Dynamic Differential Privacy for ADMM-Based Distributed Classification Learning

Tao Zhang, *Student Member, IEEE*, and Quanyan Zhu, *Member, IEEE*

***Abstract*— Privacy-preserving distributed machine learning becomes increasingly important due to the recent rapid growth of data. This paper focuses on a class of regularized empirical risk minimization machine learning problems, and develops two methods to provide differential privacy to distributed learning algorithms over a network. We first decentralize the learning algorithm using the alternating direction method of multipliers, and propose the methods of *dual variable perturbation* and *primal variable perturbation* to provide dynamic differential privacy. The two mechanisms lead to algorithms that can provide privacy guarantees under mild conditions of the convexity and differentiability of the loss function and the regularizer. We study the performance of the algorithms, and show that the dual variable perturbation outperforms its primal counterpart. To design an optimal privacy mechanism, we analyze the fundamental tradeoff between privacy and accuracy, and provide guidelines to choose privacy parameters. Numerical experiments using customer information database are performed to corroborate the results on privacy and utility tradeoffs and design.**

***Index Terms*— Machine learning, ADMM, distributed computing, privacy, differential privacy, dynamic programming.**

I. Introduction

DISTRIBUTED machine learning is a promising way to manage the deluge of data that has been witnessed recently. With the training data of size ranging from $1TB$ to $1PB$ [1], a centralized machine learning approach that collects and processes the data can lead to significant computational complexity and communications overhead. Therefore, a decentralized approach to machine learning is imperative to provide the scalability of the data processing and improve the quality of decision-making, while reducing the computational cost.

One suitable approach to decentralizing a centralized machine learning problem is *alternating direction method of multiplier* (ADMM). It enables distributed training over a network of collaborative nodes who exchange their results with the neighbors. However, the communications between two neighboring nodes create serious privacy concerns for nodes who process sensitive data including social network data, web search histories, financial information, and medical records. An adversary can observe the outcome of the learning and acquire sensitive information of the training data of individual nodes. The adversary can be either a member of the learning network who observes its neighbors or an outsider who observes the entire network. A privacy-preserving mechanism needs to automatically build into the distributed machine learning scheme to protect the internal and external adversaries throughout the entire dynamic learning process. Differential privacy is a suitable concept that provides a strong guarantee that the removal or addition of a single database item does not allow an adversary to distinguish (substantially) an individual data point [2].

In this work, we focus on a class of distributed ADMM-based *empirical risk minimization* (ERM) problems, and develop randomized algorithms that can provide differential privacy [2], [3] while keeping the learning procedure accurate. We extend the privacy concepts to dynamic differential privacy to capture the nature of distributed machine learning over networks, and propose two privacy-preserving schemes of the regularized ERM-based optimization. The first method is *dual variable perturbation* (DVP), in which we perturb the dual variable of each node at every ADMM iteration. The second is the *primal variable perturbation* (PVP) which leverages the *output perturbation* technique developed by Dwork et al. [2] by adding noise to the update process of the primal variable of each node of the ADMM-based distributed algorithm before sharing it to neighboring nodes.

We investigate the performance of the algorithms, and show that the DVP outperforms PVP. We characterize the fundamental tradeoffs between privacy and accuracy by formulating an optimization problem and use numerical experiments to demonstrate the optimal design of privacy mechanisms. The main contributions of the paper are summarized as follows:

(i) We use ADMM to decentralize regularized ERM algorithms to achieve distributed training of large datasets. Dynamic differential privacy is guaranteed for the distributed algorithm using the DVP, which adds noise to the update of the dual variable.

(ii) We develop PVP method to add noise to the primal variables when they are transmitted to neighboring nodes. This approach guarantees dynamic differential privacy in which privacy is preserved at each update.

(iii) We provide the theoretical performance guarantees of the PVP perturbations of the distributed ERM with l_2 regularization. The performance is measured by the number of sample data points required to achieve a certain criteria. Our theoretical results show that DVP is preferred for more difficult learning problems with a non-separable dataset or with small margin.

(iv) We propose a design principle to select the optimal privacy parameters by solving an optimization

Manuscript received March 22, 2016; revised July 9, 2016; accepted August 26, 2016. Date of publication September 8, 2016; date of current version November 3, 2016. This work was supported in part by the National Science Foundation under Grant CNS-1544782 and in part by the NYU Research Challenge Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Athanasios V. Vasilakos.

The authors are with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: tz636@nyu.edu; qz494@nyu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2016.2607691

problem. Numerical experiments show that the PVP outperforms the DVP at managing the privacy-accuracy tradeoff.

A. Related Work

There has been a significant amount of literature on the distributed classification learning algorithms. These works have mainly focused on either enhancing the efficiency of the learning model, or on producing a global classifier from multiple distributed local classifier trained at individual nodes. Researchers have focused on making the distributed algorithm suitable for large-scale datasets, e.g., MapReduce has been used to explore the performance improvements [4]. In addition, methods such as ADMM methods [5], voting classification [6], and mixing parameters [7] have been used to achieve distributed computation. Our approach to distributed machine learning is based on ADMM, in which the centralized problem acts as a group of coupled distributed convex optimization subproblems with the consensus constraints on the decision parameters over a network.

In privacy-preserving data mining research, the privacy can be pried through, for example, *composition attacks*, in which the adversary has some prior knowledge. Other works on data perturbation for privacy (e.g., [8], [9]) have focused on additive or multiplicative perturbations of individual samples, which might affect certain relationships among different samples in the database. A body of existing literature also have studied the differential-private machine learning. For example, Kasiviswanathan et al. have derived a general method for probabilistically approximately correct (PAC, [10]) in [11]. Many works have investigated the tradeoff privacy and accuracy while developing and exploring the theory of differential privacy (examples include [2], [12], and [13]). Also, an increasingly number of researches have studied the the cryptographic protocols for distributed differential privacy. Such works include multi-party distributed data aggregation [14] and private multi-party computations [15]. Eigner and Maffei developed the first framework for the automated verification of distributed differential privacy in [16]; they developed a system to enforce the distributed differential privacy in cryptographic protocol implementations. In this work, we extend the notion of differential privacy to a dynamic setting, and define dynamic differential privacy to capture the distributed and iterative nature of the ADMM-based distributed ERM.

B. Organization of the Paper

The rest of the paper is organized as follows. Section 2 presents the ADMM approach to decentralizing a centralized ERM problem, and describe the privacy concerns associated with the distributed machine learning. In Section 3, we present dual and primal variable perturbation algorithms to provide dynamic differential privacy. The analysis of privacy guarantee for the algorithms is discussed. Section 4 studies the performance of the privacy-preserving algorithms. Section 5 presents numerical experiments to corroborate the results and optimal design principles to the tradeoff between privacy and accuracy. Finally, Section 6 presents concluding remarks and future research directions. Table I lists the notations frequently used in this paper. Note that in Table I, $\star$ represents $dual$ or $prim$.

TABLE I

SUMMARY OF NOTATION

$Z_{C_1}(f\|\hat{D})$	Centralized regularized ERM (R-ERM)
$Z_p(f_p\|D_p)$	Distributed R-ERM
$L_p^N(t)$	Augmented Lagrange function (ALF) of Z_p
$L_p^\star(t)$	ALF of the DVP
$\hat{C}(f_p) := C^R \mathbb{E}_{(x,y)\sim\mathbb{P}^{xy}}(\mathscr{L}(yf^Tx))$	Expected loss
$\hat{Z}_p(f_p) := \hat{C}(f_p) + \rho R(f_p)$	Expected objective function (obj.)
$Z_p^\star(\cdot)$	Associated obj. of Alg-2,3 at t
$f_p^{non}(t+1) = \arg\min_{f_p} L_p^N(t)$	Non-private transcript of Algorithm 1
$f_p(t+1) = \arg\min_{f_p} L_p^\star(t)$	Private transcript of Algorithm 2/3
f^0	Non-private reference classifier
$f_p^0(t)$	Private reference classifier at time t
$\hat{f}_p(t+1) = \arg\min_{f_p} \hat{Z}_p(f_p)$,	Expected optimum
$f^* = \arg\min_{f_p} Z_p(f_p\|D_p)$	Empirical non-private optimum
$f_p^*(t) = \arg\min_{f_p} Z_p^\star(t)$	Optimum of Alg-2,3 at t

II. PROBLEM STATEMENT

Consider a connected network, which contains P nodes described by an undirected graph $G(\mathscr{P}, \mathscr{E})$ with the set of nodes $\mathscr{P} = \{1, 2, 3, \dots, P\}$, and a set of edges $\mathscr{E}$ denoting the links between connected nodes. A particular node $p \in \mathscr{P}$ only exchanges information between its neighboring node $j \in \mathscr{N}_p$, where $\mathscr{N}_p$ is the set of all neighboring nodes of node p, and $N_p = |\mathscr{N}_p|$ is the number of neighboring nodes of node p. Each node p contains a dataset $D_p = \{(x_{ip}, y_{ip}) \subset X \times Y : i = 0, 1, \dots, B_p\}$, which is of size B_p with data vector $x_{ip} \in X \subseteq \mathbb{R}^d$, and the corresponding label $y_{ip} \in Y := \{-1, 1\}$. The entire network therefore has a set of data $\hat{D} = \bigcup_{p\in\mathscr{P}} D_p$.

The target of the centralized classification algorithm is to find a classifier $f : X \to Y$ using all available data $\hat{D}$ that enables the entire network to classify any data x' input to a label $y' \in \{-1, 1\}$. Let $Z_{C_1}(f|\hat{D})$ be the objective function of a regularized empirical risk minimization problem (CR-ERM), defined as follows:

$$Z_{C_1}(f|\hat{D}) := \frac{C^R}{B_p} \sum_{p=1}^{P} \sum_{i=1}^{B_p} \hat{\mathscr{L}}(y_{ip},\ f^T x_{ip}) + \rho R(f), \qquad (1)$$

where $C^R \leq B_p$ is a regularization parameter, and $\rho > 0$ is the parameter that controls the impact of the regularizer. Suppose that $\hat{D}$ is available to the fusion center node, then we can choose the global classifier $f : X \to Y$ that minimizes the CR-ERM.

The *loss function* $\hat{\mathscr{L}}(y_{ip},\ f^T x_{ip}) : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, is used to measure the quality of the classifier trained. We focus on the specific loss function $\hat{\mathscr{L}}(y_{ip}, f^T x_{ip}) = \mathscr{L}(y_{ip} f^T x_{ip})$. The function $R(f)$ in (1) is a regularizer that prevents overfitting. We have the following assumptions on the loss, regularization functions, and the data.

Assumption 1: The loss function $\mathscr{L}$ is strictly convex and doubly differentiable of f with $|\mathscr{L}'| \leq 1$ and $|\mathscr{L}''| \leq c_1$, where c_1 is a constant. Both $\mathscr{L}$ and $\mathscr{L}'$ are continuous.

Assumption 2: The regularizer function $R(\cdot)$ is continuous differentiable and 1-strongly convex. Both $R(\cdot)$ and $\nabla R(\cdot)$ are continuous.

Assumption 3: We assume that $\|x_{ip}\| \leq 1$. Since $y_{ip} \in \{-1, 1\}$, $|y_{ip}| = 1$.

A. Distributed ERM

To decentralize CR-ERM, we introduce decision variables $\{f_p\}_{p=1}^P$, where node p determines its own classifier f_p, and impose consensus constraints $f_1 = f_2 = \ldots = f_P$ that guarantee global consistency of the classifiers. Let $\{w_{jp}\}$ be the auxiliary variables to decouple f_p of node p from its neighbors $j \in \mathcal{N}_p$. Then, the consensus-based reformulation of (1) becomes

$$\min_{\{f_p\}_{p=1}^P} \; Z_{C_2} := \frac{C^R}{B_p} \sum_{p=1}^{P} \sum_{i=1}^{B_p} \mathscr{L}(y_{ip} f_p^T x_{ip}) + \sum_{p=1}^{P} \rho R(f_p).$$
$$\text{s.t.} f_p = w_{pj},\; w_{pj} = f_j,\; p = 1, \ldots, P,\; j \in \mathcal{N}_p \tag{2}$$

where $Z_{C_2}(\{f_p\}_{p\in\mathscr{P}}|\hat{D})$ is the reformulated objective as a function of $\{f_p\}_{p=1}^P$. According to [5, Lemma 1], if $\{f_p\}_{p=1}^P$ presents a feasible solution of (2) and the network is connected, then problems (1) and (2) are equivalent, i.e., $f = f_p$, $p = 1, \ldots, P$, where f is a feasible solution of CR-ERM. Problem (2) can be solved in a distributed fashion using the alternative direction method of multiplier (ADMM) with each node $p \in \mathscr{P}$ optimizing the following distributed regularized empirical risk minimization problem (DR-ERM):

$$Z_p(f_p|D_p) := \frac{C^R}{B_p} \sum_{i=1}^{B_p} \mathscr{L}(y_{ip} f_p^T x_{ip}) + \rho R(f_p). \tag{3}$$

The augmented Lagrange function associated with the DR-ERM is:

$$\begin{aligned} &L_p^D(f_p, w_{pj}, \lambda_{pj}^k) \\ &= Z_p + \sum_{i\in\mathcal{N}_p} \left(\lambda_{pi}^a\right)^T (f_p - w_{pi}) + \sum_{i\in\mathcal{N}_p} \left(\lambda_{pi}^b\right)^T (w_{pi} - f_i) \\ &\quad + \frac{\eta}{2} \sum_{i\in\mathcal{N}_p} (\| f_p - w_{pi} \|^2 + \| w_{pi} - f_i \|^2). \end{aligned} \tag{4}$$

The distributed iterations solving (3) are:

$$f_p(t+1) = \arg\min_{f_p} L_p^D\big(f_p, w_{pj}(t), \lambda_{pj}^k(t)\big), \tag{5}$$
$$w_{pj}(t+1) = \arg\min_{w_{pj}} L_p^D\big(f_p(t+1), w_{pj}, \lambda_{pj}^k(t)\big), \tag{6}$$
$$\lambda_{pj}^a(t+1) = \lambda_{pj}^a(t) + \eta(f_p(t+1) - w_{pj}(t+1)), \quad p \in \mathscr{P}, j \in \mathcal{N}_p, \tag{7}$$
$$\lambda_{pj}^b(t+1) = \lambda_{pj}^b(t) + \eta(w_{pj}(t+1) - f_p(t+1)), \quad p \in \mathscr{P}, j \in \mathcal{N}_p. \tag{8}$$

According to [5, Lemma 2], iterations (5) to (8) can be further simplified by initializing the dual variables $\lambda_{pj}^k = \mathbf{0}_{d\times d}$, and letting $\lambda_p(t) = \sum_{j\in\mathcal{N}_p} \lambda_{pj}^k$, $p \in \mathscr{P}$, $j \in \mathcal{N}_p$, $k = a, b$, we can combine (7) and (8) into one update. Thus, we simplify (5)-(8) by introducing the following: Let $L_p^N(t)$ be the short-hand notation of $L_p^N(\{f_p\}, \{f_p(t)\}, \{\lambda_p(t)\})$ as:

$$\begin{aligned} L_p^N(t) :=\; &\frac{C^R}{B_p} \sum_{i=1}^{B_p} \mathscr{L}(y_{ip} f_p^T x_{ip}) + \rho R(f_p) + 2\lambda_p(t)^T f_p \\ &+ \eta \sum_{i\in\mathcal{N}_p} \| f_p - \frac{1}{2}(f_p(t) + f_i(t)) \|^2. \end{aligned} \tag{9}$$

Algorithm 1 Distributed ERM

```
Required: Randomly initialize $f_p, \lambda_p = \mathbf{0}_{d\times 1}$ for every $p \in \mathscr{P}$
Input: $\hat{D}$
for $t = 0, 1, 2, 3, \ldots$ do
  for $p = 0$ to $P$ do
    Compute $f_p(t+1)$ via (10).
  end for
  for $p = 0$ to $P$ do
    Broadcast $f_p(t+1)$ to all neighbors $j \in \mathcal{N}_p$.
  end for
  for $p = 0$ to $P$ do
    Compute $\lambda_p(t+1)$ via (11).
  end for
end for
Output: $f^*$.
```

The ADMM iterations (5)-(8) can be reduced to

$$f_p(t+1) = \arg\min_{f_p} L_p^N(f_p, f_p(t), \lambda_p(t)), \tag{10}$$
$$\lambda_p(t+1) = \lambda_p(t) + \frac{\eta}{2} \sum_{j\in\mathcal{N}_p} [f_p(t+1) - f_j(t+1)]. \tag{11}$$

The ADMM-based distributed ERM iterations (10)-(11) are and summarized in Algorithm 1. Every node $p \in \mathscr{P}$ updates its local $d \times 1$ estimates $f_p(t)$ and $\lambda_p(t)$. At iteration $t+1$, node p updates the local $f_p(t+1)$ through (10). Next, node p broadcasts the latest $f_p(t+1)$ to all its neighboring nodes $j \in \mathcal{N}_p$. Iteration $t+1$ finishes as each node updates the $\lambda_p(t+1)$ via (11).

Every iteration of our algorithm is still a minimization problem similar to the centralized problem (1). However, the number of variables participating in solving (10) per node per iteration is N_p, which is much smaller than the one in the centralized problem, which is $\sum_{p=1}^P N_p$. There are several methods to solve (10). For instance, projected gradient method, Newton method, and Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [17] that approximates the Newton method, to name a few.

ADMM-based distributed machine learning has benefits due to its high scalability. It also provides some degree of privacy since nodes do not communicate data directly but their decision variable f_p. However, the privacy arises when an adversary can make intelligent inferences at each step and extract the sensitive information based on his observation of the learning output of his neighboring nodes. Simple anonymization is not sufficient to address this issue as discussed in Section 1. In the following subsection, we will discuss the adversary models, and present differential privacy solutions.

B. Privacy Concerns

Although the data stored at each node is not exchanged during the entire ADMM algorithm, the potential privacy risk still exists. Suppose that the dataset D_p stored at node p contains sensitive information in data point (x_i, y_i) that is not allowed to be released to other nodes in the network or anyone else outside. Let $K : \mathbb{R}^d \to \mathbb{R}$ be the randomized version of Algorithm 1, and let $\{f_p^*\}_{p\in\mathscr{P}}$ be the output of K at all the nodes. Then, the output $\{f_p^*\}_{p\in\mathscr{P}}$ is random. In the distributed version of the algorithm, each node optimizes its local empirical risk based on its own dataset D_p. Let K_p^t be the node-p-dependent stochastic sub-algorithm of K at iteration t, and let $f_p(t)$ be the output of $K_p^t(D_p)$ at iteration t inputting D_p. Hence the output $f_p(t)$ is stochastic at each t. In this work, we consider the following attack model. The adversary can access the learning outputs of intermediate ADMM iterations as well as the final output. This type of adversary aims to obtain sensitive information about the private data point of the training dataset by observing the output $f_p(t)$ of K_p^t or f_p^* of K for all $p \in \mathscr{P}$ at every stage t of the training. We protect the privacy of distributed network using the definition of *differential privacy* in [2]. Specifically, we require that a change of any single data point in the dataset might only change the distribution of the output of the algorithm slightly, which is visible to the adversary; this is done by adding randomness to the output of the algorithm. Let D_p and D'_p be two datasets differing in one data point; i.e., let $(x_{ip}, y_{ip}) \subset D_p$, and $(x'_{ip}, y'_{ip}) \subset D'_p$, then $(x_{ip}, y_{ip}) \neq (x'_{ip}, y'_{ip})$. In other words, their *Hamming Distance*, which is defined as $H_d(D_p, D'_p) = \sum_{i=0}^{B_p} \mathbf{1}\{i : x_i \neq x'_i\}$, equals 1; i.e., $H_d(D_p, D'_p) = 1$.

To protect the privacy against the adversary, we propose the concept of dynamic differential privacy, which enables the dynamic algorithm to be privacy-preserving at every stage of the learning.

Definition 4 [Dynamic $\alpha(t)$-Differential Privacy ($\alpha(t)$-DDP)]: Consider a network of P nodes $\mathscr{P} = \{1, 2, \ldots, P\}$, and each node p has a training dataset D_p, and $\hat{D} = \bigcup_{p\in\mathscr{P}} D_p$. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a randomized version of Algorithm 1. Let $\alpha(t) = (\alpha_1(t), \alpha_2(t), \ldots, \alpha_P(t)) \in \mathbb{R}_+^P$, where $\alpha_p(t) \in \mathbb{R}_+$ is the privacy parameter of node p at iteration t. Let K_p^t be the node-p-dependent sub-algorithm of K, which corresponds to an ADMM iteration at t that outputs $f_p(t)$. Let D'_p be any dataset with $H_d(D'_p, D_p) = 1$, and $g_p(t) = K_p^t(D'_p)$. We say that the algorithm K is dynamic $\alpha_p(t)$-differentially private *(DDP) if for any dataset D'_p, and for all $p \in \mathscr{P}$ that can be observed by the adversaries, and for all possible sets of the outcomes $S \subseteq \mathbb{R}$, the following inequality holds:*

$$\Pr[f_p(t) \in S] \leq e^{\alpha_p(t)} \cdot \Pr[g_p(t) \in S], \tag{12}$$

for all $t \in \mathbb{Z}$ during a learning process. The probability is taken with respect to $f_p(t)$, the output of K_p^t at every stage t. The algorithm K is called dynamic $\alpha(t)$-differentially private *if the above conditions are satisfied.*

Definition 1 provides a suitable differential privacy concept for the adversary. For dynamic $\alpha_p(t)$-differentially private algorithms, the adversaries cannot extract additional information by observing the intermediate updates of $f_p(t)$ at each step. Clearly, the algorithm with ADMM iterations shown in (10) to (11) is not dynamic $\alpha_p(t)$-differentially private. This is because the intermediate and final optimal output f_p's are deterministic given dataset D_p. For D'_p with $H_d(D_p, D'_p) = 1$, the classifier will change completely, and the probability density $\Pr([f_p|D'_p]) = 0$, which leads to the ratio of probabilities $\frac{\Pr[f_p|D_p]}{\Pr[f_p|D'_p]} \to \infty$. In order to provide the DDP, we propose two algorithms, *dual variable perturbation* and *primal variable perturbation*, which are described in Section 3.1 and 3.2, respectively.

III. Dynamic Private Preserving

In this section, we introduce two dynamic-differential-privacy-preserving mechanisms: Dual Variable Perturbations (DVP) and Primal Variable Perturbations (PVP). Both mechanisms are shown to be $\alpha(t)$-DDP defined in Section II-B. by introducing appropriate noise on the iterative algorithms.

A. Dual Variable Perturbation

We protect the first algorithm based on *dual variable perturbation* (DVP), in which the dual variables $\{\lambda_p(t)\}_{p=1}^P$ are perturbed with a random noise vector $\epsilon_p(t) \in \mathbb{R}^d$ with the probability density function $\mathscr{K}_p(\epsilon) \sim e^{-\zeta_p(t)\|\epsilon\|}$, where $\zeta_p(t)$ is a parameter related to the value of $\alpha_p(t)$, and $\|\cdot\|$ denotes the l_2 norm. In this method, we add one additional term $\frac{\Phi}{2} \| f_p \|^2$ to the objective function (3) to ensure that the objective function associated with (13) is at least Φ-strongly convex. At each iteration, we first perturb the dual variable $\lambda_p(t)$, obtained from the last iteration, and store it in a new variable $\mu_p(t) = \lambda_p(t) + \epsilon_p(t)$. Now the corresponding node-p-based augmented Lagrange function $L_p^N(t)$ becomes $L_p^{dual}\big(f_p, f_p(t), \mu_p(t+1), \{f_i(t)\}_{i\in\mathscr{N}_p}\big)$, defined as follows, and $L_p^{dual}(t)$ is used as a short-hand notation:

$$\begin{aligned} L_p^{dual}(t) = Z_p(f_p|D_p) + \frac{\Phi}{2} \| f_p \|^2 + 2\mu_p(t+1)^T f_p \\ + \eta \sum_{i\in\mathscr{N}_p} \| f_p - \frac{1}{2}(f_p(t) + f_i(t)) \|^2 \end{aligned} \tag{13}$$

As a result, the minimizer of $L_p^{dual}(t)$ is random. At each iteration, we first perturb the dual variable $\lambda_p(t)$, obtained from the last iteration, and store it in a new variable $\mu_p(t+1)$.

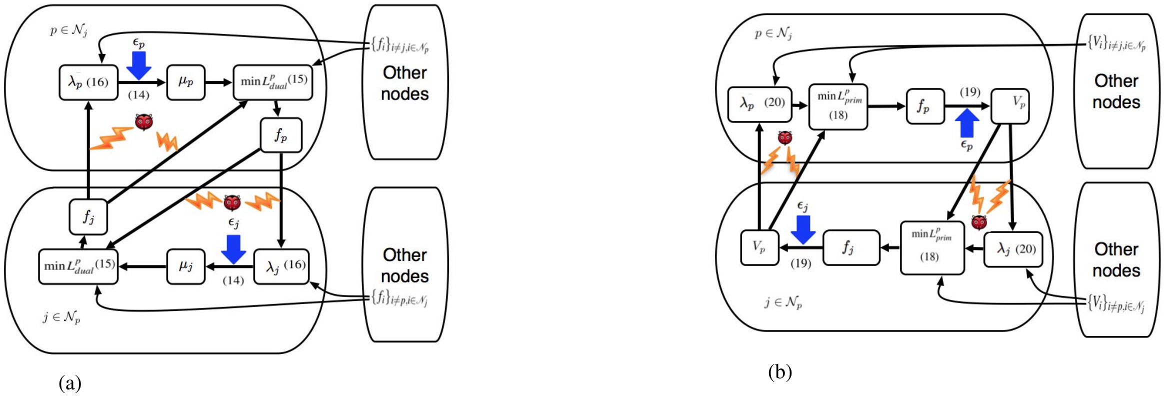
Now, the iterations (10)-(11) becomes follows:

$$\mu_p(t+1) = \lambda_p(t) + \frac{C^R}{2B_p}\epsilon_p(t+1), \tag{14}$$

$$f_p(t+1) = \arg\min_{f_p} L_p^{dual}(t), \tag{15}$$

$$\lambda_p(t+1) = \lambda_p(t) + \frac{\eta}{2} \sum_{j\in\mathscr{N}_p} [f_p(t+1) - f_j(t+1)]. \tag{16}$$

The iterations (14)-(16) are summarized as Algorithm 2, and are illustrated in Figure 1(a) and 2. The additional privacy parameter $\hat{\alpha}_p = \alpha_p(t) - \ln\left(1 + \frac{c_1}{\frac{B_p}{C^R}\left(\rho + 2\eta N_p\right)}\right)^2$ is required in the proof of Theorem 1 in Appendix A. The two cases of $\hat{\alpha}_p$ are considered to find the upper bound of the ratio of Jacobian matrices for the transformation from $f_p(t)$ to $\epsilon_p(t)$ given different datasets (see details in Appendix A). All nodes have its corresponding value of ρ. Every node $p \in \mathscr{P}$ updates its local estimates $\mu_p(t)$, $f_p(t)$ and $\lambda_p(t)$ at time t; at time $t+1$, node p first perturbs the dual variable $\lambda_p(t)$ obtained at time t to obtain $\mu_p(t+1)$ via (14), and then uses training dataset D_p to compute $f_p(t+1)$ via (15). Next, node p sends $f_p(t+1)$ to all its neighboring nodes. The $(t+1)$-th update



(a)

(b)

Fig. 1. Illustration of DVP and PVP: (a) DVP during intermediate iterations. The perturbed μ_p participates in the (15). As a result, the output f_p at each iteration is a random variable, and the transmission of f_p is differentially private. (b) PVP during intermediate iterations. The perturbed V_p is a random variable. As a result, the transmission of V_p is differentially private.

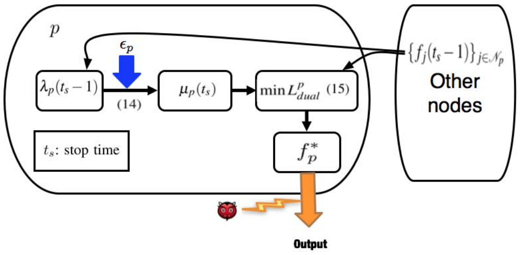


Fig. 2. The final iteration of both DVP and PVP. The perturbed μ_p participates in the (15). As a result, the output f^*_p is a random variable, and the final output is differentially private.

is done when each node updates its local $\lambda_p(t+1)$ via (16). We then have the following theorem.

Theorem 5: Under Assumption 1, 2 and 3, if the DR-ERM problem can be solved by Algorithm 2, then Algorithm 2 solving this distributed problem is dynamic α-differentially private with $\alpha_p(t)$ for each node $p \in \mathscr{P}$ at time t. Let $Q(f_p(t)|D)$ and $Q(f_p(t)|D'_p)$ be the probability density functions of $f_p(t)$ given dataset D and D'_p, respectively, with $H_d(D, D'_p) = 1$. The ratio of conditional probabilities of $f_p(t)$ is bounded as follows:

$$\frac{Q(f_p(t)|D)}{Q(f_p(t)|D'_p)} \leq e^{\alpha_p(t)}. \tag{17}$$

Proof: See Appendix A.

B. Primal Variable Perturbation

In this subsection, we provide the algorithm based on the *primal variable perturbation* (PVP), which perturbs the primal variable $\{f_p(t+1)\}_{p=0}^P$ before sending the decision to the neighboring nodes. This algorithm can also provide dynamic differential privacy defined in Definition 1. Let the node-p-based augmented Lagrange function $L_p^{prim}\big(f_p, f_p(t), \epsilon_p(t), \lambda_p(t), \{V_i(t)\}_{i \in \mathcal{N}_p}\big)$ be defined as follows, and use $L_p^{prim}(t)$ as its short hand notation:

$$L_p^{prim}(t) = \frac{C^R}{B_p}\sum_{i=1}^{B_p} \mathscr{L}(y_{ip} f_p^T x_{ip}) + \rho R(f_p) + 2\lambda_p(t)^T f_p$$
$$+\eta \sum_{i \in \mathcal{N}_p} \| f_p - \frac{1}{2}(f_p(t) + V_i(t) - \epsilon_p(t)) \|^2 .$$

Algorithm 2 Dual Variable Perturbation

```
Required: Randomly initialize $f_p, \lambda_p = \mathbf{0}_{d\times 1}$ for every $p \in \mathscr{P}$
Input: $\hat{D}$, $\{[\alpha_p(1), \alpha_p(2), ...]\}_{p=1}^P$
for $t = 0, 1, 2, 3, ...$ do
  for $p = 0$ to $P$ do
    Let $\hat{\alpha}_p = \alpha_p(t) - \ln\left(1 + \frac{c_1}{\frac{B_p}{C^R}(\rho + 2\eta N_p)}\right)^2$.
    if $\hat{\alpha}_p > 0$ then
      $\Phi = 0$.
    else
      $\Phi = \frac{c_1}{\frac{B_p}{C^R}(e^{\alpha_p(t)/4} - 1)} - \rho - 2\eta N_p$ and $\hat{\alpha}_p = \alpha_p(t)/2$.
    end if
    Draw noise $\epsilon_p(t)$ according to $\mathscr{K}_p(\epsilon) \sim e^{-\zeta_p(t)\|\epsilon\|}$ with $\zeta_p(t) = \hat{\alpha}_p$.
    Compute $\mu_p(t+1)$ via (14).
    Compute $f_p(t+1)$ via (15) with augmented Lagrange function as (13).
  end for
  for $p = 0$ to $P$ do
    Broadcast $f_p(t+1)$ to all neighbors $j \in \mathcal{N}_p$.
  end for
  for $p = 0$ to $P$ do
    Compute $\lambda_p(t+1)$ via (16).
  end for
end for
Output: $\{f^*_p\}_{p=1}^P$.
```

In this method, we divide the entire training process into two parts: (i) the intermediate iterations, and (ii) the final iteration. During the intermediate iterations, we use the unperturbed primal $f_p(t)$ obtained at time t in the augmented Lagrange function and subtract the noise vector $\epsilon_p(t)$ added at time t to reduce the noise in the minimization in (19). Note that the noise $\epsilon_p(t)$ at time t is known at time $t+1$. The privacy of releasing primal variable is not affected.

The corresponding ADMM iterations that can provide dynamic $\alpha_p(t)$-differential privacy at time t are as follows:

$$f_p(t+1) = \arg\min_{f_p} L_p^{prim}(t), \tag{18}$$

$$V_p(t+1) = f_p(t+1) + \epsilon_p(t+1), \tag{19}$$

Algorithm 3 Primal Variable Perturbation

```
Required: Randomly initialize $f_p, \lambda_p = \mathbf{0}_{d\times 1}$ for every $p \in \mathscr{P}$
Input: $\hat{D}$, $\{[\alpha_p(1), \alpha_p(2), ...]\}_{p=1}^P$.
for $t = 0, 1, 2, 3, ...$ do
  for $p = 0$ to $P$ do
    Draw noise $\epsilon_p(t)$ according to $\mathscr{K}_p(\epsilon) \sim e^{-\zeta_p(t)\|\epsilon\|}$ with $\zeta_p(t) = \frac{\rho B_p \alpha_p(t)}{2C^R}$.
    Compute $f_p(t+1)$ via (18) with augmented Lagrange function as (9).
    Compute $V_p(t+1)$ via (19).
  end for
  for $p = 0$ to $P$ do
    Broadcast $f_p(t+1)$ to all neighbors $j \in \mathscr{N}_p$.
  end for
  for $p = 0$ to $P$ do
    Compute $\lambda_p(t+1)$ via (20).
    if $t =$ stop time then
      Use the latest $\{f_p(t)\}_p$ and $\{\lambda_p(t)\}_p$ obtained as initial values, and input $\hat{D}$ to Algorithm 2 to iterate the loop once.
    end if
  end for
end for
Output: $\{f_p^*\}_{p=1}^P$.
```

$$\lambda_p(t+1) = \lambda_p(t) + \frac{\eta}{2}\sum_{j\in\mathscr{N}_p}[V_p(t+1) - V_j(t+1)], \qquad (20)$$

where $\epsilon_p(t+1)$ is the random noise vector with the density function $\mathscr{K}_p(\epsilon) \sim e^{-\zeta_p(t)\|\epsilon\|}$. The augmented Lagrange function is (9). Let t_s be the time when we enter the final iteration. When $t = t_s$ we enter the final iteration at t_s, we apply the DVP to update the variables. Specifically, we input the data sets $\hat{D}$ to DVP and use the $\{f_p(t_s-1)\}_p$ and $\{\lambda_p(t_s-1)\}_p$, obtained from (18) and (20), in iteration (14)-(16):

$$\mu_p(t_s+1) = \lambda_p(t_s) + \frac{C^R}{2B_p}\epsilon_p(t_s+1), \qquad (21)$$

$$f_p(t+1) = \arg\min_{f_p} L_p^{dual}(t_s), \qquad (22)$$

$$\lambda_p(t_s+1) = \lambda_p(t_s) + \frac{\eta}{2}\sum_{j\in\mathscr{N}_p}[f_p(t_s+1) - f_j(t_s+1)]. \qquad (23)$$

$\{f_p(t_s+1)\}_{p\in\mathscr{P}}$ is the final output of the PVP algorithm.

The iterations (18)-(20) and (21)-(23) are summarized in Algorithm 3, and are illustrated in Figure 1(b) and 2. The parameter $\zeta_p(t) = \frac{\rho B_p \alpha_p(t)}{2C^R}$ is chosen to show that Algorithm 3 is dynamic $\alpha(t)$-differentially private (see details in Appendix B). Each node $p \in \mathscr{P}$ updates $f_p(t)$, $V_p(t)$ and $\lambda_p(t)$ at time t. Then, at time $t+1$, the training dataset is used to compute $f_p(t+1)$ via (18), which is then perturbed to obtain $V_p(t+1)$ via (19). Next, $V_p(t+1)$ is distributed to all the neighboring nodes of node p. Finally, $\lambda_p(t+1)$ is updated via (20). The final iteration follows the DVP. We then have the following theorem.

Theorem 6: Under Assumption 1, 2 and 3, if the DR-ERM problem can be solved by Algorithm 3, then Algorithm 3 solving this distributed problem is dynamic $\alpha(t)$-differentially private. The ratio of conditional probabilities of $f_p(t)$ is bounded as in (17).

Proof: See Appendix B.

IV. Performance Analysis

In this section, we discuss the performance of Algorithm 2 and 3. We establish performance bounds for regularization functions with l_2 norm. Our analysis is based on the following assumptions:

Assumption 7: The data points $\{(x_{pi}, y_{pi})\}_{i=1}^{B_p}$ are drawn i.i.d. from a fixed but unknown probability distribution $\mathbb{P}^{xy}(x_{pi}, y_{pi})$ at each node $p \in \mathscr{P}$.

Assumption 8: $\epsilon_p(t)$ is drawn from $\mathscr{K}_p(\epsilon) \sim e^{-\zeta_p(t)\|\epsilon\|}$, with the same $\alpha_p(t) = \alpha(t)$ (thus the same $\zeta_p(t)$) for all $p \in \mathscr{P}$ at time $t \in \mathbb{Z}$.

We then define the expected loss of node p using classifier f_p as follows, under Assumption 4: $\hat{C}(f_p) := C^R\mathbb{E}_{(x,y)\sim\mathbb{P}^{xy}}(\mathscr{L}(yf^Tx))$, and the corresponding expected objective function $\hat{Z}$ is: $\hat{Z}_p(f_p) := \hat{C}(f_p) + \rho R(f_p)$. The performance of non-private non-distributed ERM classification learning has been already studied by, for example, Shalev et al. in [18] (also see the work of Chaudhuri et al. in [19]), which introduces a reference classifier f^0 with expected loss $\hat{C}(f^0)$, and shows that if the number of data points is sufficiently large, then the actual expected loss of the trained l_2 regularized support vector machine (SVM) classifier f_{SVM} satisfies $\hat{C}(f_{SVM}) \leq \hat{C}^0 + \alpha_{acc}$, where α_{acc} is the generalization error. We use a similar argument to study the accuracy of Algorithm 1. Let f^0 be the reference classifier of Algorithm 1. We quantify the performance of our algorithms with f^* as the final output by the number of data points required to obtain $\hat{C}(f^*) \leq \hat{C}^0 + \alpha_{acc}$.

However, instead of focusing on only the final output, we care about the learning performance at all iterations. Let $f_p^{non}(t+1) = \arg\min_{f_p} L_p^N(t)$ be the intermediate updated classifier at t, and let $f^* = \arg\min_{f_p} Z_p(f_p|D_p)$ be the final output of Algorithm 1. From Theorem 9 (see Appendix A), the sequence $\{f_p^{non}(t)\}$ is bounded and converges to the optimal value f^* as time $t \to \infty$. Note that $\{f_p^{non}(t)\}$ is a non-private classifier without added perturbations. Since the optimization is minimization, then there exists a constant $\Delta^{non}(t)$ at time t such that: $\hat{C}(f_p^{non}(t)) - \hat{C}(f^*) \leq \Delta^{non}(t)$, and substituting it to $\hat{C}(f^*) \leq \hat{C}^0 + \alpha_{acc}$, yields:

$$\hat{C}(f_p^{non}(t)) \leq \hat{C}^0 + \Delta^{non}(t) + \alpha_{acc}. \qquad (24)$$

Clearly, the above condition depends on the reference classifier f^0; actually, as shown later in this section, the number of data points depends on the l_2-norm $\| f^0 \|$ of the reference classifier. Usually, the reference classifier is chosen with an upper bound on $\| f^0 \|$, say b^0. Based on (24), we provide the following theorem about the performance of Algorithm 1.

Theorem 9: Let $R(f_p(t)) = \frac{1}{2}\| f_p(t) \|^2$, and let f^0 such that $\hat{C}(f^0) = \hat{C}^0$ for all $p \in \mathscr{P}$ at time t, and $\delta > 0$ is a positive real number. Let $f_p^{non}(t+1) = \arg\min_{f_p} L_p^N(f_p, t|D_p)$

be the output of Algorithm 1. If Assumption 1 and 4 are satisfied, then there exists a constant β_{non} such that if the number of data points, B_p in $D_p = \Big\{(x_{ip}, y_{ip}) \subset \mathbb{R}^d \times \{-1, 1\}\Big\}$, $B_p > \beta_{non}\left(\frac{C^R\|f^0\|^2\ln(\frac{1}{\delta})}{\alpha_{acc}^2}\right)$, then $f_p^{non}(t+1)$ satisfies $\mathbb{P}\big(\hat{C}(f_p^{non}(t+1)) \leq \hat{C}^0 + \alpha_{acc} + \Delta^{non}(t)\big) \geq 1 - \delta$. for all $t \in \mathbb{Z}_+$.

Proof: See Appendix C.

Note that $\alpha_{acc} \leq 1$ is required for most machine learning algorithms. In the case of SVM, if the constraints are $y_i f^T x_i \leq c_{SVM}$, for $i = 1, \ldots, n$, where n is the number of data points, then, classification margin is $c_{svm}/ \parallel f \parallel$. Thus, if we want to maximization the margin $c_{SVM}/ \parallel f^0 \parallel$ we need to choose large value of $\parallel f^0 \parallel$. Larger value of $\parallel f^0 \parallel$ is usually chosen for non-separable or with small margin. In the following section, we provide the performance guarantees of Algorithm 2 and 3.

A. Performance of Private Algorithms

Similar to Algorithm 1, we solve an optimization problem minimizing $L_p^{dual}(f_p, t|D_p)$ at each iteration. Let $f_p(t)$ and $\lambda_p(t)$ be the primal and dual variables used in minimizing $L_p^{dual}(f_p, t|D_p)$ at iteration t, respectively. Suppose that starting from iteration t, the noise vector is static with $\epsilon_p(t)$ generated at iteration t. Let $\epsilon_p^t = \epsilon_p(t)$ be the (static) value of the noise generated in Algorithm 2 at time t. To compare our private classifier at iteration t with a private reference classifier $f^0(t)$, we construct a corresponding algorithm, Alg-2, associated with Algorithm 2. However, starting from iteration $t+1$, the noise vector in Alg-2, $\epsilon_p(t') = \epsilon_p^t$ for all $t' > t$. In other words, solving Alg-2 is equivalent to solving the optimization problem with the objective function $Z_p^{dual}(f_p, t|D_p, \epsilon_p^t)$, $t \geq 0$ defined as follows:

$$Z_p^{dual}(f_p, t|D_p, \epsilon_p^t) := Z_p(f_p|D_p) + \frac{C^R}{B_p}\epsilon_p^t f_p.$$

Let $f_p'(t)$ and $\lambda'_p(t)$ be the updated variables of the ADMM-based algorithm minimizing $Z_p^{dual}(f_p, t|D_p)$ at iteration t. Then, Alg-2 can be interpreted as minimizing $Z_p^{dual}(f_p, t|D_p, \epsilon_p^t)$ with initial condition as $f_p'(0) = f_p(t)$ and $\lambda'_p(0) = \lambda_p(t)$ for all $p \in \mathscr{P}$. Let $Z_p^{dual}(f_p, t|D_p, \epsilon_p^t)$ be regarded as the associated objective function of Alg-2.

For PVP, we can also introduce a similar algorithm denoted as Alg-3. Let $\epsilon_p^{t2} = \epsilon_p(t) - \epsilon_i(t)$ be the (static) value of $\epsilon_p(t) - \epsilon_i(t)$ generated at t, for $i \in \mathscr{N}_p$. Then, the associated objective function of Alg-3 denoted by $Z_p^{prim}(f_p, t|D_p, \epsilon_p^{t2})$, $t \geq 0$, is defined as follows:

$$\begin{aligned} Z_p^{prim}(f_p, t|D_p, \epsilon_p^{t2}) &:= Z_p(f_p|D_p) \\ &-\eta \sum_{i \in \mathscr{N}_p} \Big((f_p - \frac{1}{2}(f_p(t) + f_i(t))^T \cdot (\epsilon_p^{t2}) + \frac{1}{4}(\epsilon_p^{t2})^2\Big). \end{aligned}$$

Let $Z_p^{priv}(t)$ represent $Z_p^{dual}(f_p, t|D_p, \epsilon_p^t)$ or $Z_p^{prim}(f_p, t|D_p, \epsilon_p^{t2})$. Since $Z_p^{priv}(t)$ is real and convex, then, similar to Algorithm 1, the sequence $\{f_p(t)\}$ is bounded and $f_p(t)$ converges to $f_p^*(t) = \arg\min_{f_p} Z_p^{priv}(t)$, which is a limit point of $f_p(t)$. Thus, there exists a constant $\Delta_p^{priv}(t) = \Delta_p^{dual}(t)$ or $\Delta_p^{prim}(t)$ given noise vector $\epsilon_p(t)$ such that $\hat{C}(f_p(t)) - \hat{C}(f_p^*(t)) \leq \Delta_p^{priv}(t)$. The performance analysis in Theorem 3 can also be used in DVP and PVP. Specifically, the performance is measured by the number of data points, B_p, for all $p \in \mathscr{P}$ required to obtain $\hat{C}(f_p(t)) \leq \hat{C}^0(t) + \alpha_{acc} + \Delta_p^{priv}(t)$. We say that every learned $f_p(t)$ is α_{acc}-optimal if it satisfies the above inequality.

Since in Alg-3, the perturbed primal variable $V_p(t')$ is equal to $f_p(t')$ plus a constant $\epsilon_p(t)$ generated by Algorithm 3 at iteration t, for $t' \geq 0$, we can find a constant $\Delta_p^{primV}(t)$ such that $\hat{C}(V_p(t)) - \hat{C}(V_p^*(t)) \leq \Delta_p^{primV}(t)$. Similarly, we measure the performance of V_p by the number of data points, B_p, for all $p \in \mathscr{P}$ required to achieve $\hat{C}(V_p(t)) \leq \hat{C}^0(t) + \alpha_{acc} + \Delta_p^{primV}(t)$, where $\hat{C}^0(t) = \hat{C}(f^0(t))$, and $f^0(t)$ is a reference classifier.

We now establish the performance bounds for Algorithm 2, DVP, which is summarized in the following theorem.

Theorem 10: Let $R(f_p(t)) = \frac{1}{2} \parallel f_p(t) \parallel^2$, and $f_p^0(t)$ such that $\hat{C}(f_p^0(t)) = \hat{C}^0(t)$ for all $p \in \mathscr{P}$, and a real number $\delta > 0$. If Assumption 1, 4 and 5 are satisfied, then there exists a constant β_{dual} such that if the number of data points, B_p in $D_p = \Big\{(x_{ip}, y_{ip}) \subset \mathbb{R}^d \times \{-1, 1\}\Big\}$,

$$\begin{aligned} B_p > \beta_{dual} \quad &\max\Bigg(\max_t\Big(\frac{\parallel f_p^0(t+1) \parallel d\ln(\frac{d}{\delta})}{\alpha_{acc}\alpha_p(t)}\Big), \\ &\max_t\Big(\frac{C^R c_1 \parallel f_p^0(t+1) \parallel^2}{\alpha_{acc}\alpha_p(t)}\Big), \\ &\max_t\Big(\frac{C^R \parallel f_p^0(t+1) \parallel^2 \ln(\frac{1}{\delta})}{\alpha_{acc}^2}\Big)\Bigg), \end{aligned}$$

then $f_p^*(t+1)$ satisfies $\mathbb{P}\big(\hat{C}(f_p^*(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc}\big) \geq 1 - 2\delta$.

Proof: See Appendix D.

Corollary 11: Let $f_p(t+1) = \arg\min_{f_p} L_p^{dual}(f_p, t|D_p)$ be the updated classifier of Algorithm 2 and let $f_p^0(t)$ be a reference classifier such that $\hat{C}(f_p^0(t) = \hat{C}^0(t)$. If all the conditions of Theorem 3 are satisfied, then $f_p(t+1)$ satisfies

$$\mathbb{P}\big(\hat{C}(f_p(t+1)) \leq \hat{C}^0(t) + \alpha_{acc} + \Delta_p^{dual}(t)\big) \geq 1 - 2\delta. \tag{25}$$

Proof: $\hat{C}(f_p(t)) - \hat{C}(f_p^*(t)) \leq \Delta_p^{dual}(t)$. holds for $f_p(t)$ and $f_p^*(t)$ and from Theorem 3, $\mathbb{P}\big(\hat{C}(f_p^*(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc}\big) \geq 1 - 2\delta$. Therefore, we can have (25). □

Theorem 4 and Corollary 4.1 can guarantee the privacy defined in both Definition 1. The following theorem is used to analyze the performance bound of classifier $f_p(t+1)$ in (18), which minimizes $L_p^{prim}(t)$ that involves noise vectors from $V_p(t)$ perturbed at the previous iteration.

Theorem 12: Let $R(f_p(t)) = \frac{1}{2} \parallel f_p(t) \parallel^2$, and $f_p^0(t)$ such that $\hat{C}(f_p^0(t)) = \hat{C}^0(t)$, and a real number $\delta > 0$. From Assumption 1, we have the loss function $\mathscr{L}(\cdot)$ is convex and differentiable with $\mathscr{L}'(\cdot) \leq 1$. If Assumption 4 and 5 are satisfied, then there exists a constant β_{prim}^A such that *if the number of data points, B_p in $D_p = \Big\{(x_{ip}, y_{ip}) \subset \mathbb{R}^d \times \{-1, 1\}\Big\}$,*

$$B_p > \beta^A_{prim} \max\Bigg(\max_t \Big(\frac{C^R \parallel f^0_p(t+1) \parallel^3 \eta N_p d \ln(\frac{d}{\delta})}{\alpha^2_{acc}\alpha_p(t)}\Big), \max_t \Big(\frac{C^R \parallel f^0_p(t+1) \parallel^2 \ln(\frac{1}{\delta})}{\alpha^2_{acc}}\Big)\Bigg),$$

then $f^*_p(t+1)$ satisfies $\mathbb{P}\big(\hat{C}(f^*_p(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc}\big) \geq 1 - 2\delta$.

Proof: See Appendix E.

Next, we establish the PVP performance bound of Algorithm 3. Theorem 6 and Corollary 6.1 shows the requirements under which the performance of the part 1 of Algorithm 3 is guaranteed. Corollary 6.2 combines the results from Theorem 5 and Corollary 6.2 to provide the performance bound of the part 2 of Algorithm 3.

Theorem 13: Let $R(f_p(t)) = \frac{1}{2} \parallel f_p(t) \parallel^2$, and $f^0_p(t)$ such that $\hat{C}(f^0_p(t)) = \hat{C}^0(t)$, and $\delta > 0$ is a positive real number. Let $f^*_p(t+1) = \arg\min_{f_p} Z^{prim}_p(t)$ be α_{acc}-accurate according to Theorem 4. In addition to Assumption 1, we also assume that $\mathscr{L}'$ satisfies: $|\mathscr{L}'(a) - \mathscr{L}'(b)| \leq c_4|a-b|$ for all pairs (a, b) with a constant c_4. If Assumption 4 and 5 are satisfied, then there exists a constant β^B_{prim} such that if the number of data points, B_p in $D_p = \Big\{(x_{ip}, y_{ip}) \subset \mathbb{R}^d \times \{-1, 1\}\Big\}$,

$$\begin{aligned} B_p > \beta^B_{prim} \max\Bigg(&\max_t \Big(\frac{C^R \parallel f^0_p(t+1) \parallel^3 \eta N_p d \ln(\frac{d}{\delta})}{\alpha^2_{acc}\alpha_p(t)}\Big), \\ &\max_t \Big(\frac{C^R \parallel f^0_p(t+1) \parallel^2 \ln(\frac{1}{\delta})}{\alpha^2_{acc}}\Big), \\ &\max_t \Big(\frac{4C^B \parallel f^0(t+1) \parallel d\big(\ln(\frac{d}{\delta})\big)^2}{\alpha_{acc}\alpha_p(t)}\Big), \\ &\max_t \Big(\frac{4 \parallel f^0_p(t+1) \parallel^3 \eta N_p d \ln(\frac{d}{\delta})}{\alpha^2_{acc}\alpha_p(t)}\Big), \\ &\max_t \Big(\frac{4\big(C^R\big)^{\frac{3}{2}} \parallel f^0_p(t+1) \parallel^2 d \ln(\frac{d}{\delta})}{\alpha^{3/2}_{acc}\alpha_p(t)}\Big)\Bigg), \end{aligned} \quad (26)$$

then $V^*_p(t+1) = f^*_p(t+1) + \epsilon_p(t+1)$ satisfies $\mathbb{P}\big(\hat{C}(V^*_p(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc}\big) \geq 1 - 3\delta$.

Proof: See Appendix F.

Corollary 14: Let $f_p(t+1) = \arg\min_{f_p} L^{prim}_p(f_p, t|D_p)$ be the updated classifier of Algorithm 3, and let $f^0_p(t)$ be a reference classifier such that $\hat{C}(f^0_p(t) = \hat{C}^0(t)$. If all the conditions of Theorem 5 are satisfied, then, $V_p(t+1) = f_p(t+1) + \epsilon_p(t+1)$ satisfies

$$\mathbb{P}\big(\hat{C}(V_p(t+1)) \leq \hat{C}^0(t) + \alpha_{acc} + \Delta^{primV}_p(t)\big) \geq 1 - 3\delta. \quad (27)$$

Proof: From Theorem 6, $V^*_p(t+1)$ satisfies $\mathbb{P}\big(\hat{C}(V^*_p(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc}\big) \geq 1 - 3\delta$, and since $\hat{C}(V_p(t+1)) - \hat{C}(V^*_p(t+1)) \leq \Delta^{primV}_p(t+1)$, then, we have (27). □

Corollary 15: Let f^*_p be the final output classifier of Algorithm 3 at node p, and let $f^0_p(t)$ be a reference classifier such that $\hat{C}(f^0_p(t) = \hat{C}^0(t)$. If all the conditions of Theorem 4 and 6 are satisfied, then, f^*_p satisfies

$$\mathbb{P}\big(\hat{C}(f^*_p) \leq \hat{C}^0(t) + \alpha_{acc} + \Delta^{dual}_p(t)\big) \geq 1 - 5\delta.$$

Proof: All the conditions of Theorem 6 are satisfied to guarantee the privacy during the intermediate iterations. All the conditions of Theorem 4 are satisfied so that the final update is differential privacy is provided. Combining Theorem 4 and 6 yields the results. □

From Theorem 4 and 6, we can see that, for non-separable problems or ones with a small margin, in which a larger $\parallel f^0_p(t) \parallel$ is used, the terms $\frac{1}{\alpha_{acc}}$ and $\parallel f^0_p(t) \parallel$ have a more significant influence on the requirement of datasets size for DVP than the PVP. Also, the performance of DVP is guaranteed with higher probability than PVP. Therefore, DVP is preferred for more difficult problems. Moreover, the privacy increases by trading the accuracy. It is essential to manage the tradeoff between the privacy and the accuracy, and this will be discussed in Section 5.

V. Numerical Experiment

In this section, we test Algorithm 2 and 3 with real world training dataset. The dataset used is the *Adult* dataset from UCI Machine Learning Repository [20], which contains demographic information such as age, sex, education, occupation, marital status, and native country. In the following sections, we first introduce the logistic regression and show that it is suitable for our algorithms. Second, we test the convergence of the algorithms. Then, the tradeoff between privacy and accuracy is studied. We also propose a method to select the optimum value of $\alpha_p(t)$ that balance the privacy-accuracy tradeoff by introducing a utility function of privacy. Note that in the experiments, we fix the value of $\alpha_p(t)$ (thus fix the density of noise) for each complete running of algorithms and draw the independent and identically distributed (i.i.d.) noise to all the nodes in the network at each iteration.

A. Logistic Regression

In the experiments, we use our algorithm to develop a dynamic differential private logisitic regression. The logistic regression, i.e., $\mathscr{L}_{LR}$ takes the following form:

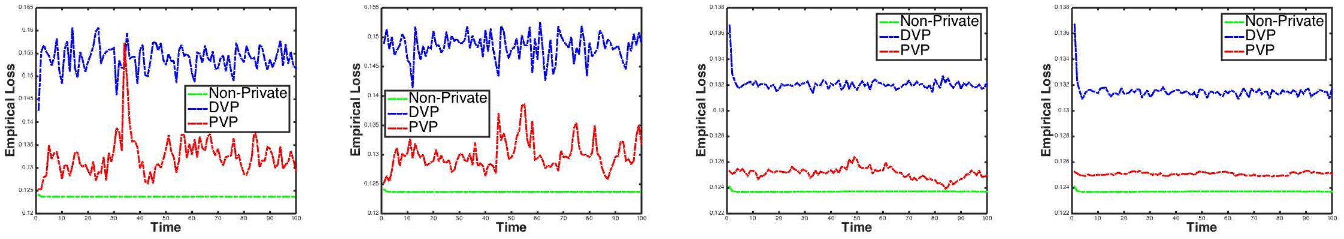
$$\mathscr{L}_{LR}(y_{ip} f^T x_{ip}) = log(1 + exp(-y_{ip} f^T_p x_{ip})).$$

The first order derivative and the second order derivative are:

$$\mathscr{L}'_{LR} = \frac{-y_{ip}x_{ip}}{1 + exp(y_{ip} f^T_p x_{ip})}$$

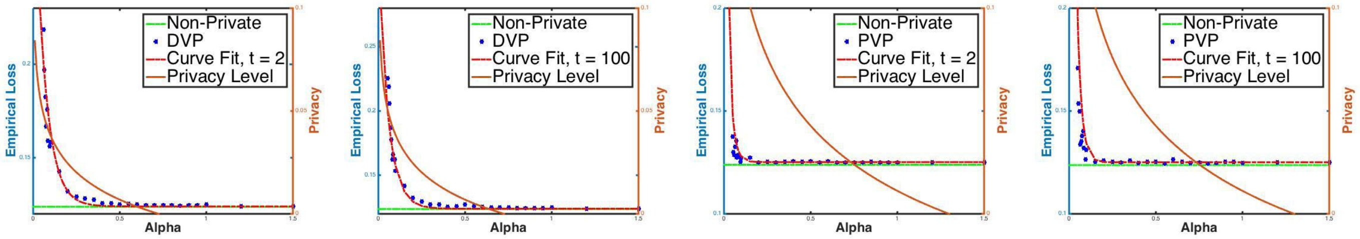
$$\mathscr{L}''_{LR} = \frac{y^2_{ip} x_{ip} x^T_{ip}}{(1 + exp(y_{ip} f^T_p x_{ip})(1 + exp(-y_{ip} f^T_p x_{ip})},$$

which can be bounded as $|\mathscr{L}'_{LR}| \leq 1$ and $|\mathscr{L}''_{LR}| \leq \frac{1}{4}$, respectively. Therefore, the loss function of logistic regression satisfies the conditions shown in Assumption 1. In this case, $R(F_p) = \frac{1}{2} \parallel f_p \parallel^2$, and $c_1 = \frac{1}{4}$. We can directly apply the loss function $\mathscr{L}_{LR}$ to Theorem 1 and 2 with $R(f) = \frac{1}{2} \parallel f_p \parallel^2$, and $c_1 = \frac{1}{4}$, and then it can provide dynamic $\alpha_p(t)$-differential privacy for all $t \in \mathbb{Z}$.



(a) (b) (c) (d)

Fig. 3. Convergence of algorithms, at iteration $t = 100$ (before the stop time) with different values of $\alpha_p(t)$. DVP with $\rho = 10^{-2.5}$ and $C^R = 1750$; PVP with $\rho = 10^{-1}$ and $C^R = 146$; Algorithm 1 (non-private) with $\rho = 10^{-10}$ and $C^R = 1750$. (a) $\alpha_p(t) = 0.01$. (b) $\alpha_p(t) = 0.1$. (c) $\alpha_p(t) = 0.5$. (d) $\alpha_p(t) = 1$.



(a) DVP: $t = 2$ (b) DVP: $t = 100$ (c) PVP: $t = 2$ (d) PVP: $t = 100$

Fig. 4. Privacy-accuracy tradeoff. (a)-(b): DVP, with $\omega_{p1} = 0.02$, $\omega_{p2} = 6$, $\omega_{p3} = 9$, $\omega_{p4} = 1$ (before the stop time); (c)-(d) PVP with $\omega_{p1} = 0.02$, $\omega_{p2} = 6$, $\omega_{p3} = 9$, $\omega_{p4} = 1$ (before the stop time).

B. Convergence

In our first set of experiments, we study the convergence of the algorithms. The convergence is tested by fixing the value of $\alpha_p(t)$ at node p for the entire running of algorithm. As shown in Figure 3, a larger $\alpha_p(t)$ leads to faster convergence of the algorithms; moreover, from Figure 3, we can see that the DVP is slightly more robust to noise than is the primal case given the same value of $\alpha_p(t)$. The empirical loss represents the accuracy of the classifier. Smaller empirical loss means higher accuracy. As can be seen, when $\alpha_p(t)$ is small, the model is more private but less accurate. Indeed, there is a tradeoff between privacy and accuracy, which will be studied in the next section.

C. Privacy-Accuracy Tradeoff

In this section, we study the privacy-accuracy tradeoff of Algorithm 2 and 3. The privacy is quantified by the value of $\alpha_p(t)$. A larger $\alpha_p(t)$ implies that the ratio of the densities of the classifier $f_p(t)$ on two different data sets is larger, which implies a higher belief of the adversary when one data point in dataset D is changed; thus, it provides lower privacy. However, the accuracy of the algorithm increases as $\alpha_p(t)$ becomes larger. Therefore, there is a decreasing monotonicity between privacy and accuracy.

We proposed a method to find an optimal value of $\alpha_p(t)$ that can balance the tradeoff between privacy and accuracy by constructing utility functions of privacy and accuracy, which need to satisfy the following assumptions:

Assumption 16: The utilities of privacy is monotonically decreasing with respect to $\alpha_p(t)$ for every $p \in \mathscr{P}$ and accuracy is monotonically increasing with respect to $\alpha_p(t)$ for every $p \in \mathscr{P}$.

In this experiment, the accuracy is measured by the empirical loss

$$\overline{C}(t) = \frac{C^R}{B_p}\sum_{i=1}^{B_p}\mathscr{L}(y_{ip}f_p(t)^T x_{ip}).$$

Let $L_{acc}(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ represent the relationship between $\alpha_p(t)$ and $\overline{C}(t)$. The function L_{acc} is obtained by curve fitting given the experimental data points $(\alpha_p(t), \overline{C}(t))$. Thus, we model the utility function by L_{acc} in this experiment. As shown in Figure 4, L_{acc} is monotonically increasing with respect to $\alpha_p(t)$ since smaller empirical loss represents higher accuracy. We then construct the utility function of privacy. The utility function of privacy is designed to meet specific requirements of privacy of the users of the algorithm. Let $U_{priv}(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ be the utility of privacy, same for every node $p \in \mathscr{P}$. Besides the decreasing monotonicity, $U_{priv}(\cdot)$ is assumed to be convex and doubly differentiable function of $\alpha_p(t)$. In our experiment, we model the utility of privacy as: $U_{priv}(\alpha_p(t)) = \omega_{p1} \cdot \ln \frac{\omega_{p2}}{\omega_{p3}\alpha_p(t)+\omega_{p4}\alpha_p^2(t)}$, where, $\omega_{pj} \in \mathbb{R}_{++}$ for $j = 1,\ 2,\ 3,\ 4$.

Given the privacy utility function $U_{priv}(\alpha_p(t))$, there exists an optimal value of $\alpha_p^*(t)$ that minimizes the following problem at time t:

$$\begin{aligned} &\min_{\alpha_p(t)} \mathscr{J}(t) = L_{acc}(\alpha_p(t)) - U_{priv}(\alpha_p(t)) \\ &s.t. 0 < \alpha_p(t) \le \alpha_U,\ \ 0 \le L_{acc}(\alpha_p(t)) \le c_3 \end{aligned} \tag{28}$$

where α_U and c_3 are the threshold values for $\alpha_p(t)$ and L_{acc}, respectively, beyond which is considered as non-private and non-accurate, respectively.

For training the classifier, we use a few fixed values of ρ and test the empirical loss $\overline{C}(t) = \frac{C^R}{B_p}\sum_{i=1}^{B_p}\mathscr{L}_{LR}(t)$ of the classifier. Then, we select the value of ρ that minimizes the empirical loss for a fixed α_p (0.3 in this experiment). We also test the non-private Algorithm 1, and the corresponding minimum ρ is obtained as the control. We choose the

TABLE II

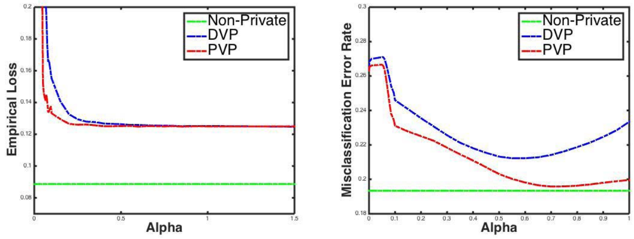
VALUES OF ρ

Algorithm	1	2	3
ρ	10^{-10}	$10^{-2.5}$	10^{-1}

TABLE III

VALUES OF C^R

Algorithm	1	2	3
ρ	1750	1750	1750



(a) (b)

Fig. 5. Privacy-accuracy tradeoff. (a): Empirical risk vs. α_p of final optimum output. (b): Misclassifications error rate vs. α_p of iteration 100.

corresponding optimal values of the regularization parameter ρ and C^R in Table II and III, respectively,

Figure 3 shows the convergence of DVP and PVP at different values of $\alpha_p(t)$ at a given iteration t. Larger values of α_p yield better convergence for both perturbations. Moreover, the DVP has a smaller variance of empirical loss than the primal perturbation does. However, a larger α_p leads to poorer privacy. Figure 4 $(a)-(b)$ shows the privacy-accuracy tradeoff of DVP at different iterations. By curve fitting, we model the function

$$L_{acc}(\alpha_p(t)) = c_4 \cdot e^{-c_5\alpha_p(t)} + c_6,$$

where $c_4,\ c_5, c_6 \in \mathbb{R}_+$. From the experimental results, we determine $c_4 = 0.2$, $c_5 = 25$, $c_6 = \min_t\{\overline{C}(t)\}$; these values are applicable at all iterations.

Figure 4 $(c)-(d)$ presents the privacy-accuracy tradeoff of PVP at different iterations. We model the function L_{acc} in the same way as DVP. In our experiment, we choose $\omega_{p1} = 0.02$, $\omega_{p2} = 6$, $\omega_{p3} = 9$, $\omega_{p4} = 1$. From Figure 3, we can see that the experimental results of $L_{acc}(\alpha_p(t))$ given $\{\alpha_p(t)\}$ for PVP experiments more oscillations than the DVP does. For iteration $t > 1$, $c_4 = 20$, $c_5 = 20$, $c_6 = \frac{1}{81}\sum_{t=20}^{100}\overline{C}(t)$. As shown in Figure 3, the empirical loss of DVP is more robust to noise than the PVP for most values of $\alpha_p(t)$. Moreover, the dual perturbation yields a lower error rate for a large range of values of $\alpha_p(t)$, which implies a better management of tradeoff between privacy and accuracy. Figure 5 shows the privacy-accuracy tradeoff of the final optimum classifier in terms of the empirical loss and misclassification error rate (MER). The MER is determined by the fraction of times the trained classifier predicts a wrong label. Since we are interested in $\alpha_p(t) < 1$, we can see that PVP performs slightly better than DVP with respect to the empirical loss for $\alpha_p(t) < 1$.

VI. Conclusion

In this work, we have developed two ADMM-based algorithms to solve a centralized regularized ERM in a distributed fashion while providing α-differential privacy for the ADMM iterations as well as the final trained output. Thus, the sensitive information stored in the training dataset at each node is protected against both the internal and the external adversaries.

Based on distributed training datasets, Algorithm 2 perturbs the dual variable $\lambda_p(t-1)$ generated at iteration $t-1$ for every node $p \in \mathscr{P}$ at iteration t before the update of the primal variable $f_p(t)$. In Algorithm 3, we perturb the primal variable $f_p(t)$, whose noisy version $V_p(t)$ is then released to the neighboring nodes. Since the primal variables are shared among all the neighboring nodes, at time t, the noise directly involved in the optimization of parameter update comes from multiple nodes; as a result, the updated primal variable has more randomness than the dual perturbation case.

In general, the accuracy decreases as privacy requirements are more stringent. The tradeoff between the privacy and accuracy is studied through numerical experiments. Our experiments are conducted using real data from UCI Machine Learning Repository.

There are two criteria to measure the performance of the algorithms. One is the performance of balancing the privacy-accuracy tradeoff; the other is the performance of learning that is quantified by the number of data points required to achieve α_{acc}-optimal results shown in Section IV. Our experiments show that the primal variable perturbation slightly outperforms the dual variable perturbation in balancing the privacy-accuracy tradeoff, while our theoretical analysis shows that the dual variable perturbation outperforms the primal variable perturbation in the performance of learning.

Appendix A
Proof of Theorem 1

Proof (Theorem 1): Let $f_p(t+1)$ be the optimal primal variable with zero duality gap. From the Assumption 1 and 2, we know that both the loss funciton $\mathscr{L}$ and the regularizer $R(\cdot)$ are differentiable and convex, and by using the Karush-Kuhn-Tucker (KKT) optimality condition (stationarity), we have the relationship between the noise $\epsilon_p(t)$ and the optimal primal variable $f_p(t+1)$ as:

$$\begin{aligned}\epsilon_p(t) = &-\sum_{i=1}^{B_p} y_{ip}\mathscr{L}'(y_{ip}f_p(t+1)^T x_{ip})x_{ip} - \frac{B_p}{C^R}\rho\nabla R(f_p)\\ &-\frac{2B_p}{C^R}\lambda_p(t) - \frac{B_p}{C^R}(\Phi + 2\eta N_p)f_p(t+1)\\ &+\frac{B_p\eta}{C^R}\sum_{i\in\mathscr{N}_p}(f_p(t)+f_i(t)).\end{aligned} \tag{29}$$

Under Assumption 1, the augmented Lagrange function $L_p^{dual}(t)$ is strictly convex, thus there is a unique value of $f_p(t+1)$ for fixed $\epsilon_p(t)$ and dataset D_p. The equation (29) shows that for any value of $f_p(t+1)$, we can find a unique value of $\epsilon_p(t)$ such that $f_p(t+1)$ is the minimizer of L_p^{dual}. Therefore, given a dataset D_p, the relation between $\epsilon_p(t)$ and $f_p(t+1)$ is bijective.

Let D_p and D'_p be two datasets with $H_d(D_p, D'_p) = 1$, $(x_i, y_i) \in D_p$ and $(x'_i, y'_i) \in D'_p$ are the corresponding two different data points. Let two matrices $\mathbf{J}_f(\epsilon_p(t)|D_p)$ and $\mathbf{J}_f(\epsilon'_p(t)|D'_p)$ denote the Jacobian matrices of mapping from $f_p(t+1)$ to $\epsilon_p(t)$ and $\epsilon'_p(t)$, respectively. Then, transformation from noise $f_p(t+1)$ to $\epsilon_p(t)$ by Jacobian yields:

$$\frac{Q(f_p(t+1)|D_p)}{Q(f_p(t+1)|D'_p)} = \frac{q(\epsilon_p(t)|D_p)}{q(\epsilon'_p(t)|D'_p)} \frac{|\det(\mathbf{J}_f(\epsilon_p(t)|D_p))|^{-1}}{|\det(\mathbf{J}_f(\epsilon'_p(t)|D'_p))|^{-1}}, \tag{30}$$

where $q(\epsilon_p(t)|D_p)$ and $q(\epsilon'_p(t)|D'_p)$ are the densities of $\epsilon_p(t)$ and $\epsilon'_p(t)$, respectively, given $f_p(t+1)$ when the datasets are D_p and D'_p, respectively.

Therefore, in order to prove the ratio of conditional densities of optimal primal variable is bounded as: $\frac{Q(f_p(t)|D)}{Q(f_p(t)|D'_p)} \le e^{\alpha_p(t)}$, we have to show: $\frac{q(\epsilon_p(t)|D_p)}{q(\epsilon'_p(t)|D'_p)} \cdot \frac{|\det(\mathbf{J}_f(\epsilon_p(t)|D_p))|^{-1}}{|\det(\mathbf{J}_f(\epsilon'_p(t)|D'_p))|^{-1}} \le e^{\alpha_p(t)}$. We first bound the ratio of the determinant of Jacobian matrices, and then the ratio of conditional densities of the noise vectors.

Let x^a be the a-th element of the vector x, and (a, b). Let $\mathbf{E} \in \mathbb{R}^{d \times d}$ be a matrix, then let $\mathbf{E}^{(a,b)}$ denote the (a, b)-th entry of the matrix $\mathbf{E}$. Thus, the (m, n)-th entry of $\mathbf{J}_f(\epsilon_p(t))$ is:

$$\begin{aligned}\mathbf{J}_f(\epsilon_p(t))^{(m,n)} &= -\sum_{i=1}^{B_p}(y_i^2\mathscr{L}''(y_i f_p(t+1)^T x_i)x_i^{(m)}x_i^{(n)} \\ &\quad -\frac{B_p}{C^R}\rho\nabla^2 R(f_p(t+1))^{(m,n)} \\ &\quad -\frac{B_p}{C^R}(\Phi + 2\eta N_p)\mathbb{1}(j = k).\end{aligned}$$

Let $\mathbf{J}_f^0(x_i, y_i) = (y_i^2\mathscr{L}''(y_i f_p(t+1)^T x_i)x_i x_i^T$, then the Jacobian matrix can be expressed as:

$$\begin{aligned}\mathbf{J}_f(\epsilon_p(t)|D_p) &= -\sum_{i=1}^{B_p}\mathbf{J}_f^0(x_i, y_i) - \frac{B_p}{C^R}\rho\nabla^2 R(f_p(t+1)) \\ &\quad -\frac{B_p}{C^R}(\Phi + 2\eta N_p)\mathbf{I}_d.\end{aligned}$$

Let $\mathbf{M} = \mathbf{J}_f^0(x'_i, y'_i) - \mathbf{J}_f^0(x_i, y_i)$, and $\mathbf{H} = -\mathbf{J}_f(\epsilon_p(t)|D_p)$, and thus $\mathbf{J}_f(\epsilon_p(t)|D'_p) = -(\mathbf{M} + \mathbf{H})$. Let $h_j(\mathbf{W})$ be the j-th largest eigenvalue of a symmetric matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ with rank θ. Then, we have the following fact: $\det(\mathbf{I} + \mathbf{W}) = \prod_j^{\theta}(1 + h_j(\mathbf{W}))$. Since the matrix $x_i x_i^T$ has rank 1, matrix $\mathbf{M}$ has rank at most 2; thus matrix $\mathbf{H}^{-1}\mathbf{M}$ has rank at most 2; therefore, we have:

$$\begin{aligned}\det(\mathbf{H} + \mathbf{M}) &= \det(\mathbf{H}) \cdot \det(\mathbf{I} + \mathbf{H}^{-1}\mathbf{M}) \\ &= \det(\mathbf{H}) \cdot (1 + h_1(\mathbf{H}^{-1}\mathbf{M}))(1 + h_2(\mathbf{H}^{-1}\mathbf{M}).\end{aligned}$$

Thus, the ratio of determinants of the Jacobian matrices can be expressed as:

$$\begin{aligned}\frac{|\det(\mathbf{J}_f(\epsilon_p(t)|D_p))|^{-1}}{|\det(\mathbf{J}_f(\epsilon'_p(t)|D'_p))|^{-1}} &= \frac{|\det(\mathbf{H} + \mathbf{M})|}{|\det(\mathbf{H})|} \\ &= |\det(\mathbf{I} + \mathbf{H}^{-1}\mathbf{M})| \\ &= (1 + h_1(\mathbf{H}^{-1}\mathbf{M}))(1 + h_2(\mathbf{H}^{-1}\mathbf{M}) \\ &= |1 + h_1(\mathbf{H}^{-1}\mathbf{M}) + h_2(\mathbf{H}^{-1}\mathbf{M}) \\ &\quad + h_1(\mathbf{H}^{-1}\mathbf{M})h_2(\mathbf{H}^{-1}\mathbf{M})|.\end{aligned}$$

Based on Assumption 2, all the eigenvalues of $\nabla^2 R(f_p(t+1))$ is greater than 1 [21]. Thus, from Assumption 1, matrix $\mathbf{H}$ has all eigenvalues at least $\frac{B_p}{C^R}\left(\rho+\Phi+2\eta N_p\right)$. Therefore, $|h_1(\mathbf{H}^{-1}\mathbf{M})| \le \frac{|h_i(\mathbf{M})|}{\frac{B_p}{C^R}\left(\rho+\Phi+2\eta N_p\right)}$. Let $\sigma_i(\mathbf{M})$ be the non-negative singular value of the symmetric matrix $\mathbf{M}$. According to [22], we have the inequality $\sum_i |h_i(\mathbf{M})| \le \sum_i \sigma_i(\mathbf{M})$. Thus, we have $|h_1(\mathbf{M})| + |h_2(\mathbf{M})| \le \sigma_1(\mathbf{M}) + \sigma_2(\mathbf{M})$. Let $\| X \|_\Sigma = \sum_i \sigma_i$ be the trace norm of X. Then, according to the *trace norm inequality*, we have $\| \mathbf{M} \|_\Sigma \le \| \mathbf{J}^0(x'_i, y'_i) \|_\Sigma + \| -\mathbf{J}^0(x_i, y_i) \|_\Sigma$. As a result, based on the upper bounds from Assumption 1 and 3, we have:

$$\begin{aligned}|h_1(\mathbf{M})| + |h_2(\mathbf{M})| &\le \| \mathbf{J}^0(x'_i, y'_i) \|_\Sigma + \| -\mathbf{J}^0(x_i, y_i) \|_\Sigma \\ &\le |(y_i^2\mathscr{L}''(y_i f_p(t+1)^T x_i)| \cdot \| x_i \| \\ &\quad + |(y_i'^2\mathscr{L}''(y'_i f_p(t+1)^T x'_i)| \cdot \| x'_i \| \le 2c_1,\end{aligned}$$

which follows $h_1(\mathbf{M})h_2(\mathbf{M}) \le c_1^2$. Finally, the ratio of determinants of Jacobian matrices is bounded as:

$$\frac{|\det(\mathbf{J}_f(\epsilon_p(t)|D_p))|^{-1}}{|\det(\mathbf{J}_f(\epsilon'_p(t)|D'_p))|^{-1}} \le (1 + \frac{c_1}{\frac{B_p}{C^R}\left(\rho + \Phi + 2\eta N_p\right))})^2 = e^{\overline{\alpha}}, \tag{31}$$

where $\overline{\alpha} = \ln\left(1 + \frac{c_1}{\frac{B_p}{C^R}\left(\rho+\Phi+2\eta N_p\right)}\right)^2$.

Now, we bound the ratio of densities of $\epsilon_p(t)$. Let $sur(E)$ be the surface area of the sphere in d dimension with radius E, and $sur(E) = sur(1) \cdot E^{d-1}$. We can write:

$$\begin{aligned}\frac{q(\epsilon_p(t)|D_p)}{q(\epsilon'_p|D'_p)} &= \frac{\mathscr{K}(\epsilon_p(t))\frac{\|\epsilon_p(t)\|^{d-1}}{sur(\|\epsilon_1(t)\|)}}{\mathscr{K}(\epsilon'_p(t))\frac{\|\epsilon'_p(t)\|^{d-1}}{sur(\|\epsilon'_p(t)\|)}} \\ &\le e^{\zeta_p(t)(\|\epsilon'_p(t)\| - \|\epsilon_p(t)\|)} \le e^{\hat{\alpha}_p},\end{aligned} \tag{32}$$

where $\hat{\alpha}_p$ is a constant satisfying the above inequality. Since we want to bound the ratio of densities of $f_p(t+1)$ as $\frac{Q(f_p(t+1)|D_p)}{Q(f_p(t+1)|D'_p)} \le e^{\alpha_p(t)}$, we need $\hat{\alpha}_p \le \alpha_p(t) - \overline{\alpha}$. For non-negative Φ, let $\hat{\alpha}_p = \alpha_p(t) - \ln\left(1 + \frac{c_1}{\frac{B_p}{C^R}\left(\rho+2\eta N_p\right)}\right)^2$. If $\hat{\alpha}_p > 0$, then we fix $\Phi = 0$, and thus $\hat{\alpha}_p = \alpha_p(t) - \overline{\alpha}$. Otherwise, let $\Phi = \frac{c_1}{\frac{B_p}{C^R}(e^{\alpha_p(t)/4} - 1)} - \rho - 2\eta N_p$, and $\hat{\alpha}_p = \frac{\alpha_p(t)}{2}$, then $\hat{\alpha}_p = \alpha_p(t) - \overline{\alpha}$. Therefore, we can have $\frac{|\det(J_f(b_1|D_p))|^{-1}}{|\det(J_f(b_2|D'_p))|^{-1}} \le e^{\alpha_p(t) - \hat{\alpha}_p}$. From the upper bounds stated in Assumption 1 and 3, the l_2 norm of the difference of ϵ_1 and ϵ_2 can be bounded as:

$$\begin{aligned}\| \epsilon'_p(t) - \epsilon_p(t) \| &= \sum_{i=1}^{B_p} \| y_{ip}\mathscr{L}'(y'_{ip} f_p(t+1)^T x'_{ip})x'_{ip} \\ &\quad -(y_{ip}\mathscr{L}'(y_{ip} f_p(t+1)^T x_{ip})x_{ip} \| \le 2.\end{aligned}$$

Thus, $\| \epsilon'_p(t) \| - \| \epsilon_p(t) \| \le \| \epsilon'_p(t) - \epsilon_p(t) \| \le 2$. Therefore, by selecting $\zeta_p(t) = \frac{\hat{\alpha}_p}{2}$, we can bound the ratio of conditional densities of $f_p(t+1)$ as $\frac{Q(f_p(t+1)|D_p)}{Q(f_p(t+1)|D'_p)} \le e^{\alpha_p(t)}$, and prove that the DVP can provide $\alpha_p(t)$-differential privacy. □

Appendix B
Proof of Theorem 2

Proof (Theorem 2): Let D_p and D'_p be two datasets with $H_d(D_p, D'_p) = 1$. Since only $V_p(t)$ is released, then our target is to prove $\frac{Q(V_p(t+1)|D_p)}{Q(V_p(t+1)|D'_p)} \leq e^{\alpha_p(t)}$. From (19), we have: $\frac{Q(V_p(t+1)|D_p)}{Q(V_p(t+1)|D'_p)} = \frac{\mathscr{K}(\epsilon_p(t))}{\mathscr{K}(\epsilon'_p(t))} = \frac{e^{-\zeta_p(t)\|\epsilon_p(t)\|}}{e^{-\zeta_p(t)\|\epsilon'_p(t)\|}}$. Therefore, in order to make the model to provide $\alpha_p(t)$-differential privacy, we need to find a $\zeta_p(t)$ that satisfies

$$\zeta_p(t)(\| \epsilon_p(t) \| - \| \epsilon'_p(t) \|) \leq \alpha_p(t). \tag{33}$$

Let $V^A = \arg\min_{V_p} L_p^{prim}(t|D_p)$, and $V^B = \arg\min_{V_p} L_p^{prim}(t|D'_p)$, where $L_{prim}(t|D)$ is the augmented Lagrange function for PVP given dataset D.

Let F, G be defined at each node $p \in \mathscr{P}$ as: $F(V_p(t)) = L_p^{prim}(t|D_p)$, and $G(V_p(t)) = L_p^{prim}(t|D'_p) - L_p^{prim}(t|D_p)$, respectively. Thus, $G(V_p) = \frac{C^R}{B_p}\sum_{i=1}^{B_p}(\mathscr{L}(y'_{ip}V_p^T x'_{ip}) - \mathscr{L}(y_{ip}V_p^T x_{ip}))$. According to Assumption 2, we can imply that $L_p^{prim}(t|D_p) = F(V_p(t))$ and $L_p^{prim}(t|D'_p) = F(V_p(t)) + G(V_p(t))$ are both ρ-strong convex. Differentating $G(V_p(t))$ with respect to $V_p(t)$ gives: $\nabla G(V_p) = \frac{C^R}{B_p}(y'_{ip}\mathscr{L}'(y'_{ip}V_p^T x'_{ip})x'_{ip} - (y_{ip}\mathscr{L}(y_{ip}V_p^T x_{ip})x_{ip}$. From Assumption 1 and 3, $\| \nabla G(V_p) \| \leq \frac{2C^R}{B_p}$. From definitions of V^A and V^B, we have: $\nabla F(V^A) = \nabla F(V^B) + \nabla F(V^B) = 0$. From [23, Lemma 14] and the fact that $F(\cdot)$ is ρ-strongly convex, weh have the following inequality: $\langle \nabla F(V^A) - F(V^B), V^A - V^B \rangle \geq \rho \| V^A - V^B \|^2$; therefore, Cauchy-Schwarz inequality yields:

$$\begin{aligned}\| V^A - V^B \| \cdot \| \nabla G(V^B) \| &\geq (V^A - V^B)^T \nabla G(V^B)\\ = \langle \nabla F(V^A) - F(V^B), V^A - V^B \rangle &\geq \rho \| V^A - V^B \|^2 .\end{aligned}$$

Dividing both sides by $\rho \| V^A - V^B \|$ gives:

$$\| V^A - V^B \| \leq \frac{1}{\rho} \| \nabla G(V^B) \| \leq \frac{2C^R}{\rho B_p}. \tag{34}$$

From (19), we have $\| V^A - V^B \| \leq \frac{1}{\rho} \| \nabla G(V^B) \| = \| \epsilon_p(t) - \epsilon'_p(t) \|$. Thus, we can bound

$$\begin{aligned}\zeta_p(t)(\| \epsilon_p(t) \| - \| \epsilon'_p(t) \|) &\leq \zeta_p(t)(\| \epsilon_p(t) - \epsilon'_p(t) \|)\\ &\leq \frac{2C^R}{B_p\rho}\zeta_p(t)\end{aligned}$$

Therefore, by choosing $\zeta_p(t) = \frac{\rho B_p \alpha_p(t)}{2C^R}$, the inequality (33) holds; thus PVP is dynamic α_p-differentially private at each node p. □

Appendix C
Proof of Theorem 3

Proof (Theorem 3): Let $\hat{f}_p(t+1) = \arg\min_{f_p} \hat{Z}_p(f_p, t)$, and $f^* = \arg\min_{f_p} Z_p(f_p, t|D_p)$. Let $f^p(t+1)$ be the actual estimated optimum obtained using Algorithm 1. We assume that $f^p(t+1)$ is very close to the actually so that $Z_p(f^p(t+1)|D_p) - Z_p(f^*|D_p) \to 0$. For the non-private ERM, [18] and [23] show that for a specific reference classifier f^0 at time $t+1$ such that $\hat{C}(f^0) = \hat{C}^0$, we have:

$$\begin{aligned}\hat{C}(f^p(t+1)) = \hat{C}^0 &+ \big(\hat{Z}_p(f^p(t+1), t) - \hat{Z}_p(\hat{f}_p(t+1), t)\big)\\ &+ \big(\hat{Z}_p(\hat{f}_p(t+1), t) - \hat{Z}_p(f^0, t)\big)\\ &+ \frac{\rho}{2} \| f^0 \|^2 - \frac{\rho}{2} \| f^p(t+1) \|^2 .\end{aligned}$$

From Sridharan et al. [24], we have, with probability at least $1 - \delta$

$$\begin{aligned}&\hat{Z}_p(f^p, t) - \hat{Z}_p(\hat{f}_p(t+1), t)\\ &\quad\leq 2\big(Z_p(f^p(t+1)|D_p) - Z_p(f^*|D_p)\big) + \mathscr{O}\Big(C^R\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big).\end{aligned}$$

Since $Z_p(f^p(t+1)|D_p) - Z_p(f^*|D_p) \to 0$, then, $\hat{Z}_p(f^p(t+1), t) - \hat{Z}_p(\hat{f}_p(t+1), t) \leq \mathscr{O}\Big(C^R\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big)$. If we choose $\rho \leq \frac{\alpha_{acc}}{\|f^0\|^2}$, then, $\frac{\rho}{2} \| f^0 \|^2 - \frac{\rho}{2} \| f^p(t+1) \|^2 \leq \frac{\alpha_{acc}}{2}$. Thus, $\hat{C}(f^p(t+1)) \leq \hat{C}^* + \mathscr{O}\Big(C^R\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big) + \frac{\alpha_{acc}}{2}$. Therefore, we can find the value of B_p by solving $\mathscr{O}\Big(C^R\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big) + \frac{\alpha_{acc}}{2} \leq \alpha_{acc}$, we obtain: $B_p > \beta_{non}\max\left(\frac{C^R\|f^0\|^2\ln(\frac{1}{\delta})}{\alpha_{acc}^2}\right)$.

If we determine different reference classifier $f_p^0(t+1)$ at different time, then we need to find the maximum value across the time and among different value of $\| f^0 \|$:

$$B_p > \beta_{non}\max\left(\max_t\Big(\frac{C^R \| f^0 \|^2 \ln(\frac{1}{\delta})}{\alpha_{acc}^2}\Big)\right).$$

Let $f_p^{non}(t+1) = \arg\min_{f_p} L_p^N(t)$. Since $\hat{C}(f_p^{non}(t+1)) = \hat{C}(f^p(t+1)) + \Delta^{non}(t)$, then, $\hat{C}(f_p^{non}(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc} + \Delta^{non}(t)$, with probability no less than $1-\delta$. □

Appendix D

In this appendix, we provide the proof of Theorem 4 with the help of Lemma 17 and 18, which are also proved later.

Proof (Theorem 4): First we define $\hat{f}_p(t+1)$ and f^* in the same way as in Appendix C. We also define $f_p^*(t+1) = \arg\min_{f_p} Z_p^{dual}(f_p, t|D_p)$, and $\hat{C}(f_p^0(t)) = \hat{C}^0(t)$ at time t. We use the analysis of [18] and [23] (also see the work of Chaudhuri et al. in [19]), and have the following:

$$\begin{aligned}&\hat{C}(f_p^*(t+1))\\ &\quad= \hat{C}^0(t+1) + \big(\hat{Z}_p(f_p^*(t+1), t) - \hat{Z}_p(\hat{f}_p(t+1), t)\big)\\ &\qquad+ \big(\hat{Z}_p(\hat{f}_p(t+1), t) - \hat{Z}_p(f_p^0(t+1), t)\big)\\ &\qquad+ \frac{\rho}{2} \| f_p^0(t+1) \|^2 - \frac{\rho}{2} \| f_p^*(t+1) \|^2 .\end{aligned} \tag{35}$$

Now we bound each terms in the right hand side of (35) as follows. From Assumption 1, we have $\mathscr{L}' \leq c_1$. By choosing $B_p > \frac{5c_1C^R\|f_p^0(t+1)\|^2}{\alpha_{acc}\alpha_p(t)}$, and $\rho > \frac{\alpha_{acc}}{2\|f_p^0(t+1)\|^2}$, and since $\alpha_p(t) \leq 1$, we have:

$$\begin{aligned}\hat{\alpha}_p &= \alpha_p(t) - \ln\left(1 + \frac{c_1}{\frac{B_p}{C^R}(\rho + 2\eta N_p)}\right)^2\\ &> \alpha_p(t) - \ln(1 + \frac{c_1C^R}{B_p\rho})^2 > \alpha_p(t) - \ln(1 + \frac{2\alpha_p(t)}{5})^2\end{aligned}$$

$$> \alpha_p(t) - \frac{4\alpha_p(t)}{5} = \frac{\alpha_p(t)}{5}.$$

Then, according to Algorithm 2, we choose the corresponding $\zeta_p(t) = \frac{\alpha_p(t)}{4}$ because $\hat{\alpha}_p > 0$. Let Λ be the event

$$\Lambda := \Big\{ Z_p(f_p^*(t+1)|D_p) \le Z_p(f^*|D_p) + \frac{16d^2\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho B_p^2 \alpha_p(t)^2} \Big\}.$$

Since $\hat{\alpha}_p > \frac{\alpha_p(t)}{2} > 0$, and applying Lemma 18 yields $\mathbb{P}_{\epsilon_p(t)}\Big(\Lambda\Big) \ge 1 - \delta$. From the work of Sridharan et al. in [24], the following inequality holds with probability $1-\delta$

$$\begin{aligned}
&\hat{Z}_p(f_p^*(t+1)) - \hat{Z}_p(\hat{f}_p(t+1)) \\
&\le 2\Big(Z_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p)\Big) + \mathscr{O}\Big(\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big) \\
&\le \frac{32d^2\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho B_p^2\alpha_p(t)^2} + \mathscr{O}\Big(\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big).
\end{aligned}$$

The big-$\mathscr{O}$ notation hides the numerical constants, which depend on the derivative of the loss function and the upper bounds of the data points shown in Assumption 3. Combining the above two processes, $\hat{Z}_p(f_p^*(t+1)) - \hat{Z}_p(\hat{f}_p(t+1))$ is bounded as shown above with probability $1-2\delta$.

From the definitions of $f_p^0(t+1)$ and $\hat{f}_p(t+1)$, we obtain $\hat{Z}_p(\hat{f}_p(t+1),t) - \hat{Z}_p(f_p^0(t+1),t) < 0$. Since $P \ge 1$, then by selecting $\rho = \frac{\alpha_{acc}}{\|f_p^0(t+1)\|^2}$, we can bound $\frac{\rho}{2} \parallel f_p^0(t+1) \parallel^2 - \frac{\rho}{2} \parallel f_p^*(t+1) \parallel^2 \le \frac{\alpha_{acc}}{2}$. Therefore, from (47), we have:

$$\hat{C}(f_p^*(t+1)) \le C_E^0 + \frac{32d^2\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho B_p^2\alpha_p(t)^2} + \mathscr{O}\Big(C^R\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big) + \frac{\alpha_{acc}}{2},$$

with $\rho = \frac{6\alpha_{acc}}{\|f_p^0(t+1)\|^2}$. The lower bounds of B_p is determined by solving the following:

$$\frac{32d^2\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho B_p^2\alpha_p(t)^2} + \mathscr{O}\Big(C^R\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big) + \frac{\alpha_{acc}}{2} \le \alpha_{acc}.$$

□

Lemma 17: Let Z be a gamma random variable with density function $\Gamma(k,\theta)$, where k is an integer, and let $\delta > 0$. Then, we have:

$$\mathbb{P}(Z < k\theta\ln(\frac{k}{\delta})) \ge 1 - \delta.$$

Proof (Lemma 17): Since Z is a gamma random variable $\Gamma(k,\theta)$, then we can express Z as $Z = \sum_{i=1}^k Z_i$, where $\{Z_i\}_{i=1}^k$ are independent exponential random variables with density function $Exp(\frac{1}{\theta})$; thus, for each Z_i we have: $\mathbb{P}(Z_i \le \theta\ln(\frac{k}{\delta})) = 1 - \frac{\delta}{k}$. Since $\{Z_i\}_{i=1}^k$ are independent, we have:

$$\mathbb{P}(Z < k\theta\ln(\frac{k}{\delta})) = \prod_{i=1}^k \mathbb{P}(Z_i \le \theta\ln(\frac{k}{\delta})) = (1-\frac{\delta}{k})^k \ge 1-\delta.$$

□

Lemma 18: Let $\hat{\alpha}_p > 0$, and $f_p^*(t+1) = \arg\min_{f_p} Z_p^{dual}(f_p, t|D_p)$, and $f^* = \arg\min_{f_p} Z_p(f_p|D_p)$. Let Λ be the event

$$\Lambda := \Big\{ Z_p(f_p^*(t+1)|D_p) \le Z_p(f^*|D_p) + \frac{16d^2\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho B_p^2\alpha_p(t)^2} \Big\}.$$

Under Assumption 1 and 2, we have: $\mathbb{P}_{\epsilon_p(t)}\Big(\Lambda\Big) \ge 1-\delta$. The probability $\mathbb{P}_{\epsilon_p(t)}$ is taken over the noise vector $\epsilon_p(t)$.

Proof (Lemma 18): Since $\hat{\alpha}_p > 0$, $\Phi = 0$, $f_p^*(t+1) = \arg\min_{f_p} Z_p^{dual}(f_p,t|D_p)$ can be expressed as: $f_p^*(t+1) = \arg\min_{f_p}\Big(Z_p(f_p|D_p) + 2\epsilon_p(t)^T f_p\Big)$. Thus, we have:

$$Z_p(f_p^*(t+1)|D_p) \le Z_p(f^*|D_p) + \frac{C^R}{B_p}\epsilon_p(t)^T(f^* - f_p^*(t+1)).$$

Firstly, we bound the l_2-norm $\parallel f^* - f_p^*(t+1) \parallel$. We use the similar procedure to establish (34) in Appendix C by setting $F(Y) = Z_p(Y|D_p)$ and $G(Y) = \frac{C^R}{B_p}\epsilon_p(t)$; thus, based on Assumption 1 and 2, we have:

$$\parallel f^* - f_p^*(t+1) \parallel \le \frac{1}{\rho} \parallel \nabla\big(2\epsilon_p(t)^T f_p\big) \parallel \le \frac{C^R \parallel \epsilon_p(t) \parallel}{B_p\rho}.$$

Cauchy-Schwarz inequality yields:

$$\begin{aligned}
&{}_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p) \\
&\le \parallel Z_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p) \parallel \\
&\le \frac{2}{B_p} \parallel \epsilon_p(t)^T(f^* - f_p^*(t+1) \parallel \le \frac{\big(C^R\big)^2 \parallel \epsilon_p(t) \parallel^2}{B_p^2\rho}.
\end{aligned}$$

Since the noise vector $\epsilon_p(t)$ is drawn from $\mathscr{K}_p(\epsilon) \sim e^{-\zeta_p(t)\|\epsilon\|}$, then $\parallel \epsilon_p(t) \parallel$ is drawn from $\Gamma(d, \frac{1}{\zeta_p(t)}) = \Gamma(d, \frac{2}{\hat{\alpha}_p})$. Then, by using Lemma 17 with $\parallel \epsilon_p(t) \parallel \le \frac{2d\ln(\frac{d}{\delta})}{\hat{\alpha}_p}$, we have:

$$\Lambda := \Big\{ Z_p(f_p^*(t+1)|D_p) \le Z_p(f^*|D_p) - \frac{4d^2\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho B_p^2\alpha_p(t)^2}.$$

with probability no less than $1-\delta$. □

Appendix E

We prove Theorem 5 here. This appendix also shows Lemma19 that are used in the proof of Theorem 5.

Proof (Theorem 5): We define $\hat{f}_p(t+1)$ and f^* as in the proof of Theorem 4 in Appendix D, and we also define $f_p^*(t+1) = \arg\min_{f_p} Z_p^{prim}(f_p,t|D_p)$. Let $\hat{C}(f_p^0(t)) = \hat{C}^0(t)$ at time t. As in Appendix D, we again use the analysis of [18] and [23] (also see the work of Chaudhuri et al. in [19]), and have the follows

$$\begin{aligned}
\hat{C}(f_p^*(t+1)) = {}& \hat{C}(f_p^0(t+1)) + \big(\hat{Z}_p(f_p^*(t+1),t) \\
& - \hat{Z}_p(\hat{f}_p(t+1),t)\big) \\
& + \big(\hat{Z}_p(\hat{f}_p(t+1),t) - \hat{Z}_p(f_p^0(t+1),t)\big) \\
& + \frac{\rho}{2} \parallel f_p^0(t+1) \parallel^2 - \frac{\rho}{2} \parallel f_p^*(t+1) \parallel^2 .
\end{aligned} \tag{36}$$

According to Theorem 2, we choose $\zeta_p(t) = \frac{\rho B_p \alpha_p(t)}{2C^R} > 0$. Thus, applying Lemma 19, we have:

$$Z_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p) \leq \frac{16\big(C^R\big)^2\eta^2N_p^2d^2\big(\ln(\frac{d}{\delta})\big)^2}{\rho^3B_p^2\alpha_p(t)^2},$$

with probability no smaller than $1-\delta$. Then, we use the result of Sridharan et al. in [24], with probability no smaller than $1-\delta$:

$$\begin{aligned}&\hat{Z}_p(f_p^*(t+1)) - \hat{Z}_p(\hat{f}_p(t+1))\\ &\quad\leq 2\Big(Z_p(f_p^*(t+1)|D_p)\\ &\qquad -Z_p(f_p^*(t+1)|D_p)\Big) + \mathscr{O}\Big(\frac{\ln(\frac{d}{\delta})}{B_p\rho}\Big)\\ &\quad\leq \frac{32\big(C^R\big)^2\eta^2N_p^2d^2\big(\ln(\frac{d}{\delta})\big)^2}{\rho^3B_p^2\alpha_p(t)^2} + \mathscr{O}\Big(\frac{\ln(\frac{1}{\delta})}{B_p\rho}\Big).\end{aligned}$$

Combining the above two processes, we have the probability no smaller than $1-2\delta$.

In order to bound the last two terms in (36), we select $\rho = \frac{\alpha_{acc}}{\|f_p^0(t+1)\|^2}$; as a result, $\frac{\rho}{2}\parallel f_p^0(t+1)\parallel^2 - \frac{\rho}{2}\parallel f_p^*(t+1)\parallel^2 \leq \frac{\alpha_{acc}}{2}$. From the definitions of $\hat{f}_p(t+1)$ and $f_p^0(t+1)$, we have: $\hat{Z}_p(\hat{f}_p(t+1),t) - \hat{Z}_p(f_p^0(t+1),t) \leq 0$. The value of B_p is determined such that $\hat{C}(f_p^*(t+1)) \leq \hat{C}^* + \alpha_{acc}$. Therefore, we find the bounds of B_p by solving

$$\frac{32\big(C^R\big)^2\eta^2N_p^2d^2\big(\ln(\frac{d}{\delta})\big)^2}{\rho^3B_p^2\alpha_p(t)^2} + \mathscr{O}\Big(\frac{C^R\ln(\frac{1}{\delta})}{B_p\rho}\Big) + \frac{\alpha_{acc}}{2} \leq \alpha_{acc},$$

with $\rho = \frac{\alpha_{acc}}{\|f_p^0(t+1)\|^2}$. □

The following Lemma is analogous to Lemma 17.

Lemma 19: Let $\zeta_p(t) > 0$, *and* $f_p^*(t+1) = \arg\min_{f_p} Z_p^{prim}(f_p, t|D_p)$, *and* $f^* = \arg\min_{f_p} Z_p(f_p|D_p)$. *Suppose that the noise vector* $\epsilon_t(t)$ *generated at time* t *has the same value of* $\alpha_p(t)$ *for all* $p \in \mathscr{P}$. *Let* Λ *be the event*

$$\begin{aligned}\Lambda := \Big\{&Z_p(f_p^*(t+1)|D_p) \leq Z_p(f^*|D_p)\\ &+ \frac{16\big(C^R\big)^2\eta^2N_p^2d^2\big(\ln(\frac{d}{\delta})\big)^2}{\rho^3B_p^2\alpha_p(t)^2}\Big\}.\end{aligned}$$

If the loss function $\mathscr{L}$ *is convex and differentiable with* $|\mathscr{L}| \leq 1$, *then, we have:* $\mathbb{P}_{\epsilon_p(t)}\Big(\Lambda\Big) \geq 1-\delta$. *The probability* $\mathbb{P}_{\epsilon_p(t)}$ *is taken over the noise vector* $\epsilon_p(t)$.

Proof (Lemma 19): Let $\epsilon^{pi}(t) = \epsilon_p(t) - \epsilon_i(t)$ with probability density $P_{\epsilon^{pi}}$. Let $f_p^*(t+1) = \arg\min_{f_p} Z_p^{prim}(f_p, t|D_p)$, and it can be expressed as: $f_p^*(t+1) = \arg\min_{f_p}\Big(Z_p(f_p|D_p) - Y_p\Big)$, where $Y_p = \eta\sum_{i\in\mathscr{N}_p}\Big((f_p - \frac{1}{2}(f_p(t)+f_i(t))^T\cdot(\epsilon^{pi}(t)) + \frac{1}{4}\big(\epsilon^{pi}(t)\big)^2\Big)$. Thus, we have:

$$\begin{aligned}Z_p(f_p^*(t+1)|D_p) &\leq Z_p(f^*|D_p)\\ &\quad -\eta\sum_{i\in\mathscr{N}_p}(f^* - f_p^*(t+1))^T\cdot\epsilon^{pi}.\end{aligned}$$

Firstly, we bound the l_2-norm $\parallel f^* - f_p^*(t+1)\parallel$. We use the similar procedure to establish (46) in Appendix D by setting $F(\cdot) = Z_p(\cdot|D_p)$ and $G(\cdot) = \eta\sum_{i\in\mathscr{N}_p}\big(\epsilon^{pi}\big)^T(\cdot)$; thus, based on Assumption 1 and 2, we have:

$$\begin{aligned}&\parallel f^* - f_p^*(t+1)\parallel\\ &\quad\leq \frac{1}{\rho}\parallel\sum_{i\in\mathscr{N}_p}\nabla(\eta N_p(f_p^*(t+1))^T\epsilon^{pi})\parallel\\ &\quad\leq \sum_{i\in\mathscr{N}_p}\frac{\eta\parallel\epsilon^{pi}(t)\parallel}{\rho} = \sum_{i\in\mathscr{N}_p}\frac{\eta\Big(\parallel\epsilon_p(t)-\epsilon_j(t)\parallel\Big)}{\rho}\\ &\quad\leq \sum_{i\in\mathscr{N}_p}\frac{\eta\Big(\parallel\epsilon_p(t)\parallel + \parallel\epsilon_j(t)\parallel\Big)}{\rho}.\end{aligned}$$

Since $\alpha_p(t)$ is the same for all $p \in \mathscr{P}$ at time t, $\zeta_j(t) = \frac{\rho B_p\alpha_p(t)}{2C^R}$ for all $j \in \mathscr{P}$. Since $\epsilon_j(t)$ is drawn from (15), then, $\parallel\epsilon_p(t)\parallel\sim\Gamma(d, \frac{1}{\zeta_p(t)})$ for all $p \in \mathscr{P}$. Let $\parallel\epsilon_{pi}\parallel^{\oplus} = \parallel\epsilon_p(t)\parallel + \parallel\epsilon_i(t)\parallel$. Thus,

$$\parallel f^* - f_p^*(t+1)\parallel \leq \sum_{i\in\mathscr{N}_p}\frac{\eta\Big(\parallel\epsilon_{pi}\parallel^{\oplus}\Big)}{\rho} = \frac{\eta N_p\Big(\parallel\epsilon_{pi}\parallel^{\oplus}\Big)}{\rho}.$$

Cauchy-Schwarz inequality yields:

$$\begin{aligned}&{}_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p)\\ &\quad\leq \parallel Z_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p)\parallel\\ &\quad\leq \frac{\eta^2N_p^2\Big(\parallel\epsilon_{pi}\parallel^{\oplus}\Big)^2}{\rho},\end{aligned}$$

From the fact that if $\{X_j\}_{j=1}^K$ are independent gamma random variables with density $\Gamma(\beta_j, h)$, then $X = \sum_{j=1}^K X_j$ is a gamma random variable with $\Gamma(\sum_j^K\beta_j, h)$, we have $P_{\|\epsilon^{pj}\|} = \Gamma(2d, \frac{2C^R}{\rho B_p\alpha_p(t)})$. Applying Lemma 17 with $\parallel\epsilon^{pj}(t)\parallel^{\oplus} \leq \frac{4C^Rd\ln(\frac{d}{\delta})}{\rho B_p\alpha_p(t)}$ yields:

$$Z_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p) \leq \frac{16\big(C^R\big)^2\eta^2N_p^2d^2\big(\ln(\frac{d}{\delta})\big)^2}{\rho^3B_p^2\alpha_p(t)^2}$$

with probability no smaller than $1-\delta$. □

Appendix F

Theorem 6 is proved in this appendix based on Lemma 20.

Proof (Theorem 6): We use $\hat{f}_p(t+1)$, f^*, and $f_p^*(t+1)$, defined in the proof of Theorem 5 in Appendix E. Now we use a reference $f_p^0(t)$ such that $\hat{C}(f_p^*(t)) = \hat{C}^0(t)$ be the reference at time $t+1$. We use the analysis of [18] and [23] (also see the work of Chaudhuri et al. in [19]), and have the follows,

$$\begin{aligned}&\hat{C}(V_p^*(t+1))\\ &\quad= \hat{C}(f_p^0(t+1)) + \big(\hat{Z}_p(V_p^*(t+1),t) - \hat{Z}_p(\hat{f}_p(t+1),t)\big)\\ &\qquad+\big(\hat{Z}_p(\hat{f}_p(t+1),t) - \hat{Z}_p(f_p^0(t+1),t)\big)\\ &\qquad+\frac{\rho}{2}\parallel f_p^0(t+1)\parallel^2 - \frac{\rho}{2}\parallel V_p^*(t+1)\parallel^2. \qquad (37)\end{aligned}$$

If $R(f_p(t)) = \frac{1}{2} \parallel f_p(t) \parallel^2$, then, $\parallel \nabla^2 R(f_p(t)) \parallel \leq 1$. Thus, we can apply Lemma 20 with $\tau = 1$:

$$Z_p^{prim}(V_p^*(t+1), t|D_p) - Z_p^{prim}(f_p^*(t+1), t|D_p) \leq \frac{4\big(C^R\big)^2 d^2 \Big(\rho + c_4 C^R\Big)\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho^2 B_p^2 \alpha_p(t)^2},$$

with probability $\geq 1 - \delta$ over the noise. In the proof of Theorem 5, we have, with probability $1 - \delta$,

$$Z_p(f_p^*(t+1)|D_p) - Z_p(f^*|D_p) \leq \frac{4\eta^2 N_p^2 d^2 \big(\ln(\frac{d}{\delta})\big)^2}{\rho^3 B_p^2 \alpha_p(t)^2}.$$

Therefore, with probability $1 - 2\delta$, we have

$${}_p(V_p^*(t+1)|D_p) - Z_p(f^*|D_p) \leq \frac{4\eta^2 N_p^2 d^2 \big(\ln(\frac{d}{\delta})\big)^2}{\rho^3 B_p^2 \alpha_p(t)^2} + \frac{4d^2\Big(\rho + c_4\Big)\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho^2 B_p^2 \alpha_p(t)^2}.$$

Sridharan et al. in [24] shows, with probability $1 - \delta$,

$$\begin{aligned}\hat{Z}_p(V_p^*(t+1)) - \hat{Z}_p(\hat{f}_p(t+1)) &\leq 2\Big(Z_p^{prim}(V_p(t+1), t|D_p) - Z_p^{prim}(f_p^*(t+1), t|D_p)\Big) \\ &\quad + \mathcal{O}\Big(C^R \frac{\ln(\frac{d}{\delta})}{B_p \rho}\Big) \\ &\leq \frac{8\big(C^R\big)^2 d^2 \Big(\rho + c_4 C^R\Big)\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho^2 B_p^2 \alpha_p(t)^2} + \frac{8\eta^2 N_p^2 d^2 \big(\ln(\frac{d}{\delta})\big)^2}{\rho^3 B_p^2 \alpha_p(t)^2} \\ &\quad + \mathcal{O}\Big(C^R \frac{\ln(\frac{1}{\delta})}{B_p \rho}\Big).\end{aligned}$$

Combining the above two inequalities, we have the probability no smaller than $1 - 3\delta$.

Since $\hat{f}_p(t+1) = \arg\min_{f_p} \hat{Z}_p(f_p, t)$, then, $\big(\hat{Z}_p(\hat{f}_p(t+1), t) - \hat{Z}_p(f_p^0(t+1), t) \leq 0$. For the last two terms, we select $\rho = \frac{\alpha_{acc}}{\| f_p^0(t+1) \|^2}$ to make them bounded by $\frac{\alpha_{acc}}{2}$.

The value of B_p is determined by solving

$$\frac{8\big(C^R\big)^2 d^2 \Big(\rho + c_4 C^R\Big)\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho^2 B_p^2 \alpha_p(t)^2} + \frac{8\eta^2 N_p^2 d^2 \big(\ln(\frac{d}{\delta})\big)^2}{\rho^3 B_p^2 \alpha_p(t)^2} + \mathcal{O}\Big(C^R \frac{\ln(\frac{1}{\delta})}{B_p \rho}\Big) + \frac{\alpha_{acc}}{2} = \alpha_{acc},$$

with $\rho = \frac{\alpha_{acc}}{\| f_p^0(t+1) \|^2}$, such that $\mathbb{P}\big(\hat{C}(V_p^*(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc}\big) \geq 1 - 3\delta$. However, the accuracy of $V_p^*(t+1)$ depends on $f_p^*(t+1)$, thus we also have to make $\mathbb{P}\big(\hat{C}(f_p^*(t+1)) \leq \hat{C}^0(t+1) + \alpha_{acc}\big) \geq 1 - 2\delta$. Combining the result of Theorem 5, we arrive at (26). □

Lemma 20: Assume $R(f_p(t))$ is doubly differentiable w.r.t. $f_p(t)$ with $\parallel \nabla^2 R(f_p(t)) \parallel \leq \tau$ for all $f_p(t)$. Suppose the loss function $\mathscr{L}$ is differentiable, $\mathscr{L}'$ is continuous, and satisfies $|\mathscr{L}'(a) - \mathscr{L}'(b)| \leq c_4|a - b|$ for all pairs (a, b) with a constant c_4. Let $f_p^*(t+1) = \arg\min_{f_p} Z_p^{prim}(f_p, t|D_p)$, and $V_p^*(t+1) = f_p^*(t+1) + \epsilon_p(t)$, where the noise vector $\epsilon_p(t)$ is drawn from (15) with the same $\alpha_p(t)$ for all $p \in \mathscr{P}$ at time t. Let Λ be the event

$$\Lambda := \Big\{Z_p^{prim}(V_p^*(t+1), t|D_p) \leq Z_p^{prim}(f_p^*(t+1), t|D_p) + \psi\Big\}$$

where $\psi = \frac{4\big(C^R\big)^2 d^2 \big(\rho\tau + c_4 C^R\big)\big(\ln(\frac{d}{\delta})\big)^2}{\rho^2 B_p^2 \alpha_p(t)^2}$. Under Assumption 1 and 2, we have: $\mathbb{P}_{\epsilon_p(t)}\Big(\Lambda\Big) \geq 1 - \delta$. The probability $\mathbb{P}_{\epsilon_p(t)}$ is taken over the noise vector $\epsilon_p(t)$.

Proof (Lemma 20): From Assumption 3, we know that the data points in dataset D_p satisfy: $\parallel x_{ip} \parallel \leq 1$, and $|y_{ip}| = 1$. From Assumption 1 and 2, $R(\cdot)$ and $\mathscr{L}$ are differentiable. Suporse $R(\cdot)$ is doubly differentiable and $\nabla^2 R(\cdot) \leq \tau$. Let $0 \leq \varphi \leq 1$, then, the *Mean Value Theorem* and Cauchy-Schwarz inequality give:

$$\begin{aligned}&Z_p^{prim}(V_p^*(t+1), t|D_p) - Z_p^{prim}(f_p^*(t+1), t|D_p) \\ &= (V_p^*(t+1) - f_p^*(t+1))^T \nabla Z_p^{prim}\Big(\varphi f_p^*(t+1) \\ &\quad + (1-\varphi)V_p^*(t+1)\Big) \leq \parallel V_p^*(t+1) - f_p^*(t+1) \parallel \\ &\quad \cdot \parallel \nabla Z_p^{prim}\Big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)\Big) \parallel.\end{aligned}$$

Let $\epsilon^{pi}(t) = \epsilon_p(t) - \epsilon_i(t)$. From the definition of $Z_p^{prim}(f_p, t|D_p)$, we have:

$$\begin{aligned}{}_p^{prim}(f_p, t|D_p) &= Z_p(f_p|D_p) \\ &\quad - \eta \sum_{i \in \mathscr{N}_p} \Big((f_p - \frac{1}{2}(f_p(t) + f_i(t))^T \\ &\quad \cdot (\epsilon^{pi}(t)) + \frac{1}{4}(\epsilon^{pi}(t))^2\Big).\end{aligned}$$

Taking the derivative of Z_p^{prim} w.r.t. f_p gives

$$\begin{aligned}\nabla Z_p^{prim}(f_p, t|D_p) &= \frac{C^R}{B_p} \sum_{i=1}^{B_p} y_{ip} \mathscr{L}'(y_{ip} f_p^T x_{ip}) x_{ip} \\ &\quad + \rho \nabla R(f_p) - \eta \sum_{j \in \mathscr{N}_p} \epsilon^{pi}(t).\end{aligned}$$

Since $\nabla Z_p^{prim}(f_p^*(t+1), t|D_p) = 0$, then, we have:

$$\begin{aligned}&\nabla Z_p^{prim}\Big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)|D_p\Big) \\ &= \nabla Z_p^{prim}(f_p^*(t+1), t|D_p) - \rho\Big(\nabla R(f_p^*(t+1)) \\ &\quad - \nabla R\big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)\big)\Big) \\ &\quad - \frac{C^R}{B_p} \sum_{i=1}^{B_p} \Big(y_{ip}\Big(\mathscr{L}'(y_{ip} f_p^*(t+1)^T x_{ip}) \\ &\quad - \mathscr{L}'(y_{ip}\big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)\big)^T x_{ip})\Big) x_{ip}\Big).\end{aligned}$$

Let

$$\begin{aligned}T &= y_{ip}\Big(\mathscr{L}'(y_{ip} f_p^*(t+1)^T x_{ip}) \\ &\quad - \mathscr{L}'(y_{ip}\big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)\big)^T x_{ip})\Big) x_{ip}.\end{aligned}$$

Based on the condition on the loss function: $|\mathscr{L}'(a) - \mathscr{L}'(b)| \leq c_4|a-b|$, we can bound T as follows:

$$\begin{aligned}
T \leq\ & |y_{ip}| \parallel x_{ip} \parallel \\
& \cdot|\mathscr{L}'(y_{ip} f_p^*(t+1)^T x_{ip}) \\
& -\mathscr{L}'(y_{ip}\big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)\big)^T x_{ip})| \\
\leq\ & |y_{ip}| \parallel x_{ip} \parallel \cdot c_4 \cdot |y_{ip}(1-\varphi)(f_p^*(t+1) - V_p^*(t+1))^T x_{ip}| \\
\leq\ & c_4 \cdot (1-\varphi)|y_{ip}|^2 \parallel x_{ip} \parallel^2 \parallel f_p^*(t+1) - V_p^*(t+1) \parallel \\
\leq\ & c_4 \cdot (1-\varphi) \parallel f_p^*(t+1) - V_p^*(t+1) \parallel .
\end{aligned}$$

Since we assume $R(\cdot)$ is doubly differentiable, we then apply the *Mean Value Theorem*:

$$\begin{aligned}
& \parallel \nabla R(f_p^*(t+1)) - \nabla R\big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)\big) \parallel \\
& \quad \leq (1-\varphi) \parallel f_p^*(t+1) - V_p^*(t+1) \parallel \cdot \parallel \nabla^2 R(\xi) \parallel,
\end{aligned}$$

where $\xi \in \mathbb{R}^d$. Therefore, we have

$$\begin{aligned}
& \nabla Z_p^{prim}\Big(\varphi f_p^*(t+1) + (1-\varphi)V_p^*(t+1)|D_p\Big) \\
& \quad \leq (1-\varphi) \parallel f_p^*(t+1) - V_p^*(t+1) \parallel \cdot\rho\cdot \parallel \nabla^2 R(\xi) \parallel \\
& \qquad + C^R c_4 \cdot (1-\varphi) \parallel f_p^*(t+1) - V_p^*(t+1) \parallel \\
& \quad \leq (1-\varphi)\cdot \parallel f_p^*(t+1) - V_p^*(t+1) \parallel \Big(\rho\tau + C^R c_4\Big) \\
& \quad \leq \parallel f_p^*(t+1) - V_p^*(t+1) \parallel \Big(\rho\tau + C^R c_4\Big).
\end{aligned}$$

Since $f_p^*(t+1) - V_p^*(t+1) = \epsilon_p(t)$, with density $\Gamma(d, \frac{2C^R}{\rho B_p \alpha_p(t)})$ then, we can apply Lemma 10 to $\parallel f_p^*(t+1) - V_p^*(t+1) \parallel$. Thus, with $\parallel f_p^*(t+1) - V_p^*(t+1) \parallel \leq \frac{2C^R d \ln(\frac{d}{\delta})}{\rho B_p \alpha_p(t)}$, we have:

$$\begin{aligned}
& Z_p^{prim}(V_p^*(t+1), t|D_p) - Z_p^{prim}(f_p^*(t+1), t|D_p) \\
& \qquad \leq \frac{4\big(C^R\big)^2 d^2 \Big(\rho\tau + c_4 C^R\Big)\Big(\ln(\frac{d}{\delta})\Big)^2}{\rho^2 B_p^2 \alpha_p(t)^2},
\end{aligned}$$

with probability no less than $1-\delta$. □

References

[1] M. Li *et al.*, "Scaling distributed machine learning with the parameter server," in *Proc. OSDI*, 2014, pp. 583–598.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory Cryptography*, Berlin, Germany: Springer, 2006, pp. 265–284.

[3] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proc. 39th Annu. ACM Symp. Theory Comput.*, 2007, pp. 75–84.

[4] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[5] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Res.*, vol. 11, pp. 1663–1707, Jan. 2010.

[6] M. Collins, "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *Proc. Conf. Empirical Methods Natural Lang. Process*, 2002, pp. 1–8.

[7] R. McDonald, K. Hall, and G. Mann, "Distributed training strategies for the structured perceptron," in *Proc. Human Lang. Technol. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 456–464.

[8] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," *Inf. Syst.*, vol. 29, no. 4, pp. 343–364, Jun. 2004.

[9] J. Kim and W. Winkler, "Multiplicative noise for masking continuous data," *Statistics*, 2003, pp. 1–6.

[10] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[11] 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, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1137/090756090

[12] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci.*, 2007, pp. 94–103.

[13] A. Blum, C. Dwork, and K. Nissim, "Practical privacy: The sulq framework," in *Proc. 24th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, 2005, pp. 128–138.

[14] R. Chen, A. Reznichenko, P. Francis, and J. Gehrke, "Towards statistical queries over distributed private user data," presented at the 9th USENIX Symp. Netw. Syst. Design Implement. (NSDI), 2012, pp. 169–182.

[15] I. Mironov, O. Pandey, O. Reingold, and S. Vadhan, "Computational differential privacy," in *Advances in Cryptology - CRYPTO*, Berlin, Germany: Springer, 2009, pp. 126–142.

[16] F. Eigner and M. Maffei, "Differential privacy by typing in security protocols," in *Proc. IEEE 26th Comput. Secur. Found. Symp. (CSF)*, 2013, pp. 272–286.

[17] Y.-H. Dai, "A perfect example for the bfgs method," *Math. Program.*, vol. 138, no. 1, pp. 501–530, Apr. 2013.

[18] S. Shalev-Shwartz and N. Srebro, "Svm optimization: Inverse dependence on training set size," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 928–935.

[19] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Res.*, vol. 12, pp. 1069–1109, Mar. 2011.

[20] A. Asuncion and D. Newman, "UCI machine learning repository," Dept. School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, Tech. Rep., 2007.

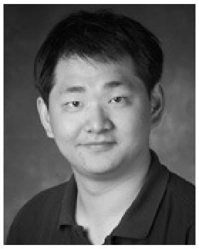
[21] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy*, 2008, pp. 111–125.

[22] P. Chilstrom, "Singular value inequalities: New approaches to conjectures," M.S. thesis, Dept. Math. Statist. Univ. North Florida, Jacksonville, FL. USA, 2013.

[23] S. Shalev-Shwartz and Y. Singer, *Online Learning: Theory, Algorithms, and Applications*. Jerusalem, Israel: Senate of the Hebrew Univ., 2007.

[24] K. Sridharan, S. Shalev-Shwartz, and N. Srebro, "Fast rates for regularized objectives," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1545–1552.

Tao Zhang received the master's degree in electrical engineering from New York University in 2016, and the bachelor's degree in electrical and electronic engineering from the Imperial College of Science, Technology and Medicine (Imperial College London), London, U.K., in 2013. His research interests include machine learning, game theory and application, and decision theory for cyber security.



Quanyan Zhu (S'04–M'12) received the B.Eng. degree (Hons.) in electrical engineering from McGill University in 2006, the M.A.Sc. degree from the University of Toronto in 2008, and the Ph.D. degree from the University of Illinois at Urbana–Champaign in 2013. He was with Princeton University. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, New York University. His current research interests include resilient and secure cyber-physical systems, adversarial signal processing, and interdependent networks. He is a recipient of the best paper awards at the 5th International Conference on Resilient Control Systems, the 18th International Conference on Information Fusion, and the 7th ACM CCS International Workshop on Managing Insider Security Threats. He spearheaded and chaired INFOCOM Workshop on Communications and Control on Smart Energy Systems, Midwest Workshop on Control and Game Theory, and the 7th Game and Decision Theory for Cyber Security.