DP-ADMM: ADMM-based Distributed Learning with Differential Privacy

Zonghao Huang, *Student Member, IEEE,* Rui Hu, *Student Member, IEEE,* Eric Chan-Tin, *Member, IEEE,* and Yanmin Gong, *Member, IEEE*

***Abstract*—Privacy-preserving distributed machine learning has become more important than ever due to the high demand of large-scale data processing. This paper focuses on a class of machine learning problems that can be formulated as regularized empirical risk minimization, and develops a privacy-preserving learning approach to such problems. We use Alternating Direction Method of Multipliers (ADMM) to decentralize the learning algorithm, and apply Gaussian mechanisms to provide differential privacy guarantee. However, simply combining ADMM and local randomization mechanisms would result in a nonconvergent algorithm with poor performance even under moderate privacy guarantees. Besides, this intuitive approach requires a strong assumption that the objective functions of the learning problems should be differentiable and strongly convex. To address these concerns, we propose an improved ADMM-based Differentially Private distributed learning algorithm, DP-ADMM, where an approximate augmented Lagrangian function and Gaussian mechanisms with time-varying variance are utilized. We also apply the moments accountant method to bound the total privacy loss. Our theoretical analysis shows that DP-ADMM can be applied to a general class of convex learning problems, provides differential privacy guarantee, and achieves a convergence rate of $O(1/\sqrt{t})$, where t is the number of iterations. Our evaluations demonstrate that our approach can achieve good convergence and accuracy with moderate privacy guarantee.**

***Index Terms*—Machine learning, ADMM, distributed computation, privacy, differential privacy, moments accountant, convergence.**

I. INTRODUCTION

DISTRIBUTED machine learning is a widely adopted and deployed approach due to high demand of large scale data processing. It allows multiple data providers to keep their datasets unexposed, and meanwhile to collaborate in one learning objective by iterative local computations and result exchanges with the global trainer. Thus, distributed machine learning can provide a degree of data privacy, help reduce the computational burden of the trainer, and improve the scalability of data processing.

One popular algorithm enabling distributed learning is Alternating Direction Method of Multipliers (ADMM) [1]. By ADMM, the learning problem is divided into several subproblems solved by each data provider independently and locally, and only intermediate parameters need to be shared with the trainer. However, data privacy risk still exists in such a setting. Recent works show that the adversary could obtain the sensitive information from the shared intermediate parameters by membership inference attacks [2] or model inversion attacks [3]. Thus, we need to consider a stronger privacy model in an ADMM-based distributed learning framework. Some previous works apply partially homomorphic cryptography in ADMM-based distributed learning [4]. However, this has a huge computational burden, leading to undesired high delays in communication. Another line of previous works employ randomization mechanisms in ADMM to guarantee differential privacy [5]–[7]. One common randomization method is to add noise to the output, which would disrupt the learning process and degrade the performance of the trained model, especially when a low total privacy loss is expected. Besides, in existing works on ADMM-based distributed learning with differential privacy [8]–[10], their privacy-preserving algorithms only apply to the learning problems with both differentiability and strongly convexity. Such weaknesses and limitations motivate us to explore further in this area.

In our paper, we focus on a class of regularized empirical risk minimization problems, and propose an improved ADMM-based Differentially Private distributed learning algorithm: DP-ADMM, which has fast convergence and good performance, and can be applied to a general class of convex learning problems. In our framework, we employ Gaussian mechanisms locally to guarantee (ϵ, δ)-differential privacy in each iteration, and employ the moments accountant method [11] to bound the total privacy loss. The key algorithmic feature of our approach is the adoption of an approximate augmented Lagrangian function and Gaussian mechanisms with time-varying variance, which enforce the algorithm to be noise-resistant and convergent, and have a bounded l_2 sensitivity. Then we analyze theoretically the privacy guarantee and the convergence of our proposed algorithm, and demonstrate that our privacy-preserving algorithm can be applied to the convex learning problems even with a non-differentiable objective function, and achieves an $O(1/\sqrt{t})$ rate of convergence, where t is the number of iterations.

Our contributions are summarized as follows:

1) We propose a differentially private distributed learning algorithm based on ADMM: DP-ADMM, where an approximate augmented Lagrangian function and Gaussian mechanisms with time-varying variance are adopted. Our proposed algorithm has low computation cost, fast convergence, and good performance.
2) We provide the privacy guarantee theorem of the proposed algorithm, proving that our approach guarantees differential privacy and it can be applied in any convex

Z. Huang, R. Hu, and Y. Gong are with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74075. (e-mail: zonghao.huang@okstate.edu; rui.hu@okstate.edu; yanmin.gong@okstate.edu)

E. Chan-Tin is with the Department of Computer Science, Loyola University Chicago, Chicago, IL 60660. (e-mail: chantin@cs.luc.edu)

learning problems. We adopt the moments accountant method to bound the total privacy leakage.

3) We give two convergence theorems of our approach under a weak assumption and a strong one respectively. Our convergence analysis shows that the proposed algorithm achieves an $O(1/\sqrt{t})$ rate of convergence.
4) Extensive experiments are carried out on real-world datasets to evaluate the accuracy and effectiveness of our algorithm in the setting with moderate total privacy loss.

The rest of our paper is organized as follows. In Section II, we give the problem statement. In Section III, we propose our approach: DP-ADMM. In Section IV and Section V, we analyze the privacy guarantee and convergence of the proposed algorithm respectively. In Section VI, we evaluate the proposed algorithm based on real-world datasets. In Section VII and VIII, we discuss the related works and conclude our work respectively.

II. Problem Statement

In this section, we first introduce the problem setting and the learning objective. Then we present a non-private distributed learning algorithm based on ADMM, and discuss the privacy concern of the distributed learning framework. Here we give a summary of notations used in this paper in Table I.

TABLE I: Notations Used in This Paper

$\mathbf{w}$	Global primal variable of classifier
$\ell(\cdot)$	Loss function
$R(\cdot)$	Regularizer
$\mathcal{D}_i$	Dataset of data provider i
$\mathbf{a}_{i,j}$	Feature vector
$b_{i,j}$	Label of $\mathbf{a}_{i,j}$
λ	Regularizer parameter
$\ell^{'}(\cdot)$	Subgradient of loss function
$R^{'}(\cdot)$	Subgradient of regularizer
$\mathbf{w}_i$	Local primal variable from data provider i
$\boldsymbol{\gamma}_i$	Local dual variable from data provider i
ρ	Penalty constant
$\mathcal{L}_\rho(\cdot)$	Augmented Lagrangian function
$\hat{\mathcal{L}}_{\rho,k}(\cdot)$	Approximate augmented Lagrangian function
$\mathbf{w}_i^k$	Primal variable from data provider i in k^{th} iteration
$\tilde{\mathbf{w}}_i^k$	Noisy version of $\mathbf{w}_i^k$ after perturbation
$\boldsymbol{\gamma}_i$	Dual variable from data provider i in k^{th} iteration
$\mathbf{w}^k$	Global primal variable in k^{th} iteration
ξ_i^k	Sampled noise from data provider i in k^{th} iteration
σ_i^2	Constant variance of Gaussian mechanism
η_i^k	Time-varying step size in k^{th} iteration
$\sigma_{i,k+1}^2$	Time-varying variance of Gaussian mechanism
D_w	L_2-norm of the optimal primal variable
$\mathbf{w}^*$	Optimal primal variable
$\nabla\ell(\cdot)$	Derivative of $\ell(\cdot)$
$\nabla R(\cdot)$	Derivative of $R(\cdot)$
$\nabla^2\ell(\cdot)$	Second-order derivative of $\ell(\cdot)$
$\nabla^2 R(\cdot)$	Second-order derivative of $R(\cdot)$
$\mathcal{D}_i^{'}$	Neighbouring dataset of $\mathcal{D}_i$

A. Problem Setting

We consider N data providers indexed by $1, 2, \ldots, N$ and a central trainer. Let $[N]$ denote the set: $\{1, 2, \ldots, N\}$. Furthermore, each data provider owns a private dataset $\mathcal{D}_i : \{(\mathbf{a}_{i,j}, b_{i,j}) :\subseteq (\mathcal{A} \times \mathcal{B})\}_{j\in[m_i]}$, where $\mathbf{a}_{i,j} \in \mathcal{A} \subseteq \mathbb{R}^d$ represents the data feature vector and $b_{i,j} \in \mathcal{B} \subseteq \{\pm 1\}$ denotes the corresponding label. Our paper only focuses on the case where the trainer and each data provider can communicate, and there is no connection between any two data providers, but our work can be extended to any connected case.

The target of our problem is to train a classifier: $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ based on $\{\mathcal{D}_i\}_{i\in[N]}$, which enables associating any given data feature vector with an accurate label. In order to train such a classifier, we model the problem as a regularized empirical risk minimization problem:

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\mathbf{a}_{i,j}, b_{i,j}, \mathbf{w}) + \lambda R(\mathbf{w}),$$

where $\ell(\cdot) : \mathbb{R}^d \times \{\pm 1\} \times \mathbb{R}^d \to \mathbb{R}$ is the loss function used to measure the quality of the trained classifier, $R(\cdot)$ refers to the regularizer introduced to prevent overfitting, and $\lambda > 0$ is the parameter controlling the impact of regularizer. In this paper, we have the following basic assumption related to our learning problem: the loss function $\ell(\cdot)$ and the regularizer $R(\cdot)$ are both convex but not necessarily differentiable. We use $\ell^{'}(\cdot)$ and $R^{'}(\cdot)$ to denote the subgradients of $\ell(\cdot)$ and $R(\cdot)$, and use $\nabla\ell(\cdot)$ and $\nabla R(\cdot)$ when they are differentiable.

B. Non-Private ADMM-Based Distributed Learning Algorithm

In a distributed setting, the learning problem is reformulated as follows:

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{N}\left(\sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\mathbf{a}_{i,j}, b_{i,j}, \mathbf{w}_i) + \frac{\lambda}{N}R(\mathbf{w}_i)\right), \tag{1a}$$

$$\text{s.t.} \quad \mathbf{w}_i = \mathbf{w}, i = 1, \ldots, N, \tag{1b}$$

where $\mathbf{w}_i \in \mathcal{W} \subseteq \mathbb{R}^d$ is the local primal variable and $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ is the global one. The objective function (1a) is decoupled and each data provider only needs to minimize the sub-problem associated with his dataset. Constraints (1b) enforce that all the local classifiers reach consensus finally.

We solve the learning problem in a distributed manner via the classical ADMM, which solves convex optimization problems by breaking them into smaller pieces that are easier to handle individually. The augmented Lagrangian function associated with the problem (1) is:

$$\mathcal{L}_\rho(\mathbf{w}_i, \mathbf{w}, \boldsymbol{\gamma}_i) = \sum_{i=1}^{N}\left(\sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\mathbf{a}_{i,j}, b_{i,j}, \mathbf{w}_i) + \frac{\lambda}{N}R(\mathbf{w}_i) - \langle\boldsymbol{\gamma}_i, \mathbf{w}_i - \mathbf{w}\rangle + \frac{\rho}{2}\|\mathbf{w}_i - \mathbf{w}\|^2\right),$$

where $\boldsymbol{\gamma}_i \in \mathbb{R}$ is the local dual variable and $\rho \in \mathbb{R}$ is the pre-defined penalty constant. ADMM solves the problem in

a Gauss-Seidel manner: minimizing $\mathcal{L}_\rho(\mathbf{w}_i, \mathbf{w}, \boldsymbol{\gamma}_i)$ w.r.t. $\mathbf{w}_i$, $\boldsymbol{\gamma}_i$, and $\mathbf{w}$ alternatively given others fixed, which is shown in Algorithm 1.

Algorithm 1 Distributed Algorithm Based on ADMM

Initialize $\mathbf{w}^0$, $\{\mathbf{w}_i^0\}_{i\in[N]}$, and $\{\boldsymbol{\gamma}_i^0\}_{i\in[N]}$.
for $k = 0, 1, 2, \ldots, T-1$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    $\mathbf{w}_i^{k+1} \leftarrow \operatorname{argmin}_{\mathbf{w}_i} \mathcal{L}_\rho(\mathbf{w}_i, \mathbf{w}^k, \boldsymbol{\gamma}_i^k)$.
  end for
  $\mathbf{w}^{k+1} \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbf{w}_i^{k+1} - \frac{1}{N}\sum_{i=1}^N \boldsymbol{\gamma}_i^k/\rho$.
  for $i = 1, 2, \ldots, N$ **do**
    $\boldsymbol{\gamma}_i^{k+1} \leftarrow \boldsymbol{\gamma}_i^k - \rho(\mathbf{w}_i^{k+1} - \mathbf{w}^{k+1})$.
  end for
end for

C. Privacy Concern

Although the individual dataset is kept local in Algorithm 1, the intermediate parameters $\{\mathbf{w}_i^k\}_{i\in[N],k\in[T]}$ from local training need to be exposed to the trainer, which may reveal data providers' private information. Thus, we need to employ additional privacy-preserving methods to control such information leakage.

In this paper, one of our goals is to provide privacy guarantee against inference attack from a strong adversary with arbitrary knowledge, who tries to infer some sensitive information about data providers from the shared messages. We assume that the adversary cannot intrude into the local datasets and have access to the data information directly. The adversary could be an outsider who eavesdrops the shared messages, or the honest-but-curious trainer who follows the protocol honestly but tends to infer the sensitive information curiously. We do not assume any trusted third party, thus a privacy-preserving mechanism should be applied locally to provide the privacy guarantee.

In order to provide privacy guarantee against such attack, we define our privacy model formally by differential privacy. Differential privacy is a strong definition of privacy, providing the protection of the dataset from attackers with arbitrary knowledge. There are two definitions including pure differential privacy: ϵ-differential privacy and relaxed differential privacy: (ϵ, δ)-differential privacy. Compared with pure differential privacy, (ϵ, δ)-differential privacy is used for the analysis of the privacy guarantee of Gaussian mechanisms [12], by which generated noise has the same distribution model as the natural noise. Especially, (ϵ, δ)-differential privacy is preferred in the application of advanced composition theorem, for example, the adaptive algorithm, which takes the output of previous step as the input of the current step. The (ϵ, δ)-differential privacy is defined as follows:

Definition 1 ((ϵ, δ)-Differential Privacy)**.** *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{O}$ satisfies (ϵ, δ)-differential privacy if for any two neighbouring datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in only one tuple, and for all outputs $\mathcal{O} \in$ range($\mathcal{M}$), the following inequality always holds:*

$$Pr[\mathcal{M}(\mathcal{D}) = \mathcal{O}] \leq e^\epsilon \cdot Pr[\mathcal{M}(\mathcal{D}') = \mathcal{O}] + \delta,$$

which means, with probability of at least $1-\delta$, the ratio of probability distributions for two neighbouring datasets is bounded by e^ϵ.

In definition 1, δ and ϵ indicate the strength of privacy protection from the mechanism. With any given δ, a privacy-preserving mechanism with a smaller ϵ gives better privacy protection. Gaussian mechanism is a common randomization mechanism used to guarantee (ϵ, δ)-differential privacy.

III. Differentially private Algorithm

A. ADMM with Differential Privacy

As described in Section II, we require to consider a local privacy-preserving mechanism in order to guarantee (ϵ, δ)-differential privacy. One intuitive way to achieve this goal is to combine perturbation mechanism and ADMM directly. Specifically, as given in Algorithm 2, in the $(k+1)^{th}$ iteration, before sharing the local primal variable $\mathbf{w}_i^{k+1}$, we apply Gaussian mechanism with pre-defined variance σ_i^2 to add noise into the shared information to guarantee differential privacy. According to [7], [13], σ_i is defined by $\frac{2S_1 N\sqrt{2\ln(1.25/\delta)}}{m_i\lambda\epsilon}$ to guarantee (ϵ, δ)-differential privacy in each iteration. A similar approach has been adopted by Zhang and Zhu [8].

Algorithm 2 Privacy-Preserving Algorithm Based on ADMM

Initialize $\mathbf{w}^0$, $\{\mathbf{w}_i^0\}_{i\in[N]}$, and $\{\boldsymbol{\gamma}_i^0\}_{i\in[N]}$.
for $k = 0, 1, 2, \ldots, T-1$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    $\mathbf{w}_i^{k+1} \leftarrow \operatorname{argmin}_{\mathbf{w}_i} \mathcal{L}_\rho(\mathbf{w}_i, \mathbf{w}^k, \boldsymbol{\gamma}_i^k)$.
    $\tilde{\mathbf{w}}_i^{k+1} \leftarrow \mathbf{w}_i^{k+1} + \mathcal{N}(0, \sigma_i^2)$.
  end for
  $\mathbf{w}^{k+1} \leftarrow \frac{1}{N}\sum_{i=1}^N \tilde{\mathbf{w}}_i^{k+1} - \frac{1}{N}\sum_{i=1}^N \boldsymbol{\gamma}_i^k/\rho$.
  for $i = 1, 2, \ldots, N$ **do**
    $\boldsymbol{\gamma}_i^{k+1} \leftarrow \boldsymbol{\gamma}_i^k - \rho(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^{k+1})$.
  end for
end for

However, directly applying perturbation would disrupt the training process and lead to a trained model with poor performance. This is especially the case when the introduced noise is large to guarantee a low total privacy leakage. Specifically, when the iteration number k is large, the result would still change dramatically due to the existence of noise. Besides, such perturbation method can only be applied when both the loss function and the regularizer are differentiable, and the regularizer is strongly convex [13]. When we consider any non-differentiable convex learning problems, this method does not work. In order to address such problems, we need to consider a better way to introduce randomness into ADMM.

B. Our Approach

Our approach is inspired by the idea that it is not necessary to solve the problem up to a very high precision in each iteration in order to guarantee the overall convergence. In this approach, instead of the augmented Lagrangian function, we employ its first order approximation with a scalar l_2-norm prox-function:

$$\begin{aligned}&\hat{\mathcal{L}}_{\rho,k}(\mathbf{w}_i, \tilde{\mathbf{w}}_i^k, \mathbf{w}^k, \boldsymbol{\gamma}_i^k)\\ =&\sum_{i=1}^{N}\Bigg(\sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\mathbf{a}_{i,j}, b_{i,j}, \tilde{\mathbf{w}}_i^k) + \frac{\lambda}{N}R(\tilde{\mathbf{w}}_i^k)\\ &+\big\langle\sum_{j=1}^{m_i}\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,j}, b_{i,j}, \tilde{\mathbf{w}}_i^k) + \frac{\lambda}{N}R^{'}(\tilde{\mathbf{w}}_i^k), \mathbf{w}_i - \tilde{\mathbf{w}}_i^k\big\rangle\\ &-\big\langle\boldsymbol{\gamma}_i^k, \mathbf{w}_i - \mathbf{w}^k\big\rangle + \frac{\rho}{2}\|\mathbf{w}_i - \mathbf{w}^k\|^2 + \frac{\|\mathbf{w}_i - \tilde{\mathbf{w}}_i^k\|^2}{2\eta_i^{k+1}}\Bigg),\end{aligned} \tag{2}$$

where $\eta_i^{k+1} \in \mathbb{R}$ is a time-varying step size, and decreases with increasing k. We minimize (2) in a Gauss-Seidel manner, and also introduce noise sampled from Gaussian mechanism with a time-varying variance $\sigma_{i,k+1}^2$ that also decreases with increasing k. The corresponding ADMM steps that provide differential privacy are as follows:

$$\mathbf{w}_i^{k+1} = \underset{\mathbf{w}_i}{\operatorname{argmin}}\ \hat{\mathcal{L}}_{\rho,k}(\mathbf{w}_i, \tilde{\mathbf{w}}_i^k, \mathbf{w}^k, \boldsymbol{\gamma}_i^k), \tag{3a}$$

$$\tilde{\mathbf{w}}_i^{k+1} = \mathbf{w}_i^{k+1} + \mathcal{N}(0, \sigma_{i,k+1}^2), \tag{3b}$$

$$\mathbf{w}^{k+1} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{w}}_i^{k+1} - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\gamma}_i^k/\rho, \tag{3c}$$

$$\boldsymbol{\gamma}_i^{k+1} = \boldsymbol{\gamma}_i^k - \rho(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^{k+1}), \tag{3d}$$

where (3c) is done in the central trainer while (3a), (3b), and (3d) are done by each data provider.

The details are given in Algorithm 3. The central trainer firstly initializes the global primal variable $\mathbf{w}^0$, and the data providers also initializes their noisy local primal variables $\{\tilde{\mathbf{w}}_i^0\}_{i\in[N]}$ and dual variables $\{\boldsymbol{\gamma}_i^0\}_{i\in[N]}$. The data providers firstly sample a noise ξ_i^{k+1} from the Gaussian mechanism with variance $\sigma_{i,k+1}^2$ and update the noisy local primal variables $\{\tilde{\mathbf{w}}_i^{k+1}\}_{i\in[N]}$ based on (3a) and (3b). Then the trainer receives the noisy local primal variables $\{\tilde{\mathbf{w}}_i^{k+1}\}_{i\in[N]}$ and the local dual variables $\{\boldsymbol{\gamma}_i^k\}_{i\in[N]}$ from data providers, and uses them for the update of the global primal variable $\mathbf{w}^{k+1}$ according to (3c). Data providers receive the updated global primal variable $\mathbf{w}^{k+1}$ from the trainer and continue to update the local dual variables $\{\boldsymbol{\gamma}_i^{k+1}\}_{i\in[N]}$. The data providers and the trainer repeatedly exchange variables until the end of their communication.

This approach is different from Algorithm 2 in three perspectives. Firstly, the approximate augmented Lagrangian function (2) used in this approach replaces the objective function with its first-order approximation at $\tilde{\mathbf{w}}_i^k$, which is similar to the stochastic mirror descent [14]. This approximation enforces the differentiability of the Lagrangian function and benefits from the ease of solving (3a). Even when the objective function is non-differentiable, we can still get a close-form solution to (3a), which achieves fast computation. More importantly, this approximation can lead to a bounded l_2 sensitivity in differential privacy guarantee without the limitation that the objective function should be differentiable and strongly convex. Thus our approach can be applied to any convex problems.

Algorithm 3 DP-ADMM

Initialize $\mathbf{w}^0$, $\{\tilde{\mathbf{w}}_i^0\}_{i\in[N]}$, and $\{\boldsymbol{\gamma}_i^0\}_{i\in[N]}$.
for $k = 0, 1, 2, \ldots, T-1$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    $\mathbf{w}_i^{k+1} \leftarrow \operatorname{argmin}_{\mathbf{w}_i}\ \hat{\mathcal{L}}_{\rho,k}(\mathbf{w}_i, \tilde{\mathbf{w}}_i^k, \mathbf{w}^k, \boldsymbol{\gamma}_i^k)$.
    $\xi_i^{k+1} \leftarrow \mathcal{N}(0, \sigma_{i,k+1}^2)$.
    $\tilde{\mathbf{w}}_i^{k+1} \leftarrow \tilde{\mathbf{w}}_i^{k+1} + \xi_i^{k+1}$.
  end for
  $\mathbf{w}^{k+1} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{w}}_i^{k+1} - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\gamma}_i^k/\rho$.
  for $i = 1, 2, \ldots, N$ **do**
    $\boldsymbol{\gamma}_i^{k+1} \leftarrow \boldsymbol{\gamma}_i^k - \rho(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^{k+1})$.
  end for
end for

Secondly, similar to linearized ADMM [15], [16], there is an l_2-norm prox-function $\|\mathbf{w}_i - \tilde{\mathbf{w}}_i^k\|^2$ but scaled by $\frac{1}{2\eta_i^{k+1}}$ added in (2), where η_i^{k+1} decreases with increasing k. Such additional part can guarantee convergent updates: when k is larger, $\frac{1}{2\eta_i^{k+1}}$ increases, then updated model $\mathbf{w}_i^{k+1}$ is closer to the previous noisy model $\tilde{\mathbf{w}}_i^k$ as a result of $\|\mathbf{w}_i - \tilde{\mathbf{w}}_i^k\|^2$. Here, η_i^{k+1} is significant for the overall convergence guarantee. In Section V, we will define η_i^{k+1} and show its importance in algorithm convergence.

Lastly, the variance $\sigma_{i,k+1}^2$ of Gaussian mechanism used in Algorithm 3 is time-varying rather than constant. It decreases when k increases. Thus, the added noise will decrease to make the updates stable. We need to emphasize that the decreased noise will not make the privacy protection weaker because the sensitivity will also decrease due to $\frac{\|\mathbf{w}_i - \tilde{\mathbf{w}}_i^k\|^2}{2\eta_i^{k+1}}$. In Section IV, we will define $\sigma_{i,k+1}^2$ specifically, and prove the privacy guarantee of Algorithm 3.

IV. Privacy Guarantee

In this section, we analyze the privacy guarantee of our proposed algorithm (Algorithm 3). In Algorithm 3, the shared messages $\{\tilde{\mathbf{w}}_i^{k+1}\}_{k=0,1,\ldots,T-1}$ may reveal the sensitive information of data provider i. Thus, we need to demonstrate that Algorithm 3 guarantees differential privacy with outputs $\{\tilde{\mathbf{w}}_i^{k+1}\}_{k=0,1,\ldots,T-1}$. We firstly define the l_2 sensitivity of $\mathbf{w}_i^{k+1}$, then analyze the privacy leakage for each iteration, and finally accumulate them by the moments accountant method.

Here we define $\mathbf{w}_{i,\mathcal{D}_i}^{k+1}$ and $\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}$ by:

$$\mathbf{w}_{i,\mathcal{D}_i}^{k+1} = -\big(\sum_{j=1}^{m_i}\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,j},b_{i,j},\tilde{\mathbf{w}}_i^k)+\frac{\lambda}{N}R^{'}(\tilde{\mathbf{w}}_i^k) - \boldsymbol{\gamma}_i^k-\rho\mathbf{w}^k-\tilde{\mathbf{w}}_i^k/\eta_i^{k+1}\big)/\big(\rho+1/\eta_i^{k+1}\big),$$

$$\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1} = -\big(\sum_{j=1}^{m_i-1}\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,j},b_{i,j},\tilde{\mathbf{w}}_i^k)+\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,m_i}^{'},b_{i,m_i}^{'},\tilde{\mathbf{w}}_i^k) +\frac{\lambda}{N}R^{'}(\tilde{\mathbf{w}}_i^k)-\boldsymbol{\gamma}_i^k-\rho\mathbf{w}^k-\tilde{\mathbf{w}}_i^k/\eta_i^{k+1}\big)/\big(\rho+1/\eta_i^{k+1}\big),$$

which would be used in the following lemma, theorem, and proofs in this section. We can easily prove that $\mathbf{w}_{i,\mathcal{D}_i}^{k+1}$ and $\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}$ are the solutions to (3a) w.r.t. $\mathcal{D}_i$ and $\mathcal{D}_i^{'}$, by obtaining the derivative of $\hat{\mathcal{L}}_{\rho,k}(\mathbf{w}_i,\tilde{\mathbf{w}}_i^k,\mathbf{w}^k,\boldsymbol{\gamma}_i^k)$: $\nabla\hat{\mathcal{L}}_{\rho,k}(\mathbf{w}_i,\tilde{\mathbf{w}}_i^k,\mathbf{w}^k,\boldsymbol{\gamma}_i^k) = \sum_{j=1}^{m_i}\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,j},b_{i,j},\tilde{\mathbf{w}}_i^k)+\frac{\lambda}{N}R^{'}(\tilde{\mathbf{w}}_i^k)-\boldsymbol{\gamma}_i^k+\rho(\mathbf{w}_i-\mathbf{w}^k)+\frac{1}{\eta_i^{k+1}}(\mathbf{w}_i-\tilde{\mathbf{w}}_i^k)$ and letting $\nabla\hat{\mathcal{L}}_{\rho,k}(\mathbf{w}_i,\tilde{\mathbf{w}}_i^k,\mathbf{w}^k,\boldsymbol{\gamma}_i^k)=0$, since $\hat{\mathcal{L}}_{\rho,k}(\mathbf{w}_i,\tilde{\mathbf{w}}_i^k,\mathbf{w}^k,\boldsymbol{\gamma}_i^k)$ is a quadratic function w.r.t. $\mathbf{w}_i$ thus convex.

A. L_2 Sensitivity

We apply Gaussian mechanism to introduce noise, which is calibrated by the l_2-norm sensitivity. Compared with the Algorithm 2 and the related work in the past [8], [13], the derivation of the sensitivity in our proposed algorithm does not require to assume the differentiability and the strongly convexity of the objective function. This is benefited from the first-order approximation of Lagrangian function.

Lemma 1. *We assume that $\|\ell^{'}(\cdot)\| \le S_1$, the l_2-norm sensitivity of $\mathbf{w}_i^{k+1}$ is defined by:*

$$\triangle_{i,2} = \max_{\mathcal{D}_i,\mathcal{D}_i^{'}}\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\| = \frac{2S_1}{m_i(\rho+1/\eta_i^{k+1})}.$$

Proof. With $\mathbf{w}_{i,\mathcal{D}_i}^{k+1}$ and $\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}$, the l_2 sensitivity of $\mathbf{w}_i^{k+1}$ is:

$$\max_{\mathcal{D}_i,\mathcal{D}_i^{'}}\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\| = \frac{\max_{\mathcal{D}_i,\mathcal{D}_i^{'}}\big\|\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,m_i},b_{i,m_i},\tilde{\mathbf{w}}_i^k)-\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,m_i}^{'},b_{i,m_i}^{'},\tilde{\mathbf{w}}_i^k)\big\|}{\rho+1/\eta_i^{k+1}},$$

where $\mathcal{D}_i$ and $\mathcal{D}_i^{'}$ are neighbouring datasets. We assume that $\|\ell^{'}(\cdot)\|$ is bounded by S_1, the sensitivity of $\mathbf{w}_i^{k+1}$ can be expressed by:

$$\max_{\mathcal{D}_i,\mathcal{D}_i^{'}}\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\| = \frac{\max_{\mathcal{D}_i,\mathcal{D}_i^{'}}\big\|\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,m_i},b_{i,m_i},\tilde{\mathbf{w}}_i^k)-\frac{1}{m_i}\ell^{'}(\mathbf{a}_{i,m_i}^{'},b_{i,m_i}^{'},\tilde{\mathbf{w}}_i^k)\big\|}{\rho+1/\eta_i^{k+1}} = \frac{2S_1}{m_i(\rho+1/\eta_i^{k+1})}.$$

□

Compared with the sensitivity of $\mathbf{w}_i^{k+1}$ in Algorithm 2: $\frac{2S_1}{m_i\lambda/N}$, the sensitivity in DP-ADMM is much smaller especially when λ is small and N is large. This shows that DP-ADMM requires less noise to guarantee the same level of privacy. More importantly, the sensitivity for DP-ADMM is affected by the time-varying η_i^{k+1}. When we set η_i^{k+1} to decrease with increasing k, the sensitivity becomes smaller with larger k, then the noise added would be smaller when ϵ is fixed. Thus, the updates would be stable with large k in spite of the existence of the noise.

B. (ϵ,δ)-Differential Privacy Guarantee

In this section, we demonstrate that each iteration of Algorithm 3 guarantees (ϵ,δ)-differential privacy.

Theorem 1. *We assume that $\|\ell^{'}(\cdot)\| \le S_1$. Let $\epsilon \in (0,1]$ be arbitrary and let ξ_i^{k+1} be sampled from Gaussian mechanism with variance $\sigma_{i,k+1}^2$ where*

$$\sigma_{i,k+1} = \frac{2S_1\sqrt{2\ln(1.25/\delta)}}{m_i\epsilon(\rho+1/\eta_i^{k+1})},$$

then each iteration of Algorithm 3 guarantees (ϵ,δ)-differential privacy. Specifically, for any neighboring dataset $\mathcal{D}_i$ and $\mathcal{D}_i^{'}$, for any output $\tilde{\mathbf{w}}_i^{k+1}$, the following inequality always holds:

$$P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i) \le e^{\epsilon}\cdot P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i^{'})+\delta.$$

Proof. The privacy loss from $\tilde{\mathbf{w}}_i^{k+1}$ is calculated by:

$$\left|\ln\frac{P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i)}{P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i^{'})}\right| = \left|\ln\frac{P(\mathbf{w}_{i,\mathcal{D}_i}^{k+1}+\xi_i^{k+1})}{P(\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}+\xi_i^{'k+1})}\right| = \left|\ln\frac{P(\xi_i^{k+1})}{P(\xi_i^{'k+1})}\right|.$$

Since ξ_i^{k+1} and $\xi_i^{'k+1}$ are sampled from $\mathcal{N}(0,\sigma_{i,k+1}^2)$,

$$\begin{aligned}&\left|\ln\frac{P(\xi_i^{k+1})}{P(\xi_i^{'k+1})}\right| = \left|\frac{\|\xi_i^{k+1}\|^2-\|\xi_i^{'k+1}\|^2}{2\sigma_{i,k+1}^2}\right| \\ &= \left|\frac{\|\xi_i^{k+1}\|^2-\|\xi_i^{k+1}+(\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1})\|^2}{2\sigma_{i,k+1}^2}\right| \\ &= \left|\frac{2\xi_i^{k+1}\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\|+\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\|^2}{2\sigma_{i,k+1}^2}\right|.\end{aligned} \tag{5}$$

We assume that $\|\ell^{'}(\cdot)\| \le S_1$, the l_2-norm sensitivity can be calculated by:

$$\max_{\mathcal{D}_i,\mathcal{D}_i^{'}}\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\| = \frac{2S_1}{m_i(\rho+1/\eta_i^{k+1})}. \tag{6}$$

Thus, let $\sigma_{i,k+1} = \frac{2S_1\sqrt{2\ln(1.25/\delta)}}{m_i\epsilon(\rho+1/\eta_i^{k+1})}$, by combining (5) and (6), we have

$$\begin{aligned}&\left|\ln\frac{P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i)}{P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i^{'})}\right| = \left|\frac{2\xi_i^{k+1}\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\|+\|\mathbf{w}_{i,\mathcal{D}_i}^{k+1}-\mathbf{w}_{i,\mathcal{D}_i^{'}}^{k+1}\|^2}{2\sigma_{i,k+1}^2}\right| \\ &\le \left|\frac{\xi_i^{k+1}m_i(\rho+1/\eta_i^{k+1})+S_1}{4\ln(1.25/\delta)S_1/\epsilon^2}\right|.\end{aligned}$$

When $|\xi_i^{k+1}| \leq \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}$, $\big|\ln\frac{P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i)}{P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i')}\big|$ is bounded by ϵ. Next, we need to prove that $P\big[|\xi_i^{k+1}| > \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\big] \leq \delta$, which requires $P\big[\xi_i^{k+1} > \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\big] \leq \delta/2$. According to the tail bound of normal distribution $\mathcal{N}(0,\sigma_{i,k+1}^2)$:

$$P[\xi_i^{k+1} > r] \leq \frac{\sigma_{i,k+1}}{r\sqrt{2\pi}} e^{-r^2/2\sigma_{i,k+1}^2},$$

let $r = \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}$, we have:

$$\begin{aligned}&P\big[\xi_i^{k+1} > \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\big]\\ \leq&\frac{2\sqrt{2\ln(1.25/\delta)}}{(4\ln(1.25/\delta)-\epsilon)\sqrt{2\pi}}\exp\Big(-\frac{(4\ln(1.25/\delta)-\epsilon)^2}{8\ln(1.25/\delta)}\Big).\end{aligned} \tag{7}$$

When δ is small (≤ 0.01) and let $\epsilon \leq 1$, we have

$$\frac{\sqrt{2\ln(1.25/\delta)}2}{(4\ln(1.25/\delta)-\epsilon)\sqrt{2\pi}} \leq \frac{\sqrt{2\ln(1.25/\delta)}2}{(4\ln(1.25/\delta)-1)\sqrt{2\pi}} < \frac{1}{\sqrt{2\pi}}. \tag{8}$$

And since:

$$\begin{aligned}-\frac{(4\ln(1.25/\delta)-\epsilon)^2}{8\ln(1.25/\delta)} &\leq -\frac{(4\ln(1.25/\delta)-1)^2}{8\ln(1.25/\delta)}\\ &= -2\ln(1.25/\delta)+1-\frac{1}{8\ln(1.25/\delta)}\\ &< -2\ln(1.25/\delta)+\frac{8}{9}\\ &< \ln(\sqrt{2\pi}\frac{\delta}{2}),\end{aligned}$$

with (7) and (8), we have:

$$\begin{aligned}&P\big[\xi_i^{k+1} > \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\big]\\ \leq&\frac{\sqrt{2\ln(1.25/\delta)}2}{(4\ln(1.25/\delta)-\epsilon)\sqrt{2\pi}}\exp(-\frac{(4\ln(1.25/\delta)-\epsilon)^2}{8\ln(1.25/\delta)})\\ <&\frac{1}{\sqrt{2\pi}}\exp(\ln(\sqrt{2\pi}\frac{\delta}{2}))\\ =&\frac{\delta}{2}.\end{aligned}$$

So far we have proved: $P\big[\xi_i^{k+1} > \big(4\ln(1.25/\delta)S_1\big)/\big(\epsilon m_i(\rho+1/\eta_i^{k+1})\big) - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\big] \leq \delta/2$ thus $P\big[|\xi_i^{k+1}| > \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\big] \leq \delta$. We define:

$$\mathbb{A}_1 = \{\xi_i^{k+1} : |\xi_i^{k+1}| \leq \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\},$$

$$\mathbb{A}_2 = \{\xi_i^{k+1} : |\xi_i^{k+1}| > \frac{4\ln(1.25/\delta)S_1}{\epsilon m_i(\rho+1/\eta_i^{k+1})} - \frac{S_1}{m_i(\rho+1/\eta_i^{k+1})}\}.$$

Thus, we obtain the desired result:

$$\begin{aligned}P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i) =&P(\mathbf{w}_{i,\mathcal{D}_i}^{k+1}+\xi_i^{k+1} : \xi_i^{k+1} \in \mathbb{A}_1)\\ &+P(\mathbf{w}_{i,\mathcal{D}_i}^{k+1}+\xi_i^{k+1} : \xi_i^{k+1} \in \mathbb{A}_2)\\ <&e^{\epsilon}\cdot P(\mathbf{w}_{i,\mathcal{D}_i'}^{k+1}+\xi_i^{'k+1})+\delta\\ =&e^{\epsilon}\cdot P(\tilde{\mathbf{w}}_i^{k+1}|\mathcal{D}_i')+\delta.\end{aligned}$$

□

The privacy guarantee of our proposed algorithm does not rely on the assumption that the loss function and regularizer are differentiable, and the regularizer is strongly convex, which is different from previous works [8], [13].

C. Total Privacy Leakage

We have proved that each iteration of the proposed algorithm is (ϵ,δ)-differentially private. Here we focus on the total privacy leakage of our algorithm. Since Algorithm 3 is a T-fold adaptive algorithm, we follow the previous works [11], [17] and use the moments accountant method to analyze the total privacy leakage.

Theorem 2 (Advanced Composition Theorem)**.** *We assume that $\|\ell'(\cdot)\| \leq S_1$ and Algorithm 3 expires after T iterations. Let $\epsilon \in (0,1]$ be arbitrary and let ξ_i^{k+1} be sampled from Gaussian mechanism with variance $\sigma_{i,k+1}^2$ where*

$$\sigma_{i,k+1} = \frac{2S_1\sqrt{2\ln(1.25/\delta)}}{m_i\epsilon(\rho+1/\eta_i^{k+1})}.$$

Algorithm 3 guarantees $(\overline{\epsilon},\delta)$-differential privacy, where there exists such a constant c_1 that $\overline{\epsilon} = c_1\sqrt{T}\epsilon$.

Proof. See Appendix A. □

V. Convergence Analysis

In this section, we analyze the convergence of the proposed algorithm with specific η_i^k defined. Let $\mathbf{w}^*$ denote the optimal result of problem (1). Firstly, we give a general convergence analysis based on the basic assumption that the objective function is convex. Then, we make an extension under stricter assumption in order to consider the relation between the convergence and ϵ.

Here we define that:

$$D_w := \|\mathbf{w}^*\|,$$

$$f_i(\mathbf{w}_i) := \sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\mathbf{a}_{i,j},b_{i,j},\mathbf{w}_i)+\frac{\lambda}{N}R(\mathbf{w}_i),$$

$$\overline{\mathbf{w}}^t := \frac{1}{t}\sum_{k=1}^{t}\mathbf{w}^k, \quad \overline{\boldsymbol{\gamma}}_i^t := \frac{1}{t}\sum_{k=1}^{t}\boldsymbol{\gamma}_i^k, \quad \overline{\mathbf{w}}_i^t := \frac{1}{t}\sum_{k=0}^{t-1}\tilde{\mathbf{w}}_i^k,$$

$$\mathbf{u}_i^k := \begin{bmatrix}\tilde{\mathbf{w}}_i^k\\ \mathbf{w}^k\\ \boldsymbol{\gamma}_i^k\end{bmatrix}, \qquad \mathbf{u}_i := \begin{bmatrix}\mathbf{w}_i\\ \mathbf{w}\\ \boldsymbol{\gamma}_i\end{bmatrix}, \quad F(\mathbf{u}_i^{k+1}) := \begin{bmatrix}-\boldsymbol{\gamma}_i^{k+1}\\ \boldsymbol{\gamma}_i^{k+1}\\ \tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}\end{bmatrix},$$

which would be used in the following analysis.

Under both two assumptions, we show that Algorithm 3 achieves an $O(1/\sqrt{t})$ rate of convergence in terms

of both the objective value and the constraint violation: $\mathbb{E}\Big[\sum_{i=1}^{N}\Big(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}^*)+\beta\|\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\|\Big)\Big]$, where $\sum_{i=1}^{N}\Big(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}^*)\Big)$ represents the distance between the current objective value and the optimal value while $\sum_{i=1}^{N}\beta\|\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\|$ measures the difference between the local model and the global one. Thus $\mathbb{E}\Big[\sum_{i=1}^{N}\Big(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}^*)+\beta\|\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\|\Big)\Big]=0$ means that our training result converges to the optimal one and all local models reach consensus.

Before giving the theorem on the convergence of Algorithm 3, we firstly start with Lemma 2, which is used to obtain the upper bound of the first order approximation based on each iteration point.

Lemma 2. *We assume that $l(\cdot)$ is a convex differentiable function. $s\geq 0$ is a scalar. For any vector $\mathbf{x}\in\mathbb{R}^d$ and $\mathbf{y}\in\mathbb{R}^d$, we denote their Bregman divergence as $D(\mathbf{x},\mathbf{y})\equiv h(\mathbf{x})-h(\mathbf{y})-\langle\nabla h(\mathbf{y}),\mathbf{x}-\mathbf{y}\rangle$, where $h(\cdot)$ is a continuously-differentiable real-valued and strictly convex function. If we define:*

$$\mathbf{x}^*:=\operatorname*{argmin}_{\mathbf{x}} l(\mathbf{x})+sD(\mathbf{x},\mathbf{y}),$$

then

$$\langle\nabla l(\mathbf{x}^*),\mathbf{x}^*-\mathbf{x}\rangle\leq s[D(\mathbf{x},\mathbf{y})-D(\mathbf{x},\mathbf{x}^*)-D(\mathbf{x}^*,\mathbf{y})].$$

Proof. According to the optimality condition,

$$\langle\nabla l(\mathbf{x}^*)+s\nabla D(\mathbf{x}^*,\mathbf{y}),\mathbf{x}-\mathbf{x}^*\rangle\geq 0.$$

Then,

$$\begin{aligned}\langle\nabla l(\mathbf{x}^*),\mathbf{x}^*-\mathbf{x}\rangle&\leq s\langle\nabla D(\mathbf{x}^*,\mathbf{y}),\mathbf{x}-\mathbf{x}^*\rangle\\&=s\langle\nabla h(\mathbf{x}^*)-\nabla h(\mathbf{y}),\mathbf{x}-\mathbf{x}^*\rangle\\&=s[D(\mathbf{x},\mathbf{y})-D(\mathbf{x},\mathbf{x}^*)-D(\mathbf{x}^*,\mathbf{y})].\end{aligned}$$

□

A. Convex Objective Function

In this section, we analyze the convergence when the objective function is convex and there is no additional assumption on it. Based on this weak assumption and Lemma 2, we give the convergence analysis.

We firstly analyze one iteration of our algorithm in Lemma 3, and then give the convergence in terms of expectation eventually in Theorem 3.

Lemma 3. *Assume that $\ell(\cdot)$ and $R(\cdot)$ are convex, with any $k\geq 1$, we have:*

$$\begin{aligned}&\sum_{i=1}^{N}\Big(f_i(\tilde{\mathbf{w}}_i^k)-f_i(\mathbf{w}_i)+(\mathbf{u}_i^{k+1}-\mathbf{u}_i)^T F(\mathbf{u}_i^{k+1})\Big)\\\leq&\sum_{i=1}^{N}\Big(\frac{\eta_i^{k+1}}{2}\|f_i'(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\&+\frac{1}{2\eta_i^{k+1}}\big(\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\|^2-\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\|^2\big)\\&+\frac{\rho}{2}\big(\|\mathbf{w}_i-\mathbf{w}^k\|^2-\|\mathbf{w}_i-\mathbf{w}^{k+1}\|^2\big)\\&-\big\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\big\rangle\\&+\frac{1}{2\rho}\big(\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^k\|^2-\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{k+1}\|^2\big)\Big).\end{aligned}$$

Proof. See Appendix B. □

Following Lemma 3, we give the convergence theorem.

Theorem 3. *Assume that $\ell(\cdot)$ and $R(\cdot)$ are convex, $\|\ell'(\cdot)\|\leq S_1$, and $\|R'(\cdot)\|\leq S_2$. Let $\eta_i^{k+1}=\frac{D_w}{(S_1+\lambda S_2/N)\sqrt{2(k+1)}}$, with any $t\geq 1$ and any $\beta\in\mathbb{R}$, we have:*

$$\begin{aligned}&\mathbb{E}\Big[\sum_{i=1}^{N}\Big(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}^*)+\beta\|\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\|\Big)\Big]\\\leq&\frac{N\sqrt{2}D_w(S_1+\lambda S_2/N)}{\sqrt{t}}+\frac{N(\rho D_w^2+\beta^2/\rho)}{2t}.\end{aligned}$$

Proof. See Appendix C. □

From Theorem 3, the results from our algorithm converges to the optimal result at a rate of $O(1/\sqrt{t})$.

B. Lipschitz Smooth Objective Function

Theorem 3 does not show how ϵ affects the convergence of our proposed algorithm. In order to explore this, we extend Theorem 3 under a stricter assumption that $\ell(\cdot)$ and $R(\cdot)$ are both twice differentiable. Under this stricter assumption, Algorithm 3 also achieves an $O(1/\sqrt{t})$ rate of convergence.

Here, we replace the definition of $\overline{\mathbf{w}}_i^t$: $\overline{\mathbf{w}}_i^t=\frac{1}{t}\sum_{k=0}^{t-1}\tilde{\mathbf{w}}_i^k$ by $\overline{\mathbf{w}}_i^t=\frac{1}{t}\sum_{k=1}^{t}\tilde{\mathbf{w}}_i^k$, so that we could obtain a close-form expression of convergence. As we analyze in Section V-A, we also give the bound of one iteration first, and then show the convergence theorem finally.

Lemma 4. *Assume $\ell(\cdot)$ and $R(\cdot)$ are convex and twice differentiable, $\|\nabla^2\ell(\cdot)\|\leq S_3$, and $\|\nabla^2 R(\cdot)\|\leq S_4$. With*

$k \geq 1$, we have:

$$\begin{aligned}
&\sum_{i=1}^{N}\left(f_i(\tilde{\mathbf{w}}_i^{k+1}) - f_i(\mathbf{w}_i) + (\mathbf{u}_i^{k+1} - \mathbf{u}_i)^T F(\mathbf{u}_i^{k+1})\right)\\
\leq &\sum_{i=1}^{N}\Bigg(\frac{1}{2(1/\eta_i^{k+1} - (S_3 + \lambda S_4/N))}\|(\rho + 1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\
&+ \frac{1}{2\eta_i^{k+1}}\left(\|\mathbf{w}_i - \tilde{\mathbf{w}}_i^k\|^2 - \|\mathbf{w}_i - \tilde{\mathbf{w}}_i^{k+1}\|^2\right)\\
&+ \frac{\rho}{2}\left(\|\mathbf{w}_i - \mathbf{w}^k\|^2 - \|\mathbf{w}_i - \mathbf{w}^{k+1}\|^2\right)\\
&- \left\langle(\rho + 1/\eta_i^{k+1})\xi_i^{k+1}, \mathbf{w}_i - \tilde{\mathbf{w}}_i^k\right\rangle\\
&+ \frac{1}{2\rho}\left(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^k\|^2 - \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^{k+1}\|^2\right)\Bigg).
\end{aligned}$$

Proof. See Appendix D. □

Based on Lemma 4, we give the following theorem.

Theorem 4. *Assume $\ell(\cdot)$ and $R(\cdot)$ are convex and twice differentiable, $\|\nabla^2\ell(\cdot)\| \leq S_3$, and $\|\nabla^2 R(\cdot)\| \leq S_4$. Let $\eta_i^{k+1} = \left(S_3 + \lambda S_4/N + 2S_1\sqrt{4(k+1)\ln(1.25/\delta)}/(m_i\epsilon D_w)\right)^{-1}$. With $t \geq 1$ and $\beta \in \mathbb{R}$, we have:*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\left(f_i(\overline{\mathbf{w}}_i^t) - f_i(\mathbf{w}^*) + \beta\|\overline{\mathbf{w}}_i^t - \overline{\mathbf{w}}^t\|\right)\right]\\
\leq &\sum_{i=1}^{N}\frac{4D_w\sqrt{\ln(1.25/\delta)}S_1}{m_i\epsilon\sqrt{t}} + \frac{ND_w^2(S_3 + \lambda S_4/N)}{2t}\\
&+ \frac{N\rho}{2t}D_w^2 + \frac{1}{t}\frac{N\beta^2}{2\rho}.
\end{aligned}$$

Proof. See Appendix E. □

VI. Performance Evaluation

In this section, we evaluate our proposed algorithm: DP-ADMM by developing a privacy-preserving logistic regression based on a real-world dataset. In order to evaluate our approach under two assumptions: convexity and Lipschitz smoothness, we consider two settings where l_1-norm and l_2-norm regularizers are applied respectively. Our experiments include a training phase and a testing phase. In the training phase, we train classifiers by our approach and the baseline algorithms based on training data, and meanwhile monitor the training process. In the testing phase, we test the accuracy of the trained classifiers based on the testing data, by comparing the predicted labels by the classifiers and their original labels.

Logistic Regression. Logistic regression is a classic machine learning technique that is commonly used in predicting dichotomous outcomes. Without loss of generality, we focus on binary class logistic regression, but our solution can be extended to the case of multiple-class logistic regression. The loss function of the logistic regression is described as:

$$\ell(\mathbf{a}_{i,j}, b_{i,j}, \mathbf{w}_i) = \log(1 + \exp(-b_{i,j}\mathbf{w}_i^T\mathbf{a}_{i,j})),$$

which is convex and differentiable. By data normalization, we enforce $\|\nabla\ell(\cdot)\| \leq 1$ and $\|\nabla^2\ell(\cdot)\| \leq \frac{1}{4}$.

Dataset. We evaluate our approach on a real-world dataset: Adult dataset [18] from UCI Machine Learning Repository.

TABLE II: Statistics of Adult Dataset.

	Pre-training	Training	Testing	Total
Number	162	21000	9000	30162

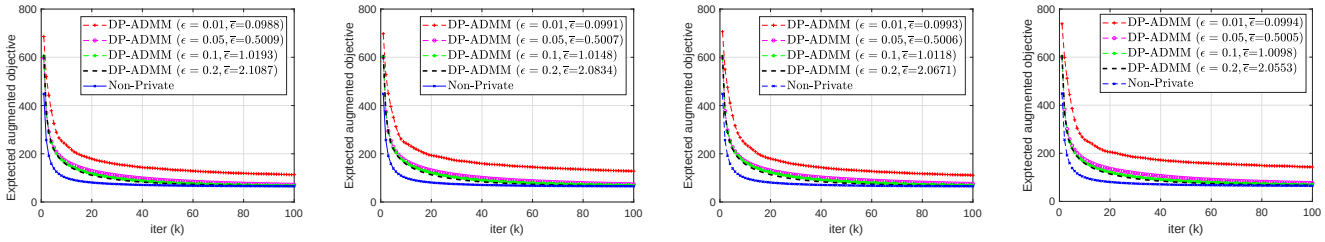
After removing incomplete data entries in Adult dataset, we get 30162 instances, each containing information such as age, sex, education, occupation, martial status, and native country (feature vector dimension $d = 14$). The corresponding label represents whether the income is above \$50000 (labeled by 1) or not (labeled by -1). Before the simulation, we firstly normalize the data. In each simulation, as shown in Table II, we sample 162 instances for pre-training to get an approximate D_w to define η_i^k, 21000 instances for training, and the remaining 9000 instances for testing. In the training process, we assume that there are 100 data providers by dividing 21000 instances into 100 groups ($N = 100$), thus each group containing 210 data points ($m_i = 210$).

Baseline algorithms. We compare our DP-ADMM (Algorithm 3) with two baseline algorithms: the non-private algorithm (Algorithm 1) and the intuitive differentially private approach (Algorithm 2). By comparing with the non-private algorithm, we evaluate the accuracy and the effectiveness of our approach. By comparing with the existing ADMM-based differentially private algorithm, we show the advantages of our work.

Setup. We set up the simulation by MATLAB in an Intel(R) Core(TM) 3.40 GHz computer with 16 GB RAM. In the simulation, we set $\rho = 1$, $\lambda = 0.17$, and the total iteration number $T = 100$. In Algorithm 2, the l_2 sensitivity is equal to $\frac{2S_1N}{\lambda m_i}$, which can be proved according to [8], [13]. Thus, we set σ_i in Algorithm 2 to be $\frac{2S_1N\sqrt{2\ln(1.25/\delta)}}{m_i\lambda\epsilon}$ to guarantee that each iteration of Algorithm 2 is (ϵ, δ)-differentially private. We evaluate our approach by setting $\epsilon = \{0.01, 0.02, 0.05, 0.07, 0.1, 0.2\}$ and $\delta = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, and use moments accountant method to obtain the corresponding $\overline{\epsilon}$. In each simulation, we run it for 10 times to get the averaged result in convergence and accuracy.

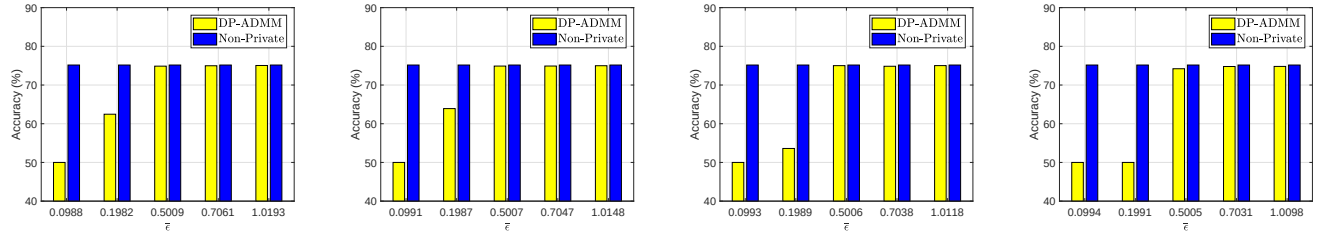
Evaluations. We mainly evaluate our approach on convergence, accuracy, and computation cost. We evaluate the convergence by the expected augmented objective value: $\mathbb{E}\left[\sum_{i=1}^{N}\left(f_i(\overline{\mathbf{w}}_i^k) + \|\overline{\mathbf{w}}_i^k - \overline{\mathbf{w}}^k\|\right)\right]$, and by the expected empirical loss: $\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\mathbf{a}_{i,j}, b_{i,j}, \mathbf{w}^k)\right]$, which measure the optimality and the quality of the training respectively. We compare our approach with baseline algorithms in computation cost by the time needed for training. With the trained classifier $\mathbf{w}$, we evaluate the accuracy of the trained model by following the logistic regression prediction model:

$$\Pr(b = 1|\mathbf{a}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{a})},$$
$$\Pr(b = -1|\mathbf{a}) = \frac{1}{1 + \exp(\mathbf{w}^T\mathbf{a})}.$$



(a) $\delta = 10^{-3}$ (b) $\delta = 10^{-4}$ (c) $\delta = 10^{-5}$ (d) $\delta = 10^{-6}$

Fig. 1: The convergence of DP-ADMM compared with the non-private algorithm (l_1-regularized logistic regression).



(a) $\delta = 10^{-3}$ (b) $\delta = 10^{-4}$ (c) $\delta = 10^{-5}$ (d) $\delta = 10^{-6}$

Fig. 2: The accuracy of DP-ADMM compared with the non-private algorithm (l_1-regularized logistic regression).

A. L_1-norm Regularizer

The l_1-norm regularizer is described as: $R(\mathbf{w}_i) = \|\mathbf{w}_i\|_1$. We obtain the DP-ADMM steps for l_1 regularized logistic regression by:

$$\begin{aligned}
\mathbf{w}_i^{k+1} \leftarrow & \Big(-\frac{1}{m_i}\sum_{j=1}^{m_i} \frac{-b_{i,j}\mathbf{a}_{i,j}}{1+\exp(b_{i,j}\tilde{\mathbf{w}}_i^{k^T}\mathbf{a}_{i,j})} - \frac{\lambda}{N}\cdot \mathrm{sgn}(\tilde{\mathbf{w}}_i^k) \\
& + \boldsymbol{\gamma}_i^k + \rho \mathbf{w}^k + \tilde{\mathbf{w}}_i^k/\eta_i^{k+1}\Big)/\big(\rho + 1/\eta_i^{k+1}\big), \\
\tilde{\mathbf{w}}_i^{k+1} \leftarrow & \tilde{\mathbf{w}}_i^{k+1} + \mathcal{N}(0, \sigma_{i,k+1}^2), \\
\mathbf{w}^{k+1} \leftarrow & \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{w}}_i^{k+1} - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\gamma}_i^k/\rho, \\
\boldsymbol{\gamma}_i^{k+1} \leftarrow & \boldsymbol{\gamma}_i^k - \rho\left(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^{k+1}\right),
\end{aligned}$$

where $\mathrm{sgn}(\cdot)$ is the sign function.

The objective function is convex but non-differentiable, thus we apply Theorem 1 and Theorem 3. Since $\|\ell'(\cdot)\| \le 1$, and $\|R'(\cdot)\| \le \sqrt{d}$ ($d = 14$), we set $S_1 = 1$, and $S_2 = \sqrt{d}$. According to Theorem 1, Theorem 3, and Theorem 2, we set η_i^{k+1} by $\frac{D_w}{(1+\lambda\sqrt{d}/N)\sqrt{2(k+1)}}$, and $\sigma_{i,k+1}$ by $\frac{2\sqrt{2\ln(1.25/\delta)}}{m_i\epsilon(1+1/\eta_i^{k+1})}$.

Since Algorithm 2 cannot be applied when the objective function is non-differentiable, we only compare our approach and the non-private algorithm (Algorithm 1) in this section. Figure 1 compares our approach under different ϵ and δ with the non-private algorithm on convergence, and shows the corresponding $\overline{\epsilon}$. By comparing with the non-private algorithm, we see that DP-ADMM is noise-resistant and convergent under different ϵ and δ. In addition, our approach achieves faster convergence under a larger ϵ or a larger δ. This indicates the privacy-and-utility trade-off of DP-ADMM: our approach has better performance when the privacy guarantee level is lower. Figure 2 shows the comparison on the accuracy of the trained model. The trained model by DP-ADMM has bad performance when $\overline{\epsilon}$ is very small, and its accuracy increases with $\overline{\epsilon}$ and δ. When $\overline{\epsilon} > 0.5$ and $\delta \ge 10^{-6}$, the trained model by our approach is nearly as good as the one in non-private setting. Thus, our approach achieves good accuracy even when the total privacy leakage is low ($\overline{\epsilon} \approx 0.5$, $\delta = 10^{-6}$).

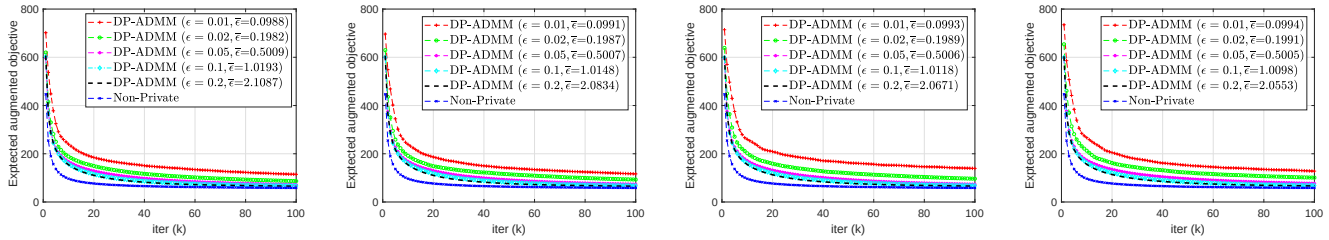
B. L_2-norm Regularizer

The l_2-norm regularizer is described as: $R(\mathbf{w}_i) = \frac{1}{2}\|\mathbf{w}_i\|^2$. Based on DP-ADMM, we update the model iteratively by:

$$\begin{aligned}
\mathbf{w}_i^{k+1} \leftarrow & \Big(-\frac{1}{m_i}\sum_{j=1}^{m_i} \frac{-b_{i,j}\mathbf{a}_{i,j}}{1+\exp(b_{i,j}\tilde{\mathbf{w}}_i^{k^T}\mathbf{a}_{i,j})} - \frac{\lambda}{N}\tilde{\mathbf{w}}_i^k + \boldsymbol{\gamma}_i^k \\
& + \rho \mathbf{w}^k + \tilde{\mathbf{w}}_i^k/\eta_i^{k+1}\Big)/(\rho + 1/\eta_i^{k+1}), \\
\tilde{\mathbf{w}}_i^{k+1} \leftarrow & \mathbf{w}_i^{k+1} + \mathcal{N}(0, \sigma_{i,k+1}^2), \\
\mathbf{w}^{k+1} \leftarrow & \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{w}}_i^{k+1} - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\gamma}_i^k/\rho, \\
\boldsymbol{\gamma}_i^{k+1} \leftarrow & \boldsymbol{\gamma}_i^k - \rho\left(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^{k+1}\right).
\end{aligned}$$

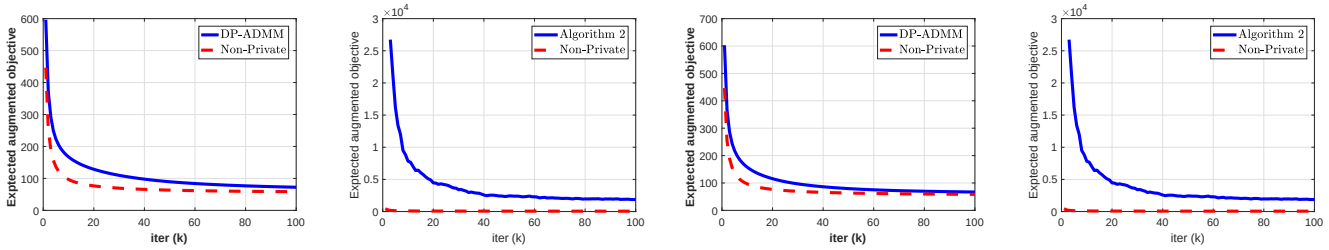
Here the objective function is convex and twice differentiable, thus we apply Theorem 1 and Theorem 4. Since $\|\nabla\ell(\cdot)\| \le 1$, $\|\nabla^2\ell(\cdot)\| \le \frac{1}{4}$, and $\|\nabla^2 R(\cdot)\| \le 1$, we set $S_1 = 1$, $S_3 = \frac{1}{4}$, and $S_4 = 1$. According to Theorem 1, Theorem 4, and Theorem 2, we set η_i^{k+1} by $\big(0.25 + \lambda/N + 2\sqrt{4(k+1)\ln(1.25/\delta)}/(m_i\epsilon D_w)\big)^{-1}$, and $\sigma_{i,k+1}$ by $\frac{2\sqrt{2\ln(1.25/\delta)}}{m_i\epsilon(1+1/\eta_i^{k+1})}$.

Here, we compare our approach with two baseline algorithms (non-private algorithm and Algorithm 2). Figure 3 compares the convergence of the proposed algorithm in different privacy guarantee level settings with the non-private algorithm, and also shows the corresponding $\overline{\epsilon}$.. It proves the noise-resistance and convergence of DP-ADMM by comparing with the non-private algorithm. Besides, it shows the trade-off



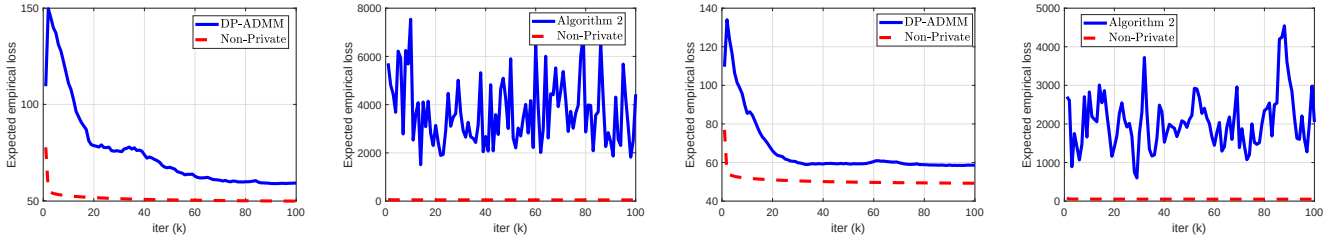
(a) $\delta = 10^{-3}$ (b) $\delta = 10^{-4}$ (c) $\delta = 10^{-5}$ (d) $\delta = 10^{-6}$

Fig. 3: The convergence of DP-ADMM compared with the non-private algorithm (l_2-regularized logistic regression).



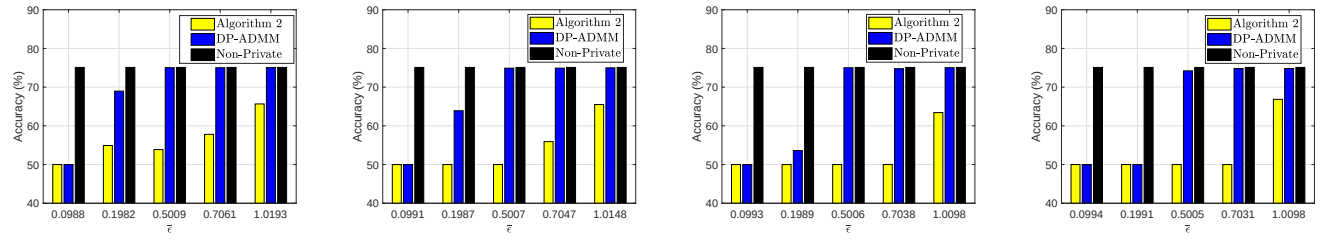
(a) $\epsilon = 0.05$, $\overline{\epsilon} = 0.5009$, $\delta = 10^{-3}$ (b) $\epsilon = 0.05$, $\overline{\epsilon} = 0.5009$, $\delta = 10^{-3}$ (c) $\epsilon = 0.1$, $\overline{\epsilon} = 1.0193$, $\delta = 10^{-3}$ (d) $\epsilon = 0.1$, $\overline{\epsilon} = 1.0193$, $\delta = 10^{-3}$

Fig. 4: The convergence of DP-ADMM compared with the existing differentially private algorithm (l_2-regularized logistic regression).



(a) $\epsilon = 0.05$, $\overline{\epsilon} = 0.5009$, $\delta = 10^{-3}$ (b) $\epsilon = 0.05$, $\overline{\epsilon} = 0.5009$, $\delta = 10^{-3}$ (c) $\epsilon = 0.1$, $\overline{\epsilon} = 1.0193$, $\delta = 10^{-3}$ (d) $\epsilon = 0.1$, $\overline{\epsilon} = 1.0193$, $\delta = 10^{-3}$

Fig. 5: The convergence of DP-ADMM compared with the existing differentially private algorithm (l_2-regularized logistic regression).



(a) $\delta = 10^{-3}$ (b) $\delta = 10^{-4}$ (c) $\delta = 10^{-5}$ (d) $\delta = 10^{-6}$

Fig. 6: Comparison of accuracy among DP-ADMM, non-private algorithm and the existing differentially private algorithm (l_2-regularized logistic regression).

between privacy and utility of DP-ADMM: under a larger ϵ or a larger δ the convergence of our approach will be faster, and it also demonstrates the relation among ϵ, δ, and the convergence shown in Theorem 4. Figure 4 and Figure 5 compare DP-ADMM and the Algorithm 2 on convergence by the expected augmented objective value and the expected empirical loss, which measure the optimality and the quality of the trained model respectively. In both two settings ($\epsilon = 0.05$, $\overline{\epsilon} = 0.5009$, $\delta = 10^{-3}$ and $\epsilon = 0.1$, $\overline{\epsilon} = 1.0193$, $\delta = 10^{-3}$), the training process of the existing ADMM-based differentially private algorithm is disrupted seriously by the noise while the training process of our approach is much more stable and much more resistant to noise. Figure 6 shows the evaluation results on accuracy: in the same privacy guarantee level, our approach achieves much better accuracy than the intuitive algorithm. Besides, it shows that the increase of $\overline{\epsilon}$ and δ

TABLE III: Computation Time (100 iterations).

	Non-Private Algorithm	Algorithm 2	DP-ADMM
$\epsilon = 0.01$	6.184117s	368.871413s	3.063915s
$\epsilon = 0.05$	6.184117s	212.488406s	3.066966s
$\epsilon = 0.1$	6.184117s	70.127314s	3.056092s

will improve the performance of the trained model by our approach. When $\overline{\epsilon} > 0.5$ and $\delta \geq 10^{-6}$, our algorithm is as accurate as the non-private algorithm, which proves that an accurate classifier could be trained by our approach with a low total privacy leakage. Table III gives the comparison in computation time, and show that DP-ADMM achieves faster computation than both two baseline algorithms. This is the result of the first-order approximation used in DP-ADMM, which enables updates by a close-form solution to (3a) rather than by searching by gradients.

VII. Related Work

The problem of privacy-preserving distributed machine learning has attracted a lot of research efforts. The existing literature related to our work could be categorized by: data privacy, privacy-preserving distributed learning, and variants of ADMM.

Data privacy. There have been tremendous research efforts on data privacy guarantee. An intuitive approach is anonymity, which protects user privacy by hiding their identity [19], [20]. However, when multiple data sources are linked together, it is easy to recover the hidden user identities [21]. Another state-of-art approach on private computation is secure multiparty computation [22], [23], by which each party can only learn the computation results and the personal data is kept private, but this approach is very computational expensive and barely used in practical systems. The third approach to guarantee privacy is the perturbation-based approach, which modifies the value of data and provides privacy at the cost of reduced data accuracy. Among the perturbation approaches, differential privacy, a formal privacy definition proposed by Dwork et al. [5] has attracted much attention.

Privacy-preserving distributed learning. Distributed learning reduces the computational burden and improves the scalability, but the privacy risk still exists from the shared information. Recently, much works try to develop a privacy-preserving distributed learning algorithm. Some of them employ cryptography-based methods in the protocol to hide the private information [4], [24]–[27], while others adopt differential privacy as their privacy model. Among the works on differentially private distributed learning, most of them focus subgradient-based algorithms with differential privacy [28]–[31]. Zhang and Zhu [8] develop a differentially private distributed learning algorithm based on ADMM, which has most similarities to our work. Distinguishing from their work, our work focuses on the total privacy loss rather than the dynamic one in order to obtain a better view on the trade-off between the data privacy and utility. Besides, our proposed approach utilizes an approximate augmented Lagrangian function and randomization mechanisms with time-varying variance, which gives a more robust way to combine differential privacy and ADMM. Zhang et al. [10] also utilize this approximate augmented Lagrangian function in the even iterations of their privacy-preserving algorithm to reduce privacy leakage. However, their work requires a strong assumption that the objective function should be twice differentiable and strongly convex, and does not give any theoretical guarantee for convergence. Our approach can be applied to a general class of convex learning problem, and achieves a convergence rate of $O(1/\sqrt{t})$ theoretically.

Variants of ADMM. In order to achieve fast convergence for complex learning problems, there are some variants of ADMM proposed recently. Linearized ADMM [15], [16] gives a better way to solve the problems that do not have close-form solutions by replacing the quadratic function in the augmented Lagrangian function with its first-order approximation. Stochastic ADMM [32], [33] is used in stochastic optimization problems, considering the nature noise in observations to provide convergent and fast updates. Our proposed algorithm: DP-ADMM guarantees differential privacy in ADMM. DP-ADMM inherits the features from the classical ADMM, linearized ADMM, and stochastic ADMM, and it is noise-resistant and convergent, and has low computation cost.

VIII. Conclusion

In this paper, we have addressed privacy issues in distributed machine learning and proposed an improved ADMM-based differentially private distributed learning algorithm, DP-ADMM, for a class of learning problems that can be formulated as convex regularized empirical risk minimization. By modifying the way classical ADMM is updating in each iteration, our novel approach is noise-resistant, convergent and computation-efficient, even under moderate privacy guarantee. We have also applied the moments accountant method to bound the total privacy loss of the proposed algorithm. Our convergence theorems have shown that our approach provides a theoretically guaranteed convergence rate of $O(1/\sqrt{t})$ with t being the number of iterations. The evaluations on real-world datasets have demonstrated the accuracy and effectiveness of our approach in the setting with moderate privacy guarantee.

Appendix A
Proof of Theorem 2

Proof. We use the log moments of the privacy loss and their linear composability to get a tight bound of the total privacy loss. The τ^{th} log moment of the privacy loss of data provider i for k^{th} iteration: $\alpha_i^k(\tau)$ could be defined by the log moment generating function at τ:

$$\alpha_i^k(\tau) = \log\left(\mathbb{E}_{\tilde{\mathbf{w}}_i^k}\left[\left(\frac{P[\tilde{\mathbf{w}}_i^k|\mathcal{D}_i]}{P[\tilde{\mathbf{w}}_i^k|\mathcal{D}_i']}\right)^{\tau}\right]\right).$$

In k^{th} iteration of Algorithm 3, we employ Gaussian mechanism with variance $\sigma_{i,k}^2$ to achieve (ϵ, δ)-differential privacy guarantee. We use μ_0 to denote the probability density function (pdf) of $\mathcal{N}(0, \sigma_{i,k}^2)$, and μ_1 to denote the pdf

of $\mathcal{N}(\frac{2S_1}{m_i(\rho+1/\eta_i^k)}, \sigma_{i,k}^2)$. We obtain the bound of $\alpha_i^k(\tau)$ by $\alpha_i^k(\tau) = \log\big(\max(E_1, E_2)\big)$, where

$$E_1 = \mathbb{E}_{z\sim\mu_0}\left[\left(\frac{\mu_0(z)}{\mu_1(z)}\right)^\tau\right] \quad \text{and} \quad E_2 = \mathbb{E}_{z\sim\mu_1}\left[\left(\frac{\mu_1(z)}{\mu_0(z)}\right)^\tau\right].$$

Since,

$$\mathbb{E}_{z\sim\mu_0}\left[(\frac{\mu_0(z)}{\mu_1(z)})^\tau\right] = \exp\left(\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}\right),$$
$$\mathbb{E}_{z\sim\mu_1}\left[(\frac{\mu_1(z)}{\mu_0(z)})^\tau\right] = \exp\left(\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}\right),$$

we have:

$$\alpha_i^k(\tau) = \frac{\tau(\tau+1)\epsilon}{4\ln(1.25/\delta)}.$$

According to the linear composability of $\alpha_i^k(\tau)$, we have the τ^{th} log moment of the overall privacy loss from i:

$$\alpha_i(\tau) = \sum_{k=1}^{T}\alpha_i^k(\tau) = T\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}.$$

Therefore, when Algorithm 3 guarantees $(\overline{\epsilon}, \delta)$-differential privacy, based on the tail bound property, we have:

$$\delta = \min_{\tau\in\mathbb{Z}^+} \quad \exp(\alpha_i(\tau) - \tau\overline{\epsilon}) = \min_{\tau\in\mathbb{Z}^+} \quad \exp\left(T\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)} - \tau\overline{\epsilon}\right).$$

Since $\delta \in (0,1)$ and there exists a positive integer τ to make $T\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)} - \tau\overline{\epsilon} < 0$, we must have:

$$\frac{T\epsilon^2}{2\ln(1.25/\delta)} < \overline{\epsilon}. \tag{15}$$

The minimum of $T\frac{x(x+1)\epsilon^2}{4\ln(1.25/\delta)} - x\overline{\epsilon}$ is $-\frac{T\epsilon^2}{16\ln(1.25/\delta)} + \frac{\overline{\epsilon}}{2} - \frac{\overline{\epsilon}^2\ln(1.25/\delta)}{T\epsilon^2}$ when $x \in \mathbb{R}$. Thus:

$$\begin{aligned}\ln(\delta) &= \min_{\tau\in\mathbb{Z}^+} \quad \left(T\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)} - \tau\overline{\epsilon}\right) \\ &\geq -\frac{T\epsilon^2}{16\ln(1.25/\delta)} + \frac{\overline{\epsilon}}{2} - \frac{\overline{\epsilon}^2\ln(1.25/\delta)}{T\epsilon^2}\end{aligned} \tag{16}$$

From (15) and (16), we obtain:

$$\begin{aligned}\ln(1/\delta) &\leq -\frac{3\overline{\epsilon}}{8} + \frac{\overline{\epsilon}^2\ln(1.25/\delta)}{T\epsilon^2} \\ &\leq \frac{\overline{\epsilon}^2\ln(1.25/\delta)}{T\epsilon^2}.\end{aligned}$$

Then,

$$\overline{\epsilon} \geq \sqrt{\frac{T\ln(1/\delta)}{\ln(1.25/\delta)}}\epsilon.$$

Thus, there exists a constant c_1, the overall privacy loss $\overline{\epsilon}$ satisfies:

$$\overline{\epsilon} = c_1\sqrt{T}\epsilon.$$

□

APPENDIX B
PROOF OF LEMMA 3

Proof. Due to the convexity of $f_i(\cdot)$, we have:

$$f_i(\tilde{\mathbf{w}}_i^k) - f_i(\mathbf{w}_i) \leq \left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k), \tilde{\mathbf{w}}_i^k - \mathbf{w}_i\right\rangle.$$

Thus,

$$\begin{aligned}
&f_i(\tilde{\mathbf{w}}_i^k) - f_i(\mathbf{w}_i) + \left\langle \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i, -\boldsymbol{\gamma}_i^{k+1}\right\rangle \\
\leq&\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k), \tilde{\mathbf{w}}_i^k - \mathbf{w}_i\right\rangle + \left\langle \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i, -\boldsymbol{\gamma}_i^{k+1}\right\rangle \\
=&\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i\right\rangle \\
&- \left\langle (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \mathbf{w}_i - \tilde{\mathbf{w}}_i^k\right\rangle \\
&+ \left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^k - \tilde{\mathbf{w}}_i^{k+1}\right\rangle \\
&+ \left\langle \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i, -\boldsymbol{\gamma}_i^{k+1}\right\rangle \\
=&\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - \boldsymbol{\gamma}_i^{k+1} - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i\right\rangle \\
&- \left\langle (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \mathbf{w}_i - \tilde{\mathbf{w}}_i^k\right\rangle \\
&+ \left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^k - \tilde{\mathbf{w}}_i^{k+1}\right\rangle.
\end{aligned} \tag{17}$$

According to the Line 10 of Algorithm 3, we have:

$$\begin{aligned}
&\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - \boldsymbol{\gamma}_i^{k+1} - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i\right\rangle \\
=&\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - \boldsymbol{\gamma}_i^k + \rho(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^k)\right. \\
&\qquad\left. - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i\right\rangle \\
&+ \left\langle \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i, \rho(\mathbf{w}^k - \mathbf{w}^{k+1})\right\rangle.
\end{aligned} \tag{18}$$

By combining (17) and (18), we obtain:

$$\begin{aligned}
&f_i(\tilde{\mathbf{w}}_i^k) - f_i(\mathbf{w}_i) + \left\langle \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i, -\boldsymbol{\gamma}_i^{k+1}\right\rangle \\
\leq&-\left\langle (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \mathbf{w}_i - \tilde{\mathbf{w}}_i^k\right\rangle \\
&+ \left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - \boldsymbol{\gamma}_i^k + \rho(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^k)\right. \\
&\qquad\left. - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i\right\rangle \\
&+ \left\langle \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i, \rho(\mathbf{w}^k - \mathbf{w}^{k+1})\right\rangle \\
&+ \left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k) - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \tilde{\mathbf{w}}_i^k - \tilde{\mathbf{w}}_i^{k+1}\right\rangle.
\end{aligned} \tag{19}$$

We handle the last three terms separately. Firstly, we have:

$$\begin{aligned}
&\left\langle \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i, \rho(\mathbf{w}^k - \mathbf{w}^{k+1})\right\rangle \\
=&\frac{\rho}{2}\big(\|\mathbf{w}_i - \mathbf{w}^k\|^2 - \|\mathbf{w}_i - \mathbf{w}^{k+1}\|^2\big) \\
&\qquad + \frac{\rho}{2}\big(\|\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^{k+1}\|^2 - \|\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^k\|^2\big) \\
\leq&\frac{\rho}{2}\big(\|\mathbf{w}_i - \mathbf{w}^k\|^2 - \|\mathbf{w}_i - \mathbf{w}^{k+1}\|^2\big) + \frac{\rho}{2}\|\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^{k+1}\|^2 \\
=&\frac{\rho}{2}\big(\|\mathbf{w}_i - \mathbf{w}^k\|^2 - \|\mathbf{w}_i - \mathbf{w}^{k+1}\|^2\big) + \frac{1}{2\rho}\|\boldsymbol{\gamma}_i^{k+1} - \boldsymbol{\gamma}_i^k\|^2.
\end{aligned} \tag{20}$$

According to the Line 4 and 6 of Algorithm 3, $\tilde{\mathbf{w}}_i^{k+1}$ is equal to the optimum of $\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k), \mathbf{w}_i - \tilde{\mathbf{w}}_i^k\right\rangle - \left\langle \boldsymbol{\gamma}_i^k, \mathbf{w}_i - \mathbf{w}^k\right\rangle + \frac{\rho}{2}\|\mathbf{w}_i - \mathbf{w}^k\|^2 + \frac{\|\mathbf{w}_i - \tilde{\mathbf{w}}_i^k\|^2}{2\eta_i^{k+1}} - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}\mathbf{w}_i$. By applying Lemma 1 where $D(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2$, $s = \frac{1}{2\eta_i^{k+1}}$, and $l(x) = \left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k), x - \tilde{\mathbf{w}}_i^k\right\rangle - \left\langle \boldsymbol{\gamma}_i^k, x - \mathbf{w}^k\right\rangle + \frac{\rho}{2}\|x - \mathbf{w}^k\|^2 - (\rho+1/\eta_i^{k+1})\xi_i^{k+1}$, we have:

$$\begin{aligned}
&\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k, \xi_i^{k+1}) - \boldsymbol{\gamma}_i^k + \rho(\tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}^k), \tilde{\mathbf{w}}_i^{k+1} - \mathbf{w}_i\right\rangle \\
\leq&\frac{1}{2\eta_i^{k+1}}\big(\|\mathbf{w}_i - \tilde{\mathbf{w}}_i^k\|^2 - \|\mathbf{w}_i - \tilde{\mathbf{w}}_i^{k+1}\|^2\big) \\
&\qquad - \frac{1}{2\eta_i^{k+1}}\|\tilde{\mathbf{w}}_i^{k+1} - \tilde{\mathbf{w}}_i^k\|^2.
\end{aligned} \tag{21}$$

Lastly, based on Young's inequality, we have:

$$\begin{aligned}&\left\langle f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^k-\tilde{\mathbf{w}}_i^{k+1}\right\rangle\\ \leq&\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\ &+\frac{1}{2\eta_i^{k+1}}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\|^2.\end{aligned}\tag{22}$$

Combining (19),(20),(21), and (22), we have:

$$\begin{aligned}&f_i(\tilde{\mathbf{w}}_i^k)-f_i(\mathbf{w}_i)+\left\langle\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i,-\boldsymbol{\gamma}_i^{k+1}\right\rangle\\ \leq&\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\ &-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\right\rangle\\ &+\frac{1}{2\eta_i^{k+1}}(\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\|^2-\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\|^2)\\ &+\frac{\rho}{2}(\|\mathbf{w}_i-\mathbf{w}^k\|^2-\|\mathbf{w}_i-\mathbf{w}^{k+1}\|^2)+\frac{1}{2\rho}\|\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i^k\|^2.\end{aligned}\tag{23}$$

Next, according to our algorithm where $\boldsymbol{\gamma}_i^{k+1}=\boldsymbol{\gamma}_i^k-\rho(\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}^{k+1})$, we have:

$$\begin{aligned}&\sum_{i=1}^N\left\langle\mathbf{w}^{k+1}-\mathbf{w},\boldsymbol{\gamma}_i^{k+1}\right\rangle\\ =&\left\langle\mathbf{w}^{k+1}-\mathbf{w},\sum_{i=1}^N(\boldsymbol{\gamma}_i^k-\rho\tilde{\mathbf{w}}_i^{k+1})+N\rho\mathbf{w}^{k+1}\right\rangle=0.\end{aligned}\tag{24}$$

And also, we could obtain:

$$\begin{aligned}&\left\langle\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i,\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}^{k+1}\right\rangle\\ =&\frac{1}{\rho}\left\langle\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i,\boldsymbol{\gamma}_i^k-\boldsymbol{\gamma}_i^{k+1}\right\rangle\\ =&\frac{1}{2\rho}(\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^k\|^2-\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{k+1}\|^2-\|\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i^k\|^2).\end{aligned}\tag{25}$$

Thus, combining (23), (24) and (25), we obtain the result in the Lemma 2:

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N\left(f_i(\tilde{\mathbf{w}}_i^k)-f_i(\mathbf{w}_i)+(\mathbf{u}_i^{k+1}-\mathbf{u}_i)^TF(\mathbf{u}_i^{k+1})\right)\\ =&\frac{1}{N}\sum_{i=1}^N\left(f_i(\tilde{\mathbf{w}}_i^k)-f_i(\mathbf{w}_i)+\left\langle-\boldsymbol{\gamma}_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i\right\rangle\right.\\ &\left.+\left\langle\boldsymbol{\gamma}_i^{k+1},\mathbf{w}^{k+1}-\mathbf{w}\right\rangle+\left\langle\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i,\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}^{k+1}\right\rangle\right)\\ \leq&\frac{1}{N}\sum_{i=1}^N\left(\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\right.\\ &+\frac{1}{2\eta_i^{k+1}}(\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\|^2-\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\|^2)\\ &+\frac{\rho}{2}(\|\mathbf{w}_i-\mathbf{w}^k\|^2-\|\mathbf{w}_i-\mathbf{w}^{k+1}\|^2)\\ &-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\right\rangle\\ &\left.+\frac{1}{2\rho}(\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^k\|^2-\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{k+1}\|^2)\right).\end{aligned}$$

□

APPENDIX C
PROOF OF THEOREM 3

Proof. According to the convexity of $f_i(\cdot)$ and the monotonicity of the operator $F(\cdot)$, we have:

$$\begin{aligned}&\sum_{i=1}^N\left(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}_i)+(\overline{\mathbf{u}}_i^t-\mathbf{u}_i)^TF(\overline{\mathbf{u}}_i^t)\right)\\ \leq&\frac{1}{t}\sum_{k=0}^{t-1}\sum_{i=1}^N\left(f_i(\tilde{\mathbf{w}}_i^k)-f_i(\mathbf{w}_i)+(\mathbf{u}_i^{k+1}-\mathbf{u}_i)^TF(\mathbf{u}_i^{k+1})\right)\\ =&\frac{1}{t}\sum_{k=0}^{t-1}\sum_{i=1}^N\left(f_i(\tilde{\mathbf{w}}_i^{k+1})-f_i(\mathbf{w}_i)+\left\langle-\boldsymbol{\gamma}_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i\right\rangle\right.\\ &\left.+\left\langle\boldsymbol{\gamma}_i^{k+1},\mathbf{w}^{k+1}-\mathbf{w}\right\rangle+\left\langle\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i,\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}^{k+1}\right\rangle\right).\end{aligned}$$

We apply Lemma 2 and let $(\mathbf{w}_i,\mathbf{w})$ be the optimal solution $(\mathbf{w}_i^*,\mathbf{w}^*)$ in the above inequality. We get: $\forall\boldsymbol{\gamma}_i$,

$$\begin{aligned}&\sum_{i=1}^N\left(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}_i^*)+\left\langle-\overline{\boldsymbol{\gamma}}_i^t,\overline{\mathbf{w}}_i^t-\mathbf{w}_i^*\right\rangle\right.\\ &\left.+\left\langle\overline{\boldsymbol{\gamma}}_i^t,\overline{\mathbf{w}}^t-\mathbf{w}^*\right\rangle+\left\langle\overline{\boldsymbol{\gamma}}_i^t-\boldsymbol{\gamma}_i,\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\right\rangle\right)\\ \leq&\sum_{i=1}^N\frac{1}{t}\sum_{k=0}^{t-1}\left(\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\right.\\ &\left.-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i^*-\tilde{\mathbf{w}}_i^k\right\rangle\right)\\ &+\frac{1}{t}\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{2\eta^t}\|\mathbf{w}_i^*-\tilde{\mathbf{w}}_i^0\|^2+\frac{\rho}{2}\|\mathbf{w}_i^*-\mathbf{w}^0\|^2\right.\\ &\left.+\frac{1}{2\rho}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^0\|^2\right)\\ =&\sum_{i=1}^N\frac{1}{t}\sum_{k=0}^{t-1}\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\ &-\sum_{i=1}^N\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i^*-\tilde{\mathbf{w}}_i^k\right\rangle\\ &+\frac{N}{t}(\frac{D_w^2}{2\eta_t}+\frac{\rho}{2}D_w^2)+\frac{1}{t}\sum_{i=1}^N\frac{1}{2\rho}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^0\|^2.\end{aligned}$$

The above inequality holds for all $\boldsymbol{\gamma}_i$, thus it also holds for $\boldsymbol{\gamma}_i\in\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}$. By letting $\boldsymbol{\gamma}_i$ be the optimal solution, we have the maximum of the left side:

$$\begin{aligned}&\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}}\sum_{i=1}^N\left(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}_i^*)+\left\langle-\overline{\boldsymbol{\gamma}}_i^t,\overline{\mathbf{w}}_i^t-\mathbf{w}_i^*\right\rangle\right.\\ &\left.+\left\langle\overline{\boldsymbol{\gamma}}_i^t,\overline{\mathbf{w}}^t-\mathbf{w}^*\right\rangle+\left\langle\overline{\boldsymbol{\gamma}}_i^t-\boldsymbol{\gamma}_i,\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\right\rangle\right)\\ =&\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}}\sum_{i=1}^N\left(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}_i)-\boldsymbol{\gamma}_i(\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t)\right)\\ =&\sum_{i=1}^N\left(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}_i)+\beta(\|\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\|)\right).\end{aligned}$$

And we also get the maximum of the right side:

$$\begin{aligned}
&\sum_{i=1}^{N}\frac{1}{t}\sum_{k=0}^{t-1}\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\
&-\sum_{i=1}^{N}\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i^*-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\frac{N}{t}(\frac{D_w^2}{2\eta_t}+\frac{\rho}{2}D_w^2)+\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\le\beta\}}\frac{1}{t}\sum_{i=1}^{N}\frac{1}{2\rho}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^0\|^2\\
=&\sum_{i=1}^{N}\frac{1}{t}\sum_{k=0}^{t-1}\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\
&-\sum_{i=1}^{N}\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i^*-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\frac{N}{t}(\frac{D_w^2}{2\eta_t}+\frac{\rho}{2}D_w^2)+\frac{N}{t}\frac{\beta^2}{2\rho}.
\end{aligned}$$

Thus, we obtain the inequality:

$$\begin{aligned}
&\sum_{i=1}^{N}\left(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}_i)+\beta\|\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\|\right)\\
\le&\sum_{i=1}^{N}\frac{1}{t}\sum_{k=0}^{t-1}\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\
&-\sum_{i=1}^{N}\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i^*-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\frac{N}{t}(\frac{D_w^2}{2\eta_t}+\frac{\rho}{2}D_w^2)+\frac{N}{t}\frac{\beta^2}{2\rho}.
\end{aligned}\tag{26}$$

Since we assume $\|\ell^{'}(\cdot)\|\le S_1$ and $\|R^{'}(\cdot)\|\le S_2$, $\mathbb{E}\big[\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|\big]=\|f_i^{'}(\tilde{\mathbf{w}}_i^k)\|=\|\ell^{'}(\cdot)+\frac{\lambda}{N}R^{'}(\cdot)\|\le S_1+\lambda S_2/N$. With $\mathbb{E}\big[\big\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\big\rangle\big]=0$ and $\eta_i^{k+1}=\frac{D_w}{(S_1+\lambda S_2/N)\sqrt{2(k+1)}}$, by taking expectation of the inequality (26), we obtain the result in the theorem:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i=1}^{N}\big(f_i(\overline{\mathbf{w}}_i^t)-f_i(\mathbf{w}_i^*)+\beta\|\overline{\mathbf{w}}_i^t-\overline{\mathbf{w}}^t\|\big)\Big]\\
\le&\sum_{i=1}^{N}\frac{1}{t}\mathbb{E}\big[\sum_{k=0}^{t-1}\frac{\eta_i^{k+1}}{2}\|f_i^{'}(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\big]\\
&+\sum_{i=1}^{N}\mathbb{E}\big[\big\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\big\rangle\big]\\
&+\frac{N}{t}(\frac{D_w^2}{2\eta_t}+\frac{\rho}{2}D_w^2)+\frac{N}{t}\frac{\beta^2}{2\rho}\\
=&\frac{N}{t}(\sum_{k=0}^{t-1}\frac{D_w(S_1+\lambda S_2/N)}{2\sqrt{2k}}+\frac{D_w^2(S_1+S_2)\sqrt{2t}/D_w}{2})\\
&+\frac{N\rho}{2t}D_w^2+\frac{N\beta^2}{2\rho t}\\
=&\frac{ND_w(S_1+\lambda S_2/N)}{2\sqrt{2}t}(\sum_{k=0}^{t-1}\frac{1}{\sqrt{k}}+2\sqrt{t})+\frac{N\rho}{2t}D_w^2+\frac{N\beta^2}{2\rho t}\\
\le&\frac{N\sqrt{2}D_w(S_1+\lambda S_2/N)}{\sqrt{t}}+\frac{N(\rho D_w^2+\beta^2/\rho)}{2t}.
\end{aligned}$$

□

APPENDIX D
PROOF OF LEMMA 4

Proof. As we assume that $\ell(\cdot)$ and $R(\cdot)$ are differentiable and convex, $\nabla\ell(\cdot)$ and $\nabla R(\cdot)$ are also differentiable, $\|\nabla^2\ell(\cdot)\|\le S_3$, and $\|\nabla^2R(\cdot)\|\le S_4$, thus we have $\|\nabla^2 f_i(\cdot)\|=\|\nabla^2\ell(\cdot)+\frac{\lambda}{N}\nabla^2R(\cdot)\|\le S_3+\lambda S_4/N$ is bounded. We have:

$$\|\nabla f_i(x)-\nabla f_i(y)\|\le(S_3+\lambda S_4/N)\|x-y\|.$$

Thus, $f_i(\cdot)$ is (S_3+S_4)-Lipschitz smooth. According to the quadratic upper bound property of Lipschitz smooth, we have:

$$\begin{aligned}
f_i(\tilde{\mathbf{w}}_i^{k+1})\le&f_i(\tilde{\mathbf{w}}_i^k)+\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k),\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\frac{S_3+\lambda S_4/N}{2}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\|^2\\
=&f_i(\tilde{\mathbf{w}}_i^k)+\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\frac{S_3+\lambda S_4/N}{2}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\|^2.
\end{aligned}\tag{27}$$

Due to the convexity of $f_i(\cdot)$, we have:

$$f_i(\tilde{\mathbf{w}}_i^k)-f_i(\mathbf{w}_i)\le\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k),\tilde{\mathbf{w}}_i^k-\mathbf{w}_i\right\rangle.\tag{28}$$

According to (27) and (28), we have:

$$\begin{aligned}
&f_i(\tilde{\mathbf{w}}_i^{k+1})-f_i(\mathbf{w}_i)+\left\langle\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i,-\boldsymbol{\gamma}_i^{k+1}\right\rangle\\
\le&f_i(\tilde{\mathbf{w}}_i^k)-f_i(\mathbf{w}_i)\\
&+\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\frac{S_3+\lambda S_4/N}{2}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\|^2+\left\langle\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i,-\boldsymbol{\gamma}_i^{k+1}\right\rangle\\
\le&\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k),\tilde{\mathbf{w}}_i^k-\mathbf{w}_i\right\rangle\\
&+\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\right\rangle\\
&+\frac{S_3+\lambda S_4/N}{2}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\|^2\\
&+\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i\right\rangle\\
&+\left\langle\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i,-\boldsymbol{\gamma}_i^{k+1}\right\rangle\\
=&-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\right\rangle\\
&+\frac{S_3+\lambda S_4/N}{2}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\|^2\\
&+\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i\right\rangle\\
&+\left\langle\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i,-\boldsymbol{\gamma}_i^{k+1}\right\rangle\\
=&-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\right\rangle\\
&+\frac{S_3+\lambda S_4/N}{2}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^k\|^2\\
&+\left\langle\nabla f_i(\tilde{\mathbf{w}}_i^k)-(\rho+1/\eta_i^{k+1})\xi_i^{k+1}-\boldsymbol{\gamma}_i^k\right.\\
&\qquad\left.+\rho(\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}^k),\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i\right\rangle\\
&+\left\langle\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i,\rho(\mathbf{w}^k-\mathbf{w}^{k+1})\right\rangle.
\end{aligned}\tag{29}$$

Based on Young's inequality,

$$\begin{aligned}
&-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\right\rangle\\
=&-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k}\right\rangle\\
&+\left\langle-(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\tilde{\mathbf{w}}_i^{k}-\tilde{\mathbf{w}}_i^{k+1}\right\rangle\\
\le&-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k}\right\rangle\\
&+\frac{1}{2(1/\eta_i^{k+1}-(S_3+\lambda S_4/N))}\|(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\
&+\frac{1/\eta_i^{k+1}-(S_3+\lambda S_4/N)}{2}\|\tilde{\mathbf{w}}_i^{k+1}-\tilde{\mathbf{w}}_i^{k}\|^2.
\end{aligned}\tag{30}$$

Combining (20), (21), (29) and (30), we have:

$$\begin{aligned}
&f_i(\tilde{\mathbf{w}}_i^{k+1})-f_i(\mathbf{w}_i)+\left\langle\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i,-\boldsymbol{\gamma}_i^{k+1}\right\rangle\\
\le&-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k}\right\rangle\\
&+\frac{1}{2(1/\eta_i^{k+1}-(S_3+\lambda S_4/N))}\|(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\
&+\frac{1}{2\eta_i^{k+1}}(\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k}\|^2-\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\|^2)\\
&+\frac{\rho}{2}(\|\mathbf{w}_i-\mathbf{w}^{k}\|^2-\|\mathbf{w}_i-\mathbf{w}^{k+1}\|^2)+\frac{1}{2\rho}\|\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i^{k}\|^2.
\end{aligned}\tag{31}$$

Combining (31), (24) and (25), we get the result as desired:

$$\begin{aligned}
&\sum_{i=1}^{N}\left(f_i(\tilde{\mathbf{w}}_i^{k+1})-f_i(\mathbf{w}_i)+(\mathbf{u}_i^{k+1}-\mathbf{u}_i)^TF(\mathbf{u}_i^{k+1})\right)\\
=&\frac{1}{N}\sum_{i=1}^{N}\Big(f_i(\tilde{\mathbf{w}}_i^{k+1})-f_i(\mathbf{w}_i)+\left\langle-\boldsymbol{\gamma}_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i\right\rangle\\
&+\left\langle\boldsymbol{\gamma}_i^{k+1},\mathbf{w}^{k+1}-\mathbf{w}\right\rangle+\left\langle\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i,\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}^{k+1}\right\rangle\Big)\\
\le&\sum_{i=1}^{N}\Big(\frac{1}{2(1/\eta_i^{k+1}-(S_3+\lambda S_4/N))}\|(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\\
&+\frac{1}{2\eta_i^{k+1}}(\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k}\|^2-\|\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k+1}\|^2)\\
&+\frac{\rho}{2}(\|\mathbf{w}_i-\mathbf{w}^{k}\|^2-\|\mathbf{w}_i-\mathbf{w}^{k+1}\|^2)\\
&-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k}\right\rangle\\
&+\frac{1}{2\rho}(\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{k}\|^2-\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{k+1}\|^2)\Big).
\end{aligned}$$

□

APPENDIX E
PROOF OF THEOREM 4

Proof. According to the convexity of $f_i(\cdot)$ and the monotonicity of $F(\cdot)$:

$$\begin{aligned}
&\sum_{i=1}^{N}\left(f_i(\overline{\mathbf{w}}_i^{t})-f_i(\mathbf{w}_i)+(\overline{\mathbf{u}}_i^{t}-\mathbf{u}_i)^TF(\overline{\mathbf{u}}_i^{t})\right)\\
\le&\frac{1}{t}\sum_{k=0}^{t-1}\sum_{i=1}^{N}\left(f_i(\tilde{\mathbf{w}}_i^{k+1})-f_i(\mathbf{w}_i)+(\mathbf{u}_i^{k}-\mathbf{u}_i)^TF(\mathbf{u}_i^{k})\right)\\
=&\frac{1}{t}\sum_{k=0}^{t-1}\sum_{i=1}^{N}\Big(f_i(\tilde{\mathbf{w}}_i^{k+1})-f_i(\mathbf{w}_i)+\left\langle-\boldsymbol{\gamma}_i^{k+1},\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}_i\right\rangle\\
&+\left\langle\boldsymbol{\gamma}_i^{k+1},\mathbf{w}^{k+1}-\mathbf{w}\right\rangle+\left\langle\boldsymbol{\gamma}_i^{k+1}-\boldsymbol{\gamma}_i,\tilde{\mathbf{w}}_i^{k+1}-\mathbf{w}^{k+1}\right\rangle\Big).
\end{aligned}$$

By applying Lemma 3 and letting $(\mathbf{w}_i,\mathbf{w})$ be the optimal solution $(\mathbf{w}_i^*,\mathbf{w}^*)$, we have:

$$\begin{aligned}
&\sum_{i=1}^{N}\Big(f_i(\overline{\mathbf{w}}_i^{t})-f_i(\mathbf{w}_i^{*})+\left\langle-\overline{\boldsymbol{\gamma}}_i^{t},\overline{\mathbf{w}}_i^{t}-\mathbf{w}_i^{*}\right\rangle\\
&\qquad+\left\langle\overline{\boldsymbol{\gamma}}_i^{t},\overline{\mathbf{w}}^{t}-\mathbf{w}^{*}\right\rangle+\left\langle\overline{\boldsymbol{\gamma}}_i^{t}-\boldsymbol{\gamma}_i,\overline{\mathbf{w}}_i^{t}-\overline{\mathbf{w}}^{t}\right\rangle\Big)\\
\le&\sum_{i=1}^{N}\frac{1}{t}\sum_{k=0}^{t-1}\Big(\frac{\|(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2}{2(1/\eta_i^{k+1}-(S_3+\lambda S_4/N))}\\
&\qquad-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i^{*}-\tilde{\mathbf{w}}_i^{k}\right\rangle\Big)\\
&+\frac{1}{t}\sum_{i=1}^{N}(\frac{1}{2\eta^t}\|\mathbf{w}_i^{*}-\tilde{\mathbf{w}}_i^{0}\|^2+\frac{\rho}{2}\|\mathbf{w}_i^{*}-\mathbf{w}^{0}\|^2+\frac{1}{2\rho}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{0}\|^2)\\
=&\sum_{i=1}^{N}\frac{1}{t}\sum_{k=0}^{t-1}\Big(\frac{\|(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2}{2(1/\eta_i^{k+1}-(S_3+\lambda S_4/N))}\\
&\qquad-\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i^{*}-\tilde{\mathbf{w}}_i^{k}\right\rangle\Big)\\
&+\frac{N}{t}(\frac{D_w^2}{2\eta_t}+\frac{\rho}{2}D_w^2)+\frac{1}{t}\sum_{i=1}^{N}\frac{1}{2\rho}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{0}\|^2.
\end{aligned}$$

The above inequality holds for all $\boldsymbol{\gamma}_i$, thus it also holds for $\boldsymbol{\gamma}_i\in\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\le\beta\}$. By letting $\boldsymbol{\gamma}_i$ be the optimum, we have

$$\begin{aligned}
&\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\le\beta\}}\sum_{i=1}^{N}\Big(f_i(\overline{\mathbf{w}}_i^{t})-f_i(\mathbf{w}_i^{*})+\left\langle-\overline{\boldsymbol{\gamma}}_i^{t},\overline{\mathbf{w}}_i^{t}-\mathbf{w}_i^{*}\right\rangle\\
&\qquad+\left\langle\overline{\boldsymbol{\gamma}}_i^{t},\overline{\mathbf{w}}^{t}-\mathbf{w}^{*}\right\rangle+\left\langle\overline{\boldsymbol{\gamma}}_i^{t}-\boldsymbol{\gamma}_i,\overline{\mathbf{w}}_i^{t}-\overline{\mathbf{w}}^{t}\right\rangle\Big)\\
=&\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\le\beta\}}\sum_{i=1}^{N}\left(f_i(\overline{\mathbf{w}}_i^{t})-f_i(\mathbf{w}_i)-\boldsymbol{\gamma}_i(\overline{\mathbf{w}}_i^{t}-\overline{\mathbf{w}}^{t})\right)\\
=&\sum_{i=1}^{N}\left(f_i(\overline{\mathbf{w}}_i^{t})-f_i(\mathbf{w}_i)+\beta\|\overline{\mathbf{w}}_i^{t}-\overline{\mathbf{w}}^{t}\|\right).
\end{aligned}\tag{32}$$

We have $\mathbb{E}\big[\left\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1},\mathbf{w}_i-\tilde{\mathbf{w}}_i^{k}\right\rangle\big]=0$ and $\mathbb{E}\big[\|(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2\big]=\sigma_{i,k+1}^2(\rho+1/\eta_i^{k+1})^2=\frac{2\ln(1.25/\delta)4S_i^2}{m_i^2\epsilon^2}$. By

taking expectation of (32) and letting $\eta_i^{k+1} = \big(S_3 + \lambda S_4/N + 2S_1\sqrt{4(k+1)\ln(1.25/\delta)}/(\epsilon m_i D_w)\big)^{-1}$, we obtain the result:

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{i=1}^{N}\big(f_i(\overline{\mathbf{w}}_i^t) - f_i(\mathbf{w}_i^*) + \beta\|\overline{\mathbf{w}}_i^t - \overline{\mathbf{w}}^t\|\big)\Big] \\
\leq&\mathbb{E}\big[\sum_{i=1}^{N}\frac{1}{t}\sum_{k=0}^{t-1}\frac{\|(\rho+1/\eta_i^{k+1})\xi_i^{k+1}\|^2}{2(1/\eta_i^{k+1} - (S_3+\lambda S_4/N))}\big] \\
&-\sum_{i=1}^{N}\mathbb{E}\big[\big\langle(\rho+1/\eta_i^{k+1})\xi_i^{k+1}, \mathbf{w}_i - \tilde{\mathbf{w}}_i^k\big\rangle\big] \\
&+\frac{N}{t}(\frac{D_w^2}{2\eta_t} + \frac{N\rho}{2}D_w^2) + \max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}}\frac{1}{t}\sum_{i=1}^{N}\frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^0\|^2 \\
=&\frac{1}{t}\sum_{i=1}^{N}\sum_{k=0}^{t-1}\frac{2\ln(1.25/\delta)4S_1^2/\epsilon^2}{\sqrt{4(k+1)\ln(1.25/\delta)}2S_1/(m_i\epsilon D_w)} \\
&+\frac{N}{t}\frac{D_w^2(S_3+\lambda S_4/N + \sqrt{4(k+1)\ln(1.25/\delta)}2S_1/(\epsilon D_w)}{2} \\
&+\frac{N\rho}{2t}D_w^2 + \frac{N}{t}\frac{\beta^2}{2\rho} \\
=&\sum_{i=1}^{N}\frac{D_w\sqrt{\ln(1.25/\delta)}S_1}{m_i\epsilon t}(\sum_{k=0}^{t-1}\frac{1}{\sqrt{k+1}} + 2\sqrt{t}) \\
&+\frac{ND_w^2(S_3+\lambda S_4/N)}{2t} + \frac{\rho N}{2t}D_w^2 + \frac{N}{t}\frac{\beta^2}{2\rho} \\
\leq&\sum_{i=1}^{N}\frac{4D_w\sqrt{\ln(1.25/\delta)}S_1}{m_i\epsilon\sqrt{t}} + \frac{ND_w^2(S_3+\lambda S_4/N)}{2t} \\
&+\frac{N\rho}{2t}D_w^2 + \frac{1}{t}\frac{N\beta^2}{2\rho}.
\end{aligned}
$$

□

REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[2] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 3–18.

[3] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1322–1333.

[4] C. Zhang, M. Ahmad, and Y. Wang, "Admm based privacy-preserving decentralized optimization," *IEEE Transactions on Information Forensics and Security*, 2018.

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

[6] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[7] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[8] T. Zhang and Q. Zhu, "Dynamic differential privacy for admm-based distributed classification learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 172–187, 2017.

[9] X. Zhang, M. M. Khalili, and M. Liu, "Improving the privacy and accuracy of admm-based distributed algorithms," *arXiv preprint arXiv:1806.02246*, 2018.

[10] ——, "Recycled admm: Improve privacy and accuracy with less computation in distributed algorithms," *arXiv preprint arXiv:1810.03197*, 2018.

[11] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.

[12] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[13] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[14] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[15] J. Yang and X. Yuan, "Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization," *Mathematics of computation*, vol. 82, no. 281, pp. 301–329, 2013.

[16] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Advances in neural information processing systems*, 2011, pp. 612–620.

[17] I. Mironov, "Renyi differential privacy," in *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*. IEEE, 2017, pp. 263–275.

[18] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[19] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Distributed computing systems, 2005. ICDCS 2005. Proceedings. 25th IEEE international conference on*. IEEE, 2005, pp. 620–629.

[20] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*. IEEE, 2010, pp. 238–243.

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

[22] J. Vaidya and C. Clifton, "Privacy-preserving decision trees over vertically partitioned data," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2005, pp. 139–152.

[23] J. Vaidya, M. Kantarcıoğlu, and C. Clifton, "Privacy-preserving naive bayes classification," *The VLDB JournalThe International Journal on Very Large Data Bases*, vol. 17, no. 4, pp. 879–898, 2008.

[24] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy preserving machine learning." *IACR Cryptology ePrint Archive*, vol. 2017, p. 281, 2017.

[25] Z. Erkin and G. Tsudik, "Private computation of spatial and temporal power consumption with smart meters." in *ACNS*, vol. 12. Springer, 2012, pp. 561–577.

[26] E. Shi, H. Chan, E. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Annual Network & Distributed System Security Symposium (NDSS)*. Internet Society., 2011.

[27] Q. Wang, S. Hu, M. Du, J. Wang, and K. Ren, "Learning privately: Privacy-preserving canonical correlation analysis for cross-media retrieval," in *INFOCOM, 2017 Proceedings IEEE*. IEEE, 2017, pp. 100–108.

[28] A. Bellet, R. Guerraoui, M. Taziki, and M. Tommasi, "Personalized and private peer-to-peer machine learning," *arXiv preprint arXiv:1705.08435*, 2017.

[29] S. Han, U. Topcu, and G. J. Pappas, "Differentially private distributed constrained optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 50–64, 2017.

[30] M. Hale and M. Egerstedt, "Differentially private cloud-based multi-agent optimization with constraints," *arXiv preprint arXiv:1708.08422*, 2017.

[31] Z. Huang, S. Mitra, and N. Vaidya, "Differentially private distributed optimization," in *Proceedings of the 2015 International Conference on Distributed Computing and Networking*. ACM, 2015, p. 4.

[32] H. Ouyang, N. He, L. Tran, and A. Gray, "Stochastic alternating direction method of multipliers," in *International Conference on Machine Learning*, 2013, pp. 80–88.

[33] S. Azadi and S. Sra, "Towards an optimal stochastic alternating direction method of multipliers," in *International Conference on Machine Learning*, 2014, pp. 620–628.