower bound in Lemma 3 since

$$\begin{aligned}\mathbb{E}\left[\mathcal{D}\left(\boldsymbol{\alpha}^{k+1}\right) - \mathcal{D}\left(\boldsymbol{\alpha}^*\right)|\mathcal{I}_k\right] &= \mathcal{D}\left(\boldsymbol{\alpha}^k\right) - \mathcal{D}\left(\boldsymbol{\alpha}^*\right)\\ &\quad - \mathbb{E}\left[\mathcal{D}\left(\boldsymbol{\alpha}^k\right) - \mathcal{D}\left(\boldsymbol{\alpha}^{k+1}\right)|\mathcal{I}_k\right].\end{aligned} \tag{13}$$

Now, we derive the upper bound of $\mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^{k+1}) - \mathcal{D}(\boldsymbol{\alpha}^*)|\mathcal{I}_k]$ based on the above theorems and (13). The following proof is similar to the proof of [7, Th. 1]. Since

$$\begin{aligned}&\mathbb{E}\left[\mathcal{D}\left(\boldsymbol{\alpha}^{k+1}\right) - \mathcal{D}\left(\boldsymbol{\alpha}^*\right)|\mathcal{I}_k\right]\\ &\leq \mathcal{D}\left(\boldsymbol{\alpha}^k\right) - \mathcal{D}\left(\boldsymbol{\alpha}^*\right) - (1-\xi)\xi\left(1-\overline{\Theta}\right)sG\left(\boldsymbol{\alpha}^k\right) - R\\ &\leq \left(1-(1-\xi)\xi(1-\overline{\Theta})s\right)\left(\mathcal{D}\left(\boldsymbol{\alpha}^k\right) - \mathcal{D}\left(\boldsymbol{\alpha}^*\right)\right) - R\end{aligned} \tag{14}$$

by recursively applying the above inequality, we arrive at

$$\begin{aligned}&\mathbb{E}\left[\mathcal{D}\left(\boldsymbol{\alpha}^{k+1}\right) - \mathcal{D}\left(\boldsymbol{\alpha}^*\right)|\mathcal{I}_k\right]\\ &\leq -\frac{R}{(1-\xi)\xi\left(1-\overline{\Theta}\right)s}\\ &\quad + \left(\left(1-(1-\xi)\xi(1-\overline{\Theta})s\right)\right)^{k+1}\left(\mathcal{D}\left(\boldsymbol{\alpha}^0\right) - \mathcal{D}\left(\boldsymbol{\alpha}^*\right)\right).\end{aligned} \tag{15}$$

If the loss function ℓ_t is $(1/\mu)$-smooth, $R = 0$ by Lemma 3. Therefore, if we denote by $\epsilon_G^k = \mathcal{D}(\boldsymbol{\alpha}^k) - \mathcal{D}(\boldsymbol{\alpha}^*)$ we have that

$$\begin{aligned}\mathbb{E}\left[\epsilon_G^k|\mathcal{I}_k\right] &\leq \left(\left(1-(1-\xi)\xi(1-\overline{\Theta})s\right)\right)^k\epsilon_G^0\\ &\leq n\exp -k(1-\xi)\xi(1-\overline{\Theta})s.\end{aligned} \tag{16}$$

Here, we have $(\mathcal{D}(\boldsymbol{\alpha}^0) - \mathcal{D}(\boldsymbol{\alpha}^*)) \leq n$ by [11, Lemma 10]. Then, the right-hand side will be smaller than some ϵ_G if

$$k \geq \frac{1}{(1-\xi)\xi\left(1-\overline{\Theta}\right)s}\log\left(\frac{n}{\epsilon_G}\right). \tag{17}$$

This concludes our proof. ■

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

B. Convergence Rate of Nonsmooth Cases

Here, we derive the convergence rate of our privacy-preserving algorithm for nonsmooth loss functions in Theorem 4.

Theorem 4 (Convergence Rate for Nonsmooth Losses): Assume the loss function ℓ_t is L-Lipschitz. Under Assumption 1, when

$$K \geq K_0 + \left\lceil \frac{1}{(1-\xi)(1-\overline{\Theta})} \max\left(1, \frac{2mL^2 \sum n_t^2}{\lambda n^2 \epsilon_G}\right) \right\rceil$$
$$K_0 \geq k_0 + \left\lceil \frac{2}{(1-\xi)(1-\overline{\Theta})} \left(\frac{4mL^2 \sum n_t^2}{\lambda n^2 \epsilon_G} - 1\right) \right\rceil$$
$$k_0 \geq \max\left(0, \left\lceil (1-\xi)\left(1-\overline{\Theta}\right) \log \frac{\left(\mathcal{D}\left(\boldsymbol{\alpha}^0\right) - \mathcal{D}(\boldsymbol{\alpha}^*)\right)}{mL^2 \sum n_t^2 / 2\lambda n^2} \right\rceil\right)$$

it holds that $\mathbb{E}(\mathcal{D}(\overline{\boldsymbol{\alpha}}) - \mathcal{D}(\boldsymbol{\alpha}^*)) \leq \epsilon_G$ at the averaged iterate $\overline{\boldsymbol{\alpha}} = (1/K - K_0) \sum_{k=K_0+1}^{K} \boldsymbol{\alpha}^k$. Here, $\overline{\Theta} := p_{\max} + (1 - p_{\max})\Theta_{\max}$.

Proof: The proof of Theorem 4 is similar to the proof of Theorem 3. According to Lemmas 2 and 3, when the loss function ℓ_t is L-Lipschitz, we follow the derivation of Theorem 9 in [11] and finally obtain that when:

$$K \geq K_0 + \left\lceil \frac{2L^2 \beta \sum n_t^2}{\lambda \xi (1-\xi)\left(1-\overline{\Theta}\right) n^2 \epsilon_G} \right\rceil$$
$$K_0 \geq k_0 + \left\lceil \frac{2}{(1-\xi)\xi\left(1-\overline{\Theta}\right)} \left(\frac{4L^2 \beta \sum n_t^2}{\lambda n^2 \epsilon_G} - 1\right) \right\rceil$$
$$k_0 \geq \max\left(0, \left\lceil (1-\xi)\xi\left(1-\overline{\Theta}\right) \log \frac{\left(\mathcal{D}\left(\boldsymbol{\alpha}^0\right) - \mathcal{D}(\boldsymbol{\alpha}^*)\right)}{2L^2 \beta \sum n_t^2 / \lambda n^2} \right\rceil\right)$$

it holds that $\mathbb{E}(\mathcal{D}(\overline{\boldsymbol{\alpha}}) - \mathcal{D}(\boldsymbol{\alpha}^*)) \leq \epsilon_D$ at the averaged iterate $\overline{\boldsymbol{\alpha}} = (1/K - K_0) \sum_{k=K_0+1}^{K} \boldsymbol{\alpha}^k$. Then, Theorem 4 follows by plugging in parameters $\beta \geq m\xi$. ■

VI. Privacy Analysis

In this section, we rigorously analyze the end-to-end privacy guarantee of our personalized federated learning scheme. For a mechanism that achieves (ϵ, δ)-DP, the corresponding privacy loss will be bounded by ϵ with probability at least $1 - \delta$. The composability property of differential privacy enables us to account the privacy loss for each access to the training data and accumulate this cost for the whole training process. Recently, some advanced composition theorems [6], [9], [12]–[14] have been proposed to achieve tighter analysis of the privacy loss for multiple iterations.

Assume that each iteration of Algorithm 1 is (ϵ, δ)-differentially private. When using the composability property, Algorithm 1 is $(K\epsilon, K\delta)$-differential private after K iterations. However, by the strong composition theorem presented in [12] and [13], Algorithm 1 will be $(\epsilon\sqrt{K \log 1/\delta}, K\delta)$-differential private, and this bound can be further tightened by combining with the privacy amplification theorem proposed in [14] which makes the composed mechanism to be $(O(q\epsilon), O(q\delta))$-differentially private if the training data at each iteration is a random sample from the data set with sampling probability q. Recently, a stronger method known as moments accountant has been proposed in [6], which saves a $\sqrt{\log 1/\delta}$ factor of the ϵ part and Kq factor of the δ part. In the following, we analyze the privacy loss of our algorithm using the moments accountant.

The application of moments accountant to our proposed scheme is not straightforward. Due to the device variability, some devices will run more local iterations (where data are randomly sampled to perform updating at each local iteration) so that more data are accessed at a global iteration, and thus more privacy is leaked. Besides, due to the node dropping, dropped devices are unable to upload their computation results to the server at a certain global iteration and thus no privacy loss occurred at that global iteration. In order to count the overall privacy loss, we let q denote the data sampling probability at each global iteration, which captures the sampling probability of data after multiple local iterations for each device. Then, we let p denote the node active probability which is equal to $1 - P(\theta_t^k = 1)$, where $P(\theta_t^k = 1)$ represents the probability of node dropping. We first account the privacy loss incurred at each global iteration. Before that, we analyze the L_2-sensitivity of $\Delta\boldsymbol{\alpha}_t$ and $\Delta\mathbf{u}_t$ that calibrate the size of Gaussian noises added to them. To calculate the sensitivities, we first show that the local objective is strongly convex.

Corollary 1: If ℓ^* is convex and differentiable with $|\ell^{*\prime}(z)| \leq 1$ for all z, then function $\operatorname{argmin}_{\Delta\boldsymbol{\alpha}_t} \mathcal{G}_t^{\beta}(\Delta\boldsymbol{\alpha}_t, \mathcal{A})$ is $(\beta n_t/2\lambda)$-strongly convex.

Proof: The proof is given in Appendix A. ■

Based on Corollary 1, we can estimate the L_2-sensitivity of $\Delta\boldsymbol{\alpha}_t$ and $\Delta\mathbf{u}_t$, whose values are given in Corollaries 2 and 3, respectively.

Corollary 2: If ℓ^* is convex and differentiable with $|\ell^{*\prime}(z)| \leq 1$ for all z, the L_2-sensitivity of $\operatorname{argmin}_{\Delta\boldsymbol{\alpha}_t} \mathcal{G}_t^{\beta}(\Delta\boldsymbol{\alpha}_t; \mathbf{w}_t, \boldsymbol{\alpha}_t)$ is at most $(1/\beta n_t)(8\lambda + \beta)$.

Proof: The proof is given in Appendix B. ■

Corollary 3: When $\Delta\boldsymbol{\alpha}_t$ is known, the L_2-sensitivity of $\Delta\mathbf{u}_t$ is at most $2\xi \|\Delta\boldsymbol{\alpha}_t\|_2$.

Proof: The proof is given in Appendix C. ■

Next, using the results in Corollaries 2 and 3, we account the privacy loss incurred at each global iteration in Lemma 4. First, we obtain the moments accountant of each global iteration of Algorithm 1, as given in Corollary 4.

Corollary 4: If $\sigma_1 \geq 1$ and the data sampling probability $q < (1/16\sigma_1)$. Then, for any integer $\gamma \in (0, \sigma_1^2 \ln(1/q\sigma_1))$, each global iteration of Algorithm 1 satisfies

$$\mu_{\mathcal{M}_{1:2}}(\gamma) \leq \frac{\gamma(\gamma+1)}{2\sigma_2^2} + \frac{q^2\gamma(\gamma+1)}{(1-q)\sigma_1^2} + O\left(q^3\gamma^3/\sigma_1^3\right).$$

Proof: The proof is given in Appendix D. ■

Lemma 4 (Per-Iteration Privacy Loss): Assume s_1 and s_2 are the sensitivity of $\Delta\boldsymbol{\alpha}_t$ and $\Delta\mathbf{u}_t$, respectively, $\delta_1 = \delta_2 = \delta \in (0, 1)$, and $\epsilon_1, \epsilon_2 \in (0, 1)$. Given the data sampling probability q, Algorithm 1 is $(\sqrt{q^2\epsilon_1^2 + \epsilon_2^2}, \delta)$-DP for device t at each global iteration if

$$\sigma_1 \geq \frac{\sqrt{2\ln\frac{1.25}{\delta}}}{\epsilon_1}, \quad \sigma_2 \geq \frac{\sqrt{2\ln\frac{1.25}{\delta}}}{\epsilon_2}.$$

Proof: We compute the moments accountant of each iteration of our algorithm according to the definitions and

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

then obtain the moments accountant for the whole algorithm by Theorem 1. Finally, we can derive the privacy guarantee of our algorithm by converting the moments accountant to a (ϵ, δ)-differentially privacy guarantee according to Theorem 2.

By Corollary 4, we have the moments accountant of iteration k that is $\mu_{\mathcal{M}_k}(\gamma) \leq [(q^2\gamma^2)/(\sigma_1^2)] + (\gamma^2/\sigma_2^2)$. By Theorem 2, we can convert $\mu_{\mathcal{M}_k}(\gamma)$ into the differentially privacy guarantee of each global iteration of our algorithm. Then, we have the following optimization problem to be solved:

$$\log\delta = \min_{\gamma} \left(q^2\gamma^2/\sigma_1^2 + \gamma^2/\sigma_2^2 - \gamma\epsilon\right)$$
$$\text{subject to } \gamma \leq \sigma_1^2 \log\left(\frac{1}{q\sigma_1}\right), \quad \gamma \in \mathbb{Z}^+$$
$$q \leq \frac{1}{16\sigma_1}, \quad \sigma_1 \geq 1.$$

Given fixed σ_1 and σ_2 and δ, if we could find some values of $\epsilon > 0$, $c_0 \in (0, 1)$ and a positive integer γ so that

$$\left(\frac{q^2}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\gamma^2 \leq c_0\gamma\epsilon \text{ and } \log\delta \geq (c_0 - 1)\gamma\epsilon$$

each iteration of Algorithm 1 can be (ϵ, δ)-differentially private. The above two inequalities are equivalent to

$$\frac{\log 1/\delta}{\epsilon(1 - c_0)} \leq \frac{c_0\epsilon}{\frac{q^2}{\sigma_1^2} + \frac{1}{\sigma_2^2}}.$$

We can see that the positive integer γ exists if $([\log(1/\delta)]/[\epsilon(1 - c_0)]) \leq ([c_0\epsilon]/[q^2/\sigma_1^2 + 1/\sigma_2^2])$, which is equivalent to the following condition:

$$\epsilon^2 \geq \frac{\log(1/\delta)\left(q^2/\sigma_1^2 + 1/\sigma_2^2\right)}{c_0(1 - c_0)}.$$

It is easy to verify that there exists a constant c_1 such that

$$\epsilon^2 = c_1 \log(1/\delta)\left(q^2/\sigma_1^2 + 1/\sigma_2^2\right).$$

Corollary 4 follows by plugging $\sigma_1 = \sqrt{2\ln(1.25/\delta)}/\epsilon_1$ and $\sigma_2 = \sqrt{2\ln(1.25/\delta)}/\epsilon_2$ into the above equality. ■

Now, we account the privacy loss of the whole algorithm given the node active probability p. We give the conclusion in Theorem 5 in terms of a privacy guarantee for Algorithm 1, which helps us to allocate the Gaussian noise directly.

Theorem 5 (End-to-End Privacy Loss): There exist constants c_1 and c_2 so that given the node active probability p and the number of global iteration K, for any $\epsilon < c_1p^2K$, Algorithm 1 is (ϵ, δ)-differential private for any $\delta > 0$ if we choose

$$\sigma_1 \geq c_2\frac{p\sqrt{\left(q^2 + r^2\right)K\log(1/\delta)}}{\epsilon}$$
$$\sigma_2 \geq c_2\frac{p\sqrt{\left(q^2 + r^2\right)K\log(1/\delta)}}{r\epsilon}$$

where the parameter r refers to the ratio of privacy budgets at steps 6 and 7, i.e., $\epsilon_2 = r\epsilon_1$ with $r \geq 0$.

Proof: Assume each iteration of Algorithm 1 is (ϵ, δ)-differentially private, and the iteration number is K. Given the node active probability p, we have that Algorithm 1 is $(\epsilon p\sqrt{K}, \delta)$-differentially private using the moments accountant. Then, the result follows by an application of Lemma 4 with ϵ_2 being represented by $\epsilon_2 = r\epsilon_1$. ■

VII. Evaluation

In this section, we evaluate our approach against device heterogeneity and different privacy budgets. We first show the convergence properties of our approach and then study the impact of device heterogeneity and the tradeoff between accuracy and privacy.

A. Experimental Setting

We evaluate our approach on the Human Activity Recognition Using Smartphones Data Set (HAR data set) [15]. It is collected by monitoring six different activities (walking, walking upstairs, walking downstairs, sitting, standing, and laying) of 30 individuals, using the accelerometer and gyroscope embedded in the mobile phone. The data set includes 10 299 instances in total with 561 features, and 210–306 instances per individual. All data are normalized locally by l_2-normalization. We train models for each individual and predict between sitting and other activities using 75% of the data for training and 25% for testing.

We use the hinge loss $\ell(u) = \max(0, 1 - yu)$ as the loss function. It is L-Lipschitz, and its dual is $\ell^*(-\alpha) = -\alpha y$ with $\alpha y \in [0, 1]$. We use the stochastic dual coordinate ascent (SDCA) as the local solver which selects one coordinate to update randomly at each iteration [16]. Since SDCA samples the data point with probability $1/n_t$ at each iteration, the data sampling probability q is n_{iter}/n_t, where n_{iter} is the local iteration number of device t. The node active probability $p = 1 - P(\theta_t^k = 1)$, where $P(\theta_t^k = 1)$ is the probability of node dropping of device t at global iteration k.

We evaluate our approach in both homogeneous and heterogeneous scenarios. In the homogeneous scenario, the server waits for all devices to upload their updates at each time, and no device will drop out during the process. In the heterogeneous scenario, all devices have to upload their updates in a fixed global clock cycle at each time, and each device will drop out with a certain probability $P(\theta_t^k = 1)$. We simulate the device variability via varying the local iteration numbers of each device. We use $\tau \in [0, 1]$ to measure the device variability level. The local iteration numbers of devices are uniformly distributed between $(1 - \tau)n_{\min}$ and $n_{\min}$, where $n_{\min}$ is the minimum number of local data points across devices. Besides, in each scenario, we compare our approach with the baseline approach, i.e., the nonprivate personalized federated learning scheme, where no noise is added to the updates.

For each experiment, the number of inner global iterations K is set as 2000 and the number of outer global iterations H is set as 10. Besides, we use a fivefold cross-validation to choose the best hyperparameters λ and γ. We train and test all models for ten times and report the average results.

B. Numerical Results

1) Convergence Properties of Our Approach: We evaluate the convergence properties of our approach in both the

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

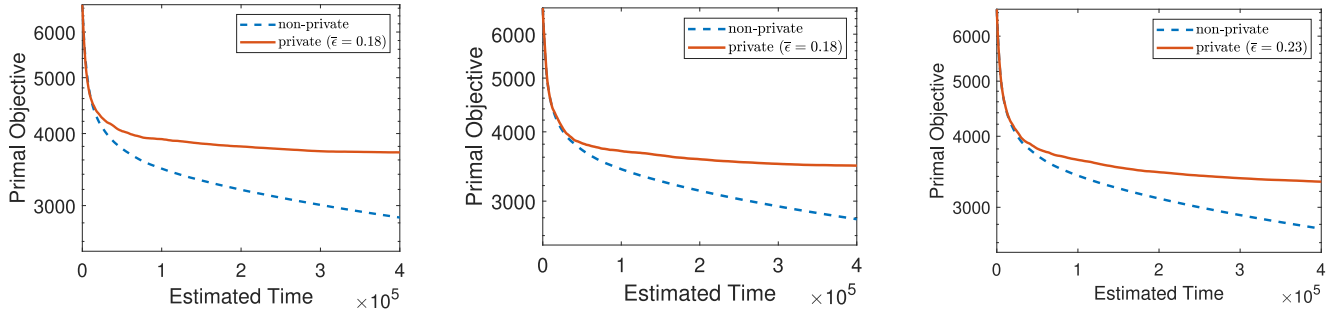


Fig. 2. Convergence properties of our approach: (a) homogeneous setting; (b) heterogeneous setting with device variability ($\tau = 0.5$); and (c) heterogeneous setting with device variability ($\tau = 0.5$) and node dropping [$P(\theta_t^k = 1) = 0.2$].

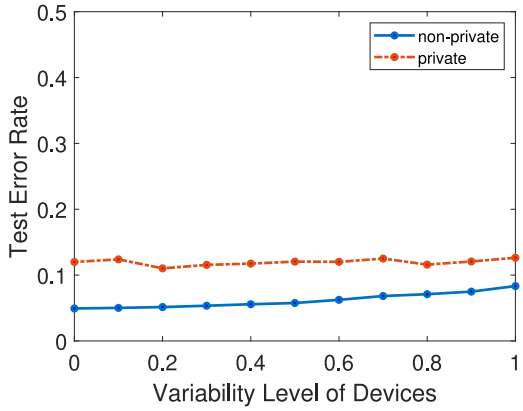


Fig. 3. Impact of device variability.

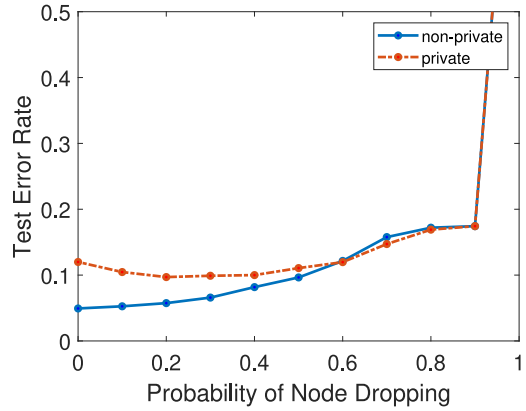


Fig. 4. Impact of the drop out of devices during the training.

homogeneous and heterogeneous scenarios considering device variability and node dropping. In our approach, we achieve (ϵ, δ)-differential privacy for each user using Algorithm 1. Specifically, we set $\epsilon = 8$ and $\delta = 10^{-3}$ and calculate the privacy budget of each global iteration $\bar{\epsilon}$ for each experiment.

We first compare the learning progress of our approach with the baseline approach (i.e., the nonprivate personalized federated learning scheme) under homogeneous and heterogeneous scenarios as shown in Fig. 2. Specifically, Fig. 2(a) shows the change of the primal objective value with respect to the overall running time (which is proportional to the iteration number) in the homogeneous scenario. We can observe that our approach will converge quickly. However, due to the addition of random noise at each iteration in our approach, it will only converge to a suboptimal value compared with the nonprivate approach, which matches the intuition that differential privacy guarantee comes with a utility loss. Similar results are observed under the heterogeneous scenarios, which are shown in Fig. 2(b) (with device variability) and Fig. 2(c) (with both device variability and dropping).

2) Impact of Device Heterogeneity: In this set of experiments, we evaluate the test error rate of the learned models in our approach under different device heterogeneity settings. We first study the impact of device variability by measuring the test error rate with respect to different variability levels of devices τ as shown in Fig. 3. As we can observe from the figure, our approach is robust to the device variability and the error rate of the learned model is almost stable even when the variability level increases. In comparison, the nonprivate baseline approach will be affected more by the device variability. The reason is that in our approach, the randomness introduced by the device variability reduces the size of noises and thus eliminates the impact of privacy on the accuracy.

Then, we study the impact of node dropping by assigning the probability $P(\theta_t^k = 1)$ to be a random number within [0, 1] as shown in Fig. 4. The results show that the error rate of the learned models in our approach first decreases to 0.097 when $P(\theta_t^k = 1) = 0.2$ and then increases as more nodes start to drop out. In our approach, the randomness introduced by the node dropping reduces the size of noises which means the test error rate will decrease. According to the test error rate of the baseline approach, the test error rate increases as the probability of node dropping increases. Thus, there exists an optimal point, i.e., when $P(\theta_t^k = 1) = 0.2$, which generates the minimum test error rate. We can see that node dropping does not always make things worse but brings us benefits sometimes. Therefore, sometimes extra dropping of updates will be needed in order to achieve better accuracy of models while preserving the privacy.

3) Tradeoff Between Accuracy and Privacy: In the last set of experiments, we measure the error rate of learned models corresponding to different privacy budgets, and in each case,

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

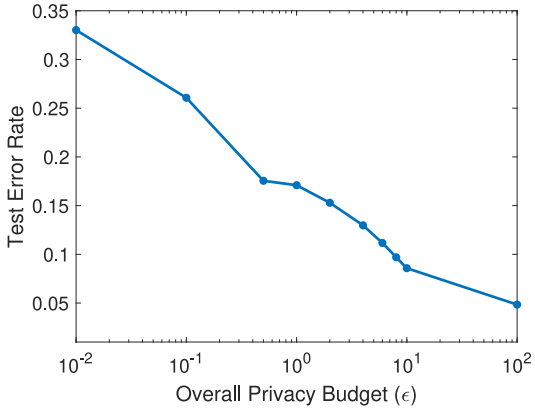


Fig. 5. Tradeoff between accuracy and privacy.

we set $\tau = 1$ and $P(\theta_1^k = 1)$ as the optimal probability that minimizes the error rate. Here, the privacy budget ϵ indicates the overall end-to-end privacy loss for one user and smaller ϵ implies higher privacy. As shown in Fig. 5, by increasing the value of ϵ from 0.01 to 100 with $\delta = 10^{-3}$, the corresponding test error rate keeps decreasing, matching the intuition that higher privacy corresponds to lower utility.

VIII. Related Work

Federated learning uses multiple devices to collaboratively train a shared model in an iterative manner while keeping all the data on devices. Specifically, all devices update a model downloaded from a central server independently using their own data and then upload the updates to the server to improve the shared model. Most of the work in federated learning has focused on the consensus problems [1], [2], [17]–[22] which are aimed to learn one global model distributedly. In contrast, we tackle the case where multiple personalized models are trained collaboratively based on relationships among all participants, which is known as multitask learning.

Multitask learning can be generally categorized into two categories based on how they capture relationships amongst tasks. The first category (e.g., [7], [8], [23], and [24]) assumes that the relationships are not known beforehand and can be learned from the data sets of tasks. On the other hand, the second category (e.g., [25] and [26]) assumes that a clustered, sparse, or low-rank structure between the tasks is known as *a priori*. In this article, we focus on the first category, which is more general and the relationships amongst tasks may not be known beforehand in practice. Moreover, different from the traditional multitask learning approaches where all learning tasks are performed on a single machine, we consider a federated learning setting where learning tasks are performed on different edge devices and a cloud server will coordinate the learning process. Since devices are heterogeneous in practice, the training process becomes much more complex and challenging. Besides, some works recently studied the problem of collaborative learning of personalized models, however, they did not address any privacy issue [7], [24], [27]. Although some private and personalized learning scheme with a fully decentralized architecture like [28] has been proposed, the architecture with central servers will be more efficient especially for applications that are large scale and require high system agility. Moreover, they did not consider the device heterogeneity in the real world.

For a distributed system coordinated by a central server, the privacy issue arises when an honest-but-curious server or device has access to the data or models. There exist several kinds of attacks in addition to the direct access of raw data: reconstruction attacks which recover training data from learned knowledge [3], model inversion attacks which create an adversarial example that resemble those used to create the model based on the responses received from that model [29], and membership inference attacks which determine if the sample was a member of the training set through querying the model [4].

Differential privacy [9] is especially effective in preventing membership inference attacks and reconstruction attacks. The differentially private approaches in machine learning can be categorized according to the object it perturbs: one is to directly add noises to the training data [30], another is to add noises to the output of training at each iteration or at the end [6]. But many of these approaches are not designed for a distributed system where data are stored on local devices. Therefore, these privacy guarantees are achieved for the whole data set without including the personalized privacy concern. In this article, we make our distributed personalized learning process to be differentially private by perturbing the output of training at each iteration and achieve personalized differential privacy for each user under the consideration of device heterogeneity in the real world.

IX. Conclusion

In this article, we have studied the problem of learning multiple personalized classifiers collaboratively in a privacy-preserving manner. We have considered privacy in the (ϵ, δ)-differential privacy model and provided a privacy-preserving algorithm for the personalized federated learning. We bounded the privacy loss by exploiting the existing system uncertainty caused by the device heterogeneity. The proposed approach is robust to device heterogeneity and the perturbation of noises. We have evaluated our approach on real mobile sensing data, showing the impact of device heterogeneity and the tradeoff between privacy and accuracy.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat.*, 2017, pp. 1273–1282.

[2] J. Konecnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016. [Online]. Available: arXiv:1610.02527.

[3] M. Al-Rubaie and J. M. Chang, "Reconstruction attacks against mobile-based continuous authentication systems in the cloud," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 12, pp. 2648–2663, Dec. 2016.

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

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.

[5] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2016. [Online]. Available: arXiv:1610.05755.

[6] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 308–318.

[7] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. 31st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4427–4437.

[8] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," 2012. [Online]. Available: arXiv:1203.3536.

[9] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[10] C. Ma, V. Smith, M. Jaggi, M. I. Jordan, P. Richtárik, and M. Takáč, "Adding vs. averaging in distributed primal-dual optimization," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 1973–1982.

[11] V. Smith, S. Forte, C. Ma, M. Takáč, M. I. Jordan, and M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 8590–8638, 2017.

[12] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Proc. IEEE 51st Annu. Symp. Found. Comput. Sci.*, Las Vegas, NV, USA, 2010, pp. 51–60.

[13] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," 2016. [Online]. Available: arXiv:1603.01887.

[14] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, 2011.

[15] J. Reyes-Oritz, D. Anguita, A. Ghio, L. Oneto, and X. Parra, *UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set*, School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2012.

[16] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss," *J. Mach. Learn. Res.*, vol. 14, pp. 567–599, Feb. 2013.

[17] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016. [Online]. Available: arXiv:1610.05492.

[18] Y. Guo and Y. Gong, "Practical collaborative learning for crowdsensing in the Internet of Things with differential privacy," in *Proc. IEEE Conf. Commun. Netw. Security (CNS)*, Beijing, China, 2018, pp. 1–9.

[19] Y. Gong, Y. Fang, and Y. Guo, "Private data analytics on biomedical sensing data via distributed computation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 3, pp. 431–444, May/Jun. 2016.

[20] Y. Gong, Y. Fang, and Y. Guo, "Privacy-preserving collaborative learning for mobile health monitoring," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, 2015, pp. 1–6.

[21] R. Hu, Y. Gong, and Y. Guo, "CPFed: Communication-efficient and privacy-preserving federated learning," 2020. [Online]. Available: arXiv:2003.13761.

[22] R. Hu, Y. Guo, E. P. Ratazzi, and Y. Gong, "Differentially private federated learning for resource-constrained Internet of Things," 2020. [Online]. Available: arXiv:2003.12705.

[23] L. Jacob, J.-P. Vert, and F. R. Bach, "Clustered multi-task learning: A convex formulation," in *Proc. 21st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2009, pp. 745–752.

[24] S. Liu, S. J. Pan, and Q. Ho, "Distributed multi-task relationship learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discover. Data Mining*, 2017, pp. 937–946.

[25] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discover. Data Mining*, 2004, pp. 109–117.

[26] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. 19th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2007, pp. 41–48.

[27] J. Wang, M. Kolar, and N. Srebro, "Distributed multi-task learning with shared representation," 2016. [Online]. Available: arXiv:1603.02185.

[28] P. Vanhaesebrouck, A. Bellet, and M. Tommasi, "Decentralized collaborative learning of personalized models over networks," in *Proc. 20th Int. Conf. Artif. Intell. Stat.*, 2017, pp. 509–517.

[29] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Security*, 2015, pp. 1322–1333.

[30] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, "Analyze gauss: Optimal bounds for privacy-preserving principal component analysis," in *Proc. 46th Annu. ACM Symp. Theory Comput.*, 2014, pp. 11–20.

Rui Hu (Student Member, IEEE) received the B.Eng. degree in electrical engineering from Jinan University, Guangzhou, China, in 2017. She is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Texas at San Antonio, San Antonio, TX, USA.

Her research interest includes security and privacy in machine learning, social networks, and cyber–physical systems.

Dr. Hu received the Dorrough Distinguished Graduate Fellowship and the Graduate Student Professional Development Awards in 2018 and 2019, respectively.

Yuanxiong Guo (Senior Member, IEEE) received the B.Eng. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2009, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2012 and 2014, respectively.

He is an Assistant Professor with the Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX, USA. His research interests include data analytics, security, and privacy.

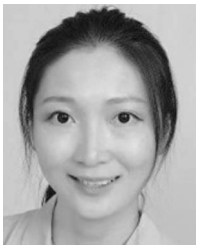
Hongning Li (Member, IEEE) received the B.S. degree from the School of Science, Xidian University, Xi'an, China, in 2007, the M.S. degree from the School of Telecommunications Engineering, Xidian University in 2010, and the Ph.D. degree from the School of Computer Science and Technology, Xidian University in 2014.

Her research interests include wireless networks and security, and security and privacy in cognitive radio networks.

Qingqi Pei (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and cryptography from Xidian University, Xi'an, China, in 1998, 2005, and 2008, respectively.

He is currently a Professor and a member of the State Key Laboratory of Integrated Services Networks, Xidian University. His research interests focus on digital contents protection and wireless networks and security.

Prof. Pei is a Professional Member of ACM and a Senior Member of the Chinese Institute of Electronics and China Computer Federation.



Yanmin Gong (Member, IEEE) received the B.Eng. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2012, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2016.

She is an Assistant Professor with the Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX, USA. Her research interests include security and privacy, wireless networks, and machine learning.

Authorized licensed use limited to: University of Texas at San Antonio. Downloaded on November 27,2020 at 21:03:36 UTC from IEEE Xplore. Restrictions apply.