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

Zonghao Huang, Rui Hu, Yuanxiong Guo, Eric Chan-Tin, and Yanmin Gong

***Abstract*—Alternating direction method of multipliers (ADMM) is a widely used tool for machine learning in distributed settings, where a machine learning model is trained over distributed data sources through an interactive process of local computation and message passing. Such an iterative process could cause privacy concerns of data owners. The goal of this paper is to provide differential privacy for ADMM-based distributed machine learning. Prior approaches on differentially private ADMM exhibit low utility under high privacy guarantee and assume the objective functions of the learning problems to be smooth and strongly convex. To address these concerns, we propose a novel differentially private ADMM-based distributed learning algorithm called DP-ADMM, which combines an approximate augmented Lagrangian function with time-varying Gaussian noise addition in the iterative process to achieve higher utility for general objective functions under the same differential privacy guarantee. We also apply the moments accountant method to analyze the end-to-end privacy loss. The theoretical analysis shows that DP-ADMM can be applied to a wider class of distributed learning problems, is provably convergent, and offers an explicit utility-privacy tradeoff. To our knowledge, this is the first paper to provide explicit convergence and utility properties for differentially private ADMM-based distributed learning algorithms. The evaluation results demonstrate that our approach can achieve good convergence and model accuracy under high end-to-end differential privacy guarantee.**

***Index Terms*—Machine learning, ADMM, distributed algorithms, privacy, differential privacy, and moments accountant.**

I. INTRODUCTION

DISTRIBUTED machine learning is a widely adopted approach due to the high demand of large-scale and distributed data processing. It allows multiple entities to keep their datasets unexposed, and meanwhile to collaborate in a common learning objective (usually formulated as a regularized empirical risk minimization problem) by iterative local computation and message passing. Therefore, distributed machine learning helps to reduce computational burden and improves both robustness and scalability of data processing. As pointed out in recent studies [1], [2], existing approaches to decentralizing an optimization problem mainly consist of subgradient-based algorithms [3], [4], alternating direction method of multipliers (ADMM) based algorithms [5]–[8], and composite of sub-gradient descent and ADMM [9]. It has been shown that ADMM-based algorithms can converge at the rate of $O(1/t)$ while subgradient-based algorithms typically converge at the rate of $O(1/\sqrt{t})$, where t is the number of iterations [10]. Therefore, ADMM has become a popular method for designing distributed versions of a machine learning algorithm [5], [8], [11], and our work focuses on ADMM-based distributed learning.

With ADMM, the learning problem is divided into several sub-problems solved by agents independently and locally, and only intermediate parameters need to be shared. However, the iterative process of ADMM involves privacy leakage, and the adversary can obtain the sensitive information from the shared model parameters as shown in [12], [13]. Thus, we aim to limit the privacy leakage during the iterative process of ADMM using differential privacy. Differential privacy is a widely used privacy definition [14]–[16] and can be guaranteed in ADMM through adding noise to the exchanged messages. However, in existing studies on ADMM-based distributed learning with differential privacy [1], [2], [17]–[19], noise addition would disrupt the learning process and severely degrade the performance of the trained model, especially when large noise is needed to provide high privacy protection. Besides, their privacy-preserving algorithms only apply to the learning problems with both smoothness and strongly convexity assumptions about the objective functions. Such weaknesses and limitations motivate us to explore further in this area.

In this paper, we mainly focus on using ADMM to enable distributed learning while guaranteeing differential privacy, and propose a novel differentially private ADMM-based distributed learning algorithm called DP-ADMM, which has good convergence properties, low computational cost, and an explicit and improved utility-privacy tradeoff, and can be applied to a wide class of distributed learning problems. The key algorithmic feature of DP-ADMM is the combination of an approximate augmented Lagrangian function and time-varying Gaussian noise addition in the iterative process, which enables the algorithm to be noise-resilient and provably convergent. The moments accountant method [20] is used to analyze the end-to-end privacy guarantee of DP-ADMM. We also rigorously analyze the convergence rate and utility bound of our approach. To our knowledge, this is the first paper to provide explicit convergence and utility properties for differentially private ADMM-based distributed learning algorithms.

The main contributions of this paper are summarized as follows:

The work of R. Hu and Y. Gong is supported by National Science Foundation under grant CNS-1850523. Z. Huang is with Oklahoma State University, Stillwater, OK 74075. E-mail: zonghao.huang@okstate.edu R. Hu, Y. Guo, and Y. Gong are with The University of Texas at San Antonio, San Antonio, TX 78249. Email:{rui.hu@my., yuanxiong.guo@, yanmin.gong@}utsa.edu E. Chan-Tin is with Loyola University Chicago, Chicago, IL 60660. E-mail: chantin@cs.luc.edu

This paper has supplementary downloadable material available at https://ieeexplore.ieee.org provided by the author. This includes a PDF file containing mathematical proofs.

1) We design a novel differentially private ADMM-based distributed learning algorithm called DP-ADMM, which combines an approximate augmented Lagrangian function with time-varying Gaussian noise addition in the iterative process to achieve higher utility for more general objective functions than prior works under the same differential privacy guarantee.
2) Different from previous studies providing only differential privacy guarantee for each iteration, we use the moments accountant method to analyze the total privacy loss and provide a tight end-to-end differential privacy guarantee for DP-ADMM.
3) We provide rigorous convergence and utility analysis of the proposed DP-ADMM. To our knowledge, this is the first paper to provide explicit convergence and utility properties for differentially private ADMM-based distributed learning algorithms.
4) We conduct extensive simulations based on real-world datasets to validate the effectiveness of DP-ADMM in distributed learning settings.

The rest of the paper is organized as follows. In Section II, we present our problem statement. In Section III, we describe a differentially private standard ADMM-based algorithm and propose our DP-ADMM. In Section IV and Section V, we theoretically analyze our privacy guarantee and convergence and utility properties of DP-ADMM, respectively. The numerical results of DP-ADMM based on real-world datasets are shown in Section VI. Section VII discusses the related work, and Section VIII concludes the paper.

II. Problem Statement

In this section, we first introduce the problem setting. Then we present the standard ADMM-based distributed learning algorithm and discuss the associated privacy concern. A summary of notations used in this paper is listed in Table I.

A. Problem Setting

We consider a set of agents $[n] := \{1, \ldots, n\}$ and a central aggregator. Each agent $i \in [n]$ has a private training dataset $\mathcal{D}_i := \{(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}) : \forall j \in [m_i]\}$, where m_i is the number of training samples in the dataset $\mathcal{D}_i$, $\boldsymbol{a}_{i,j} \in \mathbb{R}^d$ is the d-dimensional data feature vector of the j-th training sample, and $\boldsymbol{b}_{i,j} \in \mathbb{R}^p$ is the corresponding p-dimensional data label. In this paper, we consider a star network topology where each agent can communicate with the central aggregator and the aggregator is responsible for message passing and aggregation. Note that our approach can be generalized to other network topologies where agents are connected with their neighbors without a central aggregator, as discussed in [1], [2], [17].

The goal of our problem is to train a supervised learning model on the aggregated dataset $\{\mathcal{D}_i\}_{i\in[n]}$, which enables predicting a label for any new data feature vector. The learning objective can be formulated as the following regularized empirical risk minimization problem:

$$\min_{\boldsymbol{w}} \quad \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{1}{m_i}\ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \boldsymbol{w}) + \lambda R(\boldsymbol{w}), \tag{1}$$

TABLE I: List of notations

$\boldsymbol{a}_{i,j}$	Data feature vector
$\boldsymbol{b}_{i,j}$	Data label
$\ell(\cdot)$	Loss function
$R(\cdot)$	Regularizer function
λ	Regularizer parameter
$\ell^{'}(\cdot)$	Subgradient of loss function
$R^{'}(\cdot)$	Subgradient of regularizer
$\nabla\ell(\cdot)$	Gradient of loss function
$\nabla R(\cdot)$	Gradient of regularizer
$\boldsymbol{w}$	Global machine learning model
$\boldsymbol{w}_i$	Local learning model from agent i
$\boldsymbol{\gamma}_i$	Dual variable from agent i
ρ	Penalty parameter
$\mathcal{L}_\rho(\cdot)$	Augmented Lagrangian function
$\hat{\mathcal{L}}_{\rho,k}(\cdot)$	Approximate augmented Lagrangian function
$\boldsymbol{w}_i^k$	Primal variable from agent i in k-th iteration
$\tilde{\boldsymbol{w}}_i^k$	Noisy version of $\boldsymbol{w}_i^k$ after perturbation
$\boldsymbol{\gamma}_i^k$	Dual variable from agent i in k-th iteration
$\boldsymbol{w}^k$	Global variable in k-th iteration
$\boldsymbol{\xi}_i^k$	Sampled noise from agent i in k-th iteration
σ_i^2	Constant variance of Gaussian mechanism
η_i^k	Time-varying step size in k-th iteration
$\sigma_{i,k}^2$	Time-varying variance of Gaussian mechanism

where $\boldsymbol{w} \in \mathbb{R}^{d\times p}$ is the trained machine learning model, $\ell(\cdot) : \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^{d\times p} \to \mathbb{R}$ is the loss function used to measure the quality of the trained model, $R(\cdot)$ refers to the regularizer function introduced to prevent overfitting, and $\lambda > 0$ is the regularizer parameter controlling the impact of regularizer. Note that the problem formulation (1) can represent a wide range of machine learning tasks by choosing different loss functions. For instance, the loss function of binary logistic regression is:

$$\ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \boldsymbol{w}) = \ln\left(1 + \exp(-\boldsymbol{b}_{i,j}\boldsymbol{w}^\intercal \boldsymbol{a}_{i,j})\right), \tag{2}$$

and the loss function of multi-class logistic regression is:

$$\ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \boldsymbol{w}) = \sum_{h=1}^{p} \boldsymbol{b}_{i,j}^{(h)} \ln\left(\frac{\sum_{l=1}^{p}\exp(\boldsymbol{w}^{(l)\intercal}\boldsymbol{a}_{i,j})}{\exp(\boldsymbol{w}^{(h)\intercal}\boldsymbol{a}_{i,j})}\right). \tag{3}$$

In this paper, we assume that the loss function $\ell(\cdot)$ and the regularizer function $R(\cdot)$ are both convex but not necessarily smooth. Throughout this paper, we use $\ell^{'}(\cdot)$ and $R^{'}(\cdot)$ to denote the sub-gradient of $\ell(\cdot)$ and $R(\cdot)$ respectively. When we consider smooth functions, we use $\nabla\ell(\cdot)$ and $\nabla R(\cdot)$ instead.

B. ADMM-Based Distributed Learning Algorithm

To apply ADMM, we re-formulate the problem (1) as:

$$\min_{\{\boldsymbol{w}_i\}_{i\in[n]}} \quad \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \boldsymbol{w}_i) + \frac{\lambda}{n}R(\boldsymbol{w}_i)\right), \tag{4a}$$

$$\text{s.t.} \quad \boldsymbol{w}_i = \boldsymbol{w}, i = 1, \ldots, n, \tag{4b}$$

where $\boldsymbol{w}_i \in \mathbb{R}^{d\times p}$ is the local model, and $\boldsymbol{w} \in \mathbb{R}^{d\times p}$ is the global one. The objective function (4a) is decoupled and each agent only needs to minimize the sub-problem associated with its dataset. Constraints (4b) enforce that all the local models reach consensus finally.

In standard ADMM, the augmented Lagrangian function associated with the problem (4) is:

$$\mathcal{L}_\rho(\boldsymbol{w}, \{\boldsymbol{w}_i\}_{i\in[n]}, \{\boldsymbol{\gamma}_i\}_{i\in[n]}) = \sum_{i=1}^{n} \mathcal{L}_{\rho,i}(\boldsymbol{w}_i, \boldsymbol{w}, \boldsymbol{\gamma}_i), \quad (5)$$

where

$$\begin{aligned}\mathcal{L}_{\rho,i}(\boldsymbol{w}_i, \boldsymbol{w}, \boldsymbol{\gamma}_i) = &\sum_{j=1}^{m_i} \frac{1}{m_i} \ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \boldsymbol{w}_i) + \frac{\lambda}{n} R(\boldsymbol{w}_i) \\ &- \langle \boldsymbol{\gamma}_i, \boldsymbol{w}_i - \boldsymbol{w} \rangle + \frac{\rho}{2} \|\boldsymbol{w}_i - \boldsymbol{w}\|^2.\end{aligned} \quad (6)$$

In (6), $\{\boldsymbol{\gamma}_i\}_{i\in[n]} \in \mathbb{R}^{d\times p\times n}$ are the dual variables associated with constraints (4b) and $\rho > 0$ is the penalty parameter. The standard ADMM solves the problem (4) in a Gauss-Seidel manner by minimizing (5) w.r.t. $\{\boldsymbol{w}_i\}_{i\in[n]}$ and $\boldsymbol{w}$ alternatively followed by a dual update of $\{\boldsymbol{\gamma}_i\}_{i\in[n]}$. The ADMM-based distributed algorithm is shown in Algorithm 1.

Algorithm 1 ADMM-Based Distributed Algorithm

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

C. Privacy Concern

In Algorithm 1, the intermediate parameters $\{\boldsymbol{w}_i^k\}_{i\in[n],k\in[t]}$ need to be shared with the aggregator, which may reveal the agents' private information as demonstrated by model inversion attacks [21]. Thus, we need to develop privacy-preserving methods to control such information leakage. The main goal of this paper is to provide privacy protection against inference attacks from an adversary, who tries to infer sensitive information about the agents' private datasets from the shared messages. We assume that the adversary can neither intrude into the local datasets nor have access to the datasets directly. The adversary could be an outsider who eavesdrops the shared messages, or the honest-but-curious aggregator who follows the protocol honestly but tends to infer the sensitive information. We do not assume any trusted third party, thus a privacy-preserving mechanism should be applied locally by each agent to provide privacy protection.

In order to provide privacy guarantee against such attacks, we define our privacy model formally by the notion of differential privacy [14]. Specifically, we adopt the (ϵ, δ)-differential privacy defined as follows:

Definition 1 ((ϵ, δ)-Differential Privacy)**.** *A randomized mechanism $\mathcal{M}$ is (ϵ, δ)-differentially private if for any two neighbouring datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in only one tuple, and for any subsets of outputs $\mathcal{O} \subseteq$ range($\mathcal{M}$):*

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq e^\epsilon \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{O}] + \delta, \quad (7)$$

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

In Definition 1, the parameters δ and ϵ are privacy budgets indicating the strength of privacy protection from the mechanism. Smaller ϵ or δ indicates better privacy protection. Gaussian mechanism is a common randomization method used to guarantee (ϵ, δ)-differential privacy, where noise sampled from normal distribution is added to the output. In this paper, we use $\mathcal{MN}_{d,p}(0, \sigma^2\mathbf{I}_d, \sigma^2\mathbf{I}_p)$ to denote the matrix normal distribution with variance σ^2.

III. ADMM with Differential Privacy

In this section, we achieve differential privacy under the framework of ADMM. First, we introduce an intuitive method by directly combining standard ADMM and primal variable perturbation (PVP) and discuss the weaknesses of this method. Then we propose our new approach to achieving differential privacy in ADMM with an improved utility-privacy tradeoff.

A. ADMM with Primal Variable Perturbation (PVP)

As described in Section II, we need to use a local privacy-preserving mechanism in order to guarantee (ϵ, δ)-differential privacy for each agent. An intuitive way to achieve this goal is to combine the primal variable perturbation mechanism (PVP) and standard ADMM directly as proposed in [17]. Specifically, as given in Algorithm 2, at the k-th iteration, after obtaining the local primal variable $\boldsymbol{w}_i^k$, we apply Gaussian mechanism with a pre-defined variance σ_i^2 to perturb it and share the noisy primal variable $\tilde{\boldsymbol{w}}_i^k$, which can guarantee differential privacy. According to [22], [23], by assuming the smoothness of loss function $l(\cdot)$ and regularizer function $R(\cdot)$, strongly convexity of regularizer $R(\cdot)$, and the bounded l_2 norm of the derivative of loss function by c_1, the l_2 sensitivity of $\boldsymbol{w}_i^k$ update function in standard ADMM is $2c_1/\big(m_i(\lambda/n + \rho)\big)$ as proved in Appendix A. Therefore, the noise magnitude $\sigma_i = 2c_1\sqrt{2\ln(1.25/\delta)}/\big((\lambda/n + \rho)m_i\epsilon\big)$ can achieve (ϵ, δ)-differential privacy in each iteration.

Algorithm 2 ADMM with PVP

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

However, the added noise from the perturbation mechanism would disrupt the learning process, break the convergence property of the iterative process, and lead to a trained model with poor performance. This is especially the case when

the privacy budget is small. Specifically, when the iteration number k is large, the trained model would keep changing dramatically due to the existence of large noise. Besides, the above perturbation method can only be applied when the objective function is smooth and the regularizer is strongly convex [17], [23]. In order to address such problems, we need to consider an alternative way to preserving differential privacy of ADMM-based distributed learning algorithms.

B. Our Approach

Our approach is inspired by the intuition 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 our approach, instead of using the exact augmented Lagrangian function, we employ its first-order approximation with a scalar l_2-norm prox-function. Here we define:

$$\begin{aligned}&\hat{\mathcal{L}}_{\rho,k,i}(\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_i^{k-1}, \boldsymbol{w}, \boldsymbol{\gamma}_i)\\ &= \sum_{j=1}^{m_i} \frac{1}{m_i} \ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \tilde{\boldsymbol{w}}_i^{k-1}) + \frac{\lambda}{n} R(\tilde{\boldsymbol{w}}_i^{k-1})\\ &+ \big\langle \sum_{j=1}^{m_i} \frac{1}{m_i} \ell^{'}(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \tilde{\boldsymbol{w}}_i^{k-1}) + \frac{\lambda}{n} R^{'}(\tilde{\boldsymbol{w}}_i^{k-1}), \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1} \big\rangle\\ &- \big\langle \boldsymbol{\gamma}_i, \boldsymbol{w}_i - \boldsymbol{w} \big\rangle + \frac{\rho}{2} \|\boldsymbol{w}_i - \boldsymbol{w}\|^2 + \frac{\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2}{2\eta_i^k},\end{aligned} \tag{8}$$

where $\eta_i^k \in \mathbb{R}$ is the time-varying step size, and it decreases as the iteration number k increases.

The proposed approximate augmented Lagrangian function used in our approach is defined by:

$$\begin{aligned}&\hat{\mathcal{L}}_{\rho,k}(\{\boldsymbol{w}_i\}_{i\in[n]}, \{\tilde{\boldsymbol{w}}_i^{k-1}\}_{i\in[n]}, \boldsymbol{w}, \{\boldsymbol{\gamma}_i\}_{i\in[n]})\\ &\qquad = \sum_{i=1}^{n} \hat{\mathcal{L}}_{\rho,k,i}(\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_i^{k-1}, \boldsymbol{w}, \boldsymbol{\gamma}_i).\end{aligned} \tag{9}$$

Our approach minimizes (9) in a Gauss-Seidel manner and adds zero-mean Gaussian noise with time-varying variance $\sigma_{i,k}^2$ that decreases as the iteration number k increases.

The resulting ADMM steps that provide differential privacy are as follows:

$$\boldsymbol{w}_i^k = \underset{\boldsymbol{w}_i}{\operatorname{argmin}}\ \hat{\mathcal{L}}_{\rho,k,i}(\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_i^{k-1}, \boldsymbol{w}^{k-1}, \boldsymbol{\gamma}_i^{k-1}), \tag{10a}$$

$$\tilde{\boldsymbol{w}}_i^k = \boldsymbol{w}_i^k + \mathcal{MN}_{d,p}(0, \sigma_{i,k}^2 \mathbf{I}_d, \sigma_{i,k}^2 \mathbf{I}_p), \tag{10b}$$

$$\boldsymbol{w}^k = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{w}}_i^k - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\gamma}_i^{k-1}/\rho, \tag{10c}$$

$$\boldsymbol{\gamma}_i^k = \boldsymbol{\gamma}_i^{k-1} - \rho(\tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k), \tag{10d}$$

where (10c) is computed at the aggregator while (10a), (10b) and (10d) are performed at each agent.

The details are given in Algorithm 3. The central aggregator firstly initializes the global variable $\boldsymbol{w}^0$, and the agents also initialize their noisy primal variables $\{\tilde{\boldsymbol{w}}_i^0\}_{i\in[n]}$ and dual variables $\{\boldsymbol{\gamma}_i^0\}_{i\in[n]}$. At the beginning of each iteration k, each agent i first samples a zero-mean Gaussian noise $\boldsymbol{\xi}_i^k$ with variance $\sigma_{i,k}^2$ and updates the noisy primal variable $\tilde{\boldsymbol{w}}_i^k$ based on (10a) and (10b). Then the aggregator receives the noisy primal variables $\{\tilde{\boldsymbol{w}}_i^k\}_{i\in[n]}$ and the dual variables $\{\boldsymbol{\gamma}_i^{k-1}\}_{i\in[n]}$ from the agents, and uses them to update the global variable $\boldsymbol{w}^k$ according to (10c). After that, agents receive the updated global variable $\boldsymbol{w}^k$ from the aggregator and continue to update the dual variables $\{\boldsymbol{\gamma}_i^k\}_{i\in[n]}$ by (10d). The iterative process will continue until reaching t iterations.

Algorithm 3 DP-ADMM

Initialize $\boldsymbol{w}^0$, $\{\tilde{\boldsymbol{w}}_i^0\}_{i\in[n]}$, and $\{\boldsymbol{\gamma}_i^0\}_{i\in[n]}$.
for $k = 1, 2, \ldots, t$ **do**
  for $i = 1, 2, \ldots, n$ **do**
    $\boldsymbol{w}_i^k \leftarrow \operatorname{argmin}_{\boldsymbol{w}_i}\ \hat{\mathcal{L}}_{\rho,k,i}(\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_i^{k-1}, \boldsymbol{w}^{k-1}, \boldsymbol{\gamma}_i^{k-1})$.
    $\boldsymbol{\xi}_i^k \leftarrow \mathcal{MN}_{d,p}(0, \sigma_{i,k}^2 \mathbf{I}_d, \sigma_{i,k}^2 \mathbf{I}_p)$.
    $\tilde{\boldsymbol{w}}_i^k \leftarrow \boldsymbol{w}_i^k + \boldsymbol{\xi}_i^k$.
  end for
  $\boldsymbol{w}^k \leftarrow \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{w}}_i^k - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\gamma}_i^{k-1}/\rho$.
  for $i = 1, 2, \ldots, n$ **do**
    $\boldsymbol{\gamma}_i^k \leftarrow \boldsymbol{\gamma}_i^{k-1} - \rho(\tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k)$.
  end for
end for

Algorithm 3 is different from Algorithm 2 in three aspects. Firstly, the approximate augmented Lagrangian function used in this approach replaces the objective function with its first-order approximation at $\tilde{\boldsymbol{w}}_i^{k-1}$, which is similar to the stochastic mirror descent [24]. This approximation enforces the smoothness of the Lagrangian function and makes it easy to solve (10a). Even when the objective function is non-smooth, we can still get a closed-form solution to (10a), 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 smooth and strongly convex. Thus our approach can be applied to any convex problems. We demonstrate this in Section IV.

Secondly, similar to linearized ADMM [25], [26], there is an l_2-norm prox-function $\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2$ but scaled by $1/2\eta_i^k$ added in (8), where the step size η_i^k decreases when the iteration number k increases. Such additional part can guarantee the consistency between the updated model $\boldsymbol{w}_i^k$ and the previous one, especially when k is large. Thus, as k increases, the updated model would change more smoothly. Note that the time-varying step-size η_i^k is significant for the overall convergence guarantee. In Section V, we will define η_i^k and show its importance in algorithmic convergence.

Lastly, the variance $\sigma_{i,k}^2$ of Gaussian mechanism used in Algorithm 3 is time-varying rather than constant as adopted in prior studies [20]. It decreases when the iteration number k increases. The motivation of using Gaussian mechanism with time-varying variance is to mitigate the negative effect from noise and guarantee the convergence property of our approach. As explained before, the added noise would disrupt the learning process. By using the Gaussian mechanism with time-varying variance, the added noise will decrease when the iteration number k increases. Therefore, the negative affect from the added noise will be mitigated, enabling the updates

to be stable. In Section IV, we would define the magnitude of time-varying variance $\sigma_{i,k}^2$ to achieve differential privacy.

IV. Privacy Guarantee

In this section, we analyze the privacy guarantee of the proposed DP-ADMM. In DP-ADMM, the shared messages $\{\tilde{\boldsymbol{w}}_i^k\}_{k\in[t]}$ may reveal the sensitive information of agent i, which has been discussed in Section II. Thus, we need to demonstrate that DP-ADMM guarantees differential privacy with outputs $\{\tilde{\boldsymbol{w}}_i^k\}_{k\in[t]}$. We first estimate the l_2 norm sensitivity of $\boldsymbol{w}_i^k$ update function, then analyze the privacy leakage from the shared primal variable $\tilde{\boldsymbol{w}}_i^k$ in each iteration, and finally compute the end-to-end differential privacy guarantee across t iterations using the moments accountant method. Here we use $\boldsymbol{w}_{i,\mathcal{D}_i}^k$ and $\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k$ to denote the local primal variables updated from two neighboring datasets $\mathcal{D}_i$ and $\mathcal{D}_i^{'}$.

A. L_2-norm Sensitivity

In our approach, we apply Gaussian mechanism to add noise whose magnitude is calibrated by the l_2-norm sensitivity. Note that compared with Algorithm 2 and prior works [1], [2], [17], the derivation of the sensitivity in our proposed algorithm does not require the assumption of smoothness and strong convexity of the objective function due to the first-order approximation used in the approximate augmented Lagrangian function.

Lemma 1. *Assume that $\|\ell^{'}(\cdot)\| \leq c_1$. The l_2-norm sensitivity of local primal variable $\boldsymbol{w}_i^k$ update function is given by:*

$$\max_{\mathcal{D}_i,\mathcal{D}_i^{'}} \|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\| = \frac{2c_1}{m_i(\rho+1/\eta_i^k)}. \tag{11}$$

Proof. Since $\hat{\mathcal{L}}_{\rho,k,i}(\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_i^{k-1}, \boldsymbol{w}^{k-1}, \boldsymbol{\gamma}_i^{k-1})$ in the first step of DP-ADMM (10a) is a quadratic function w.r.t. $\boldsymbol{w}_i$ and therefore convex, we could obtain that:

$$\begin{aligned}\boldsymbol{w}_{i,\mathcal{D}_i}^k =& \Bigg(-\sum_{j=1}^{m_i} \frac{1}{m_i}\ell^{'}(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \tilde{\boldsymbol{w}}_i^{k-1}) - \frac{\lambda}{n}R^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) \\ &+\boldsymbol{\gamma}_i^{k-1} + \rho\boldsymbol{w}^{k-1} + \frac{\tilde{\boldsymbol{w}}_i^{k-1}}{\eta_i^k}\Bigg)\Big(\rho + 1/\eta_i^k\Big)^{-1}, \quad (12a)\\ \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k =& \Bigg(-\sum_{j=1}^{m_i-1} \frac{1}{m_i}\ell^{'}(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \tilde{\boldsymbol{w}}_i^{k-1}) \\ &-\frac{1}{m_i}\ell^{'}(a_{i,m_i}^{'}, b_{i,m_i}^{'}, \tilde{\boldsymbol{w}}_i^{k-1}) - \frac{\lambda}{n}R^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) \\ &+\boldsymbol{\gamma}_i^{k-1} + \rho\boldsymbol{w}^{k-1} + \frac{\tilde{\boldsymbol{w}}_i^{k-1}}{\eta_i^k}\Bigg)\Big(\rho + 1/\eta_i^k\Big)^{-1}, \quad (12b)\end{aligned}$$

by computing the derivative of (8) with inputs $\boldsymbol{w}^{k-1}$ and $\boldsymbol{\gamma}_i^{k-1}$ and letting $\nabla\hat{\mathcal{L}}_{\rho,k,i}(\boldsymbol{w}_i, \tilde{\boldsymbol{w}}_i^{k-1}, \boldsymbol{w}^{k-1}, \boldsymbol{\gamma}_i^{k-1})$ to be 0.

With $\boldsymbol{w}_{i,\mathcal{D}_i}^k$ and $\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k$ calculated by (12a) and (12b) respectively, the l_2-norm sensitivity of primal variable $\boldsymbol{w}_i^k$ update function is defined by:

$$\begin{aligned}&\max_{\mathcal{D}_i,\mathcal{D}_i^{'}} \|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\| \\ &= \max_{\mathcal{D}_i,\mathcal{D}_i^{'}} \frac{\big\|\ell^{'}(\boldsymbol{a}_{i,m_i}, \boldsymbol{b}_{i,m_i}, \tilde{\boldsymbol{w}}_i^{k-1}) - \ell^{'}(\boldsymbol{a}_{i,m_i}^{'}, \boldsymbol{b}_{i,m_i}^{'}, \tilde{\boldsymbol{w}}_i^{k-1})\big\|}{m_i(\rho+1/\eta_i^k)}.\end{aligned} \tag{13}$$

Since $\|\ell^{'}(\cdot)\|$ is bounded by c_1, the sensitivity of $\boldsymbol{w}_i^k$ update function is given by $2c_1/\big(m_i(\rho+1/\eta_i^k)\big)$. □

Lemma 1 shows that the sensitivity of $\boldsymbol{w}_i^k$ update function in our approach is affected by the time-varying η_i^k. When we set η_i^k 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 in spite of the existence of the noise.

B. (ϵ, δ)-Differential Privacy Guarantee

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

Theorem 1. *Assume that $\|\ell^{'}(\cdot)\| \leq c_1$. Let $\epsilon \in (0, 1]$ be arbitrary and $\boldsymbol{\xi}_i^k$ be the noise sampled from Gaussian mechanism with variance $\sigma_{i,k}^2$ where*

$$\sigma_{i,k} = \frac{2c_1\sqrt{2\ln(1.25/\delta)}}{m_i\epsilon(\rho+1/\eta_i^k)}. \tag{14}$$

Each iteration of DP-ADMM guarantees (ϵ, δ)-differential privacy. Specifically, for any neighboring datasets $\mathcal{D}_i$ and $\mathcal{D}_i^{'}$, for any output $\tilde{\boldsymbol{w}}_i^k$, the following inequality always holds:

$$\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i] \leq e^{\epsilon}\cdot\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i^{'}] + \delta. \tag{15}$$

Proof. The privacy loss from $\tilde{\boldsymbol{w}}_i^k$ is calculated as

$$\left|\ln\frac{\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i]}{\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i^{'}]}\right| = \left|\ln\frac{\Pr[\tilde{w}_i^{k^{(h,l)}}|\mathcal{D}_i]}{\Pr[\tilde{w}_i^{k^{(h,l)}}|\mathcal{D}_i^{'}]}\right| = \left|\ln\frac{\Pr[\xi_i^{k^{(h,l)}}]}{\Pr[\xi_i^{k,{'}^{(h,l)}}]}\right|, \tag{16}$$

where $\xi_i^{k^{(h,l)}}$ and $\xi_i^{k,{'}^{(h,l)}}$ are the (h,l)-entry of $\boldsymbol{\xi}_i^k$ and $\boldsymbol{\xi}_i^{k,{'}}$, and are sampled from $\mathcal{N}(0, \sigma_{i,k}^2)$. This leads to:

$$\begin{aligned}&\left|\ln\frac{\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i]}{\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i^{'}]}\right| = \Big|\frac{1}{2\sigma_{i,k}^2}\big(\|\xi_i^{k^{(h,l)}}\|^2 - \|\xi_i^{k,{'}^{(h,l)}}\|^2\big)\Big| \\ =&\Big|\frac{1}{2\sigma_{i,k}^2}\big(\|\xi_i^{k^{(h,l)}}\|^2 - \|\xi_i^{k^{(h,l)}} + (w_{i,\mathcal{D}_i}^{k^{(h,l)}} - w_{i,\mathcal{D}_i^{'}}^{k^{(h,l)}})\|^2\big)\Big| \\ =&\Big|\frac{1}{2\sigma_{i,k}^2}\big(2\xi_i^{k^{(h,l)}}\|w_{i,\mathcal{D}_i}^{k^{(h,l)}} - w_{i,\mathcal{D}_i^{'}}^{k^{(h,l)}}\| + \|w_{i,\mathcal{D}_i}^{k^{(h,l)}} - w_{i,\mathcal{D}_i^{'}}^{k^{(h,l)}}\|^2\big)\Big|.\end{aligned} \tag{17}$$

Since $\|\ell^{'}(\cdot)\| \leq c_1$, according to Lemma 1, we have $\|w_{i,\mathcal{D}_i}^{k^{(h,l)}} - w_{i,\mathcal{D}_i^{'}}^{k^{(h,l)}}\| < \|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\| \leq 2c_1/\big(m_i(\rho+1/\eta_i^k)\big)$. Thus, by letting $\sigma_{i,k} = 2c_1\sqrt{2\ln(1.25/\delta)}/\big(m_i\epsilon(\rho+1/\eta_i^k)\big)$, we have

$$\left|\ln\frac{\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i]}{\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i^{'}]}\right| \leq \left|\frac{\xi_i^{k^{(h,l)}}m_i(\rho+1/\eta_i^k) + c_1}{4\ln(1.25/\delta)c_1/\epsilon^2}\right|. \tag{18}$$

When $|\xi_i^{k^{(h,l)}}| \leq \big(4\ln(1.25/\delta)c_1/\epsilon - c_1\big)/\big(\epsilon m_i(\rho+1/\eta_i^k)\big)$, $\big|\ln\big(\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i]/\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i^{'}]\big)\big|$ is bounded by ϵ. Next, we need to prove that $\Pr\big[|\xi_i^{k^{(h,l)}}| > \big(4\ln(1.25/\delta)c_1/\epsilon - c_1\big)/\big(\epsilon m_i(\rho+1/\eta_i^k)\big)\big] \leq \delta$, which requires $\Pr\big[\xi_i^{k^{(h,l)}} > \big(4\ln(1.25/\delta)c_1/\epsilon - c_1\big)/\big(\epsilon m_i(\rho+1/\eta_i^k)\big)\big] \leq \delta/2$. According to the tail bound of normal distribution $\mathcal{N}(0, \sigma_{i,k}^2)$, we have

$$\Pr\big[\xi_i^{k^{(h,l)}} > r\big] \leq \frac{\sigma_{i,k}}{r\sqrt{2\pi}}e^{-r^2/2\sigma_{i,k}^2}. \tag{19}$$

By letting $r = \big(4\ln(1.25/\delta)c_1/\epsilon - c_1\big)/\big(\epsilon m_i(\rho + 1/\eta_i^k)\big)$ in the above inequality, we have:

$$\begin{aligned}&\Pr\left[\xi_i^{k^{(h,l)}} > \frac{4\ln(1.25/\delta)c_1/\epsilon - c_1}{m_i(\rho + 1/\eta_i^k)}\right]\\ \leq&\frac{2\sqrt{2\ln(1.25/\delta)}}{(4\ln(1.25/\delta) - \epsilon)\sqrt{2\pi}}\exp\left(-\frac{(4\ln(1.25/\delta) - \epsilon)^2}{8\ln(1.25/\delta)}\right).\end{aligned} \tag{20}$$

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

$$\frac{2\sqrt{2\ln(1.25/\delta)}}{(4\ln(1.25/\delta) - \epsilon)\sqrt{2\pi}} < \frac{1}{\sqrt{2\pi}}, \tag{21}$$

and

$$-\frac{\big(4\ln(1.25/\delta) - \epsilon\big)^2}{8\ln(1.25/\delta)} < \ln(\sqrt{2\pi}\frac{\delta}{2}). \tag{22}$$

As a result, we have:

$$\Pr\left[\xi_i^{k^{(h,l)}} > \frac{4\ln(1.25/\delta)c_1/\epsilon - c_1}{m_i(\rho + 1/\eta_i^k)}\right] < \frac{\delta}{2}. \tag{23}$$

So far we have proved that $\Pr\big[\xi_i^{k^{(h,l)}} > \big(4\ln(1.25/\delta)c_1/\epsilon - c_1\big)/\big(\epsilon m_i(\rho + 1/\eta_i^k)\big)\big] \leq \delta/2$, thus we can prove that $\Pr\big[|\xi_i^{k^{(h,l)}}| > \big(4\ln(1.25/\delta)c_1/\epsilon - c_1\big)/\big(\epsilon m_i(\rho + 1/\eta_i^k)\big)\big] \leq \delta$. We define:

$$\mathbb{A}_1 = \{\xi_i^{k^{(h,l)}} : |\xi_i^{k^{(h,l)}}| \leq \frac{4\ln(1.25/\delta)c_1/\epsilon - c_1}{m_i(\rho + 1/\eta_i^k)}\}, \tag{24a}$$

$$\mathbb{A}_2 = \{\xi_i^{k^{(h,l)}} : |\xi_i^{k^{(h,l)}}| > \frac{4\ln(1.25/\delta)c_1/\epsilon - c_1}{m_i(\rho + 1/\eta_i^k)}\}. \tag{24b}$$

Therefore, we obtain the result:

$$\begin{aligned}\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i] =& \Pr[w_{i,\mathcal{D}_i}^{k^{(h,l)}} + \xi_i^{k^{(h,l)}} : \xi_i^{k^{(h,l)}} \in \mathbb{A}_1]\\ &+ \Pr[w_{i,\mathcal{D}_i}^{k^{(h,l)}} + \xi_i^{k^{(h,l)}} : \xi_i^{k^{(h,l)}} \in \mathbb{A}_2]\\ <& e^{\epsilon}\cdot\Pr[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i^{'}] + \delta,\end{aligned} \tag{25}$$

which proves that each iteration of DP-ADMM guarantees (ϵ, δ)-differential privacy. □

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 prior studies [20], [27] and use the moments accountant method to analyze the total privacy leakage.

Theorem 2 (Advanced Composition Theorem)**.** *Assume $\|\ell^{'}(\cdot)\| \leq c_1$. Let $\epsilon \in (0, 1]$ be arbitrary and $\boldsymbol{\xi}_i^k$ be sampled from Gaussian mechanism with variance $\sigma_{i,k}^2$ where*

$$\sigma_{i,k} = \frac{2c_1\sqrt{2\ln(1.25/\delta)}}{m_i\epsilon(\rho + 1/\eta_i^k)}. \tag{26}$$

Then Algorithm 3 guarantees $(\bar{\epsilon}, \delta)$-differential privacy, where $\bar{\epsilon} = c_0\sqrt{t}\epsilon$ for some constant c_0.

Proof. See Appendix B. □

V. Convergence Analysis

In this section, we analyze the convergence of the proposed DP-ADMM. Let $\boldsymbol{w}^*$ denote the optimal solution of problem (4), and c_w denote $\|\boldsymbol{w}^*\|$. Firstly, we analyze the convergence property based on the general assumption that the objective function is convex and non-smooth. Secondly, we refine the convergence property under a stricter assumption that the objective function is convex and smooth.

We define the following notations to be used for the analysis:

$$f_i(\boldsymbol{w}_i) = \sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \boldsymbol{w}_i) + \frac{\lambda}{n}R(\boldsymbol{w}_i),$$

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

$$\boldsymbol{u}_i^k = \begin{bmatrix}\tilde{\boldsymbol{w}}_i^k\\ \boldsymbol{w}^k\\ \boldsymbol{\gamma}_i^k\end{bmatrix}, \quad \boldsymbol{u}_i = \begin{bmatrix}\boldsymbol{w}_i\\ \boldsymbol{w}\\ \boldsymbol{\gamma}_i\end{bmatrix}, \quad F(\boldsymbol{u}_i^k) = \begin{bmatrix}-\boldsymbol{\gamma}_i^k\\ \boldsymbol{\gamma}_i^k\\ \tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k\end{bmatrix}.$$

We show that DP-ADMM achieves an $O(1/\sqrt{t})$ rate of convergence in terms of both the objective value and the constraint violation: $\sum_{i=1}^n\big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}^*) + \beta\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\|\big)$, where $\sum_{i=1}^n\big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}^*)\big)$ represents the distance between the current objective value and the optimal value while $\sum_{i=1}^n\beta\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\|$ measures the difference between the local model and the global one. Therefore, when we have $\sum_{i=1}^n\big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}^*) + \beta\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\|\big) = 0$, our training result converges to the optimal one and all local models reach consensus.

A. Non-Smooth Convex Objective Function

In this section, we analyze the convergence when the objective function is convex but non-smooth. We firstly analyze a single iteration of our algorithm in Lemma 2 and then give the convergence result of DP-ADMM in Theorem 3.

Lemma 2. *Assume $\ell(\cdot)$ and $R(\cdot)$ are convex. For any $k \geq 1$, we have:*

$$\begin{aligned}&\sum_{i=1}^n\Big(f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + (\boldsymbol{u}_i^k - \boldsymbol{u}_i)^{\intercal}F(\boldsymbol{u}_i^k)\Big)\\ \leq&\sum_{i=1}^n\bigg(\frac{\eta_i^k}{2}\big\|f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\big\|^2 - \frac{\rho}{2}\|\boldsymbol{w}_i - \boldsymbol{w}^k\|^2\\ &+ \frac{\rho}{2}\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2 - \big(\rho + 1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k, \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\\ &+ \frac{1}{2\eta_i^k}\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2 - \frac{1}{2\eta_i^k}\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^k\|^2\\ &+ \frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^{k-1}\|^2 - \frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^k\|^2\bigg).\end{aligned} \tag{28}$$

Proof. See Appendix D. □

Based on Lemma 2, we give the following convergence theorem.

Theorem 3. *Assume $\ell(\cdot)$ and $R(\cdot)$ are convex, $\|\ell^{'}(\cdot)\| \leq c_1$, and $\|R^{'}(\cdot)\| \leq c_2$. Let*

$$\eta_i^k = \frac{c_w}{\sqrt{2k}}\left((c_1+\lambda c_2/n)^2 + \frac{8dpc_1^2\ln(1.25/\delta)}{m_i^2\epsilon^2}\right)^{-\frac{1}{2}}. \quad (29)$$

Define

$$M_1(\epsilon,\delta) = \sum_{i=1}^{n} c_w\sqrt{2(c_1+\lambda c_2/n)^2 + \frac{16dpc_1^2\ln(1.25/\delta)}{m_i^2\epsilon^2}}, \quad (30)$$

and

$$M_2 = \frac{n(\rho c_w^2+\beta^2/\rho)}{2}. \quad (31)$$

For any $t \geq 1$ and β, we have:

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{n}\left(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}^*) + \beta\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\|\right)\right]\\ \leq\ & \frac{M_1(\epsilon,\delta)}{\sqrt{t}} + \frac{M_2}{t}.\end{aligned} \quad (32)$$

Proof. See Appendix E. □

Theorem 3 shows an explicit utility-privacy trade-off of our approach: when privacy guarantee is weaker (larger ϵ and δ), our approach has better utility. In addition, it demonstrates that our algorithm converges at a rate of $O(1/\sqrt{t})$.

B. Smooth Convex Objective Function

In this section, we refine Theorem 3 under a stricter assumption that $\ell(\cdot)$ and $R(\cdot)$ are both convex and smooth. Here, we replace the definition of $\bar{\boldsymbol{w}}_i^t$: $\bar{\boldsymbol{w}}_i^t = \frac{1}{t}\sum_{k=0}^{t-1}\tilde{\boldsymbol{w}}_i^k$ by $\bar{\boldsymbol{w}}_i^t = \frac{1}{t}\sum_{k=1}^{t}\tilde{\boldsymbol{w}}_i^k$. Similar to Section V-A, we first focus on a single iteration and then give the final convergence result.

Lemma 3. *Assume $\ell(\cdot)$ and $R(\cdot)$ are convex and smooth, $\|\nabla^2\ell(\cdot)\| \leq c_3$, and $\|\nabla^2 R(\cdot)\| \leq c_4$. For any $k \geq 1$, we have:*

$$\begin{aligned}&\sum_{i=1}^{n}\left(f_i(\tilde{\boldsymbol{w}}_i^k) - f_i(\boldsymbol{w}_i) + (\boldsymbol{u}_i^k - \boldsymbol{u}_i)^{\intercal}F(\boldsymbol{u}_i^k)\right)\\ \leq&\sum_{i=1}^{n}\Bigg(\frac{\left(\rho+1/\eta_i^k\right)^2}{2/\eta_i^k - 2(c_3+\lambda c_4/n)}\left\|\boldsymbol{\xi}_i^k\right\|^2 - \frac{1}{2\eta_i^k}\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^k\|^2\\ &+\frac{1}{2\eta_i^k}\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2 - \left(\rho+1/\eta_i^k\right)\langle\boldsymbol{\xi}_i^k, \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\rangle\\ &+\frac{\rho}{2}\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2 - \frac{\rho}{2}\|\boldsymbol{w}_i - \boldsymbol{w}^k\|^2\\ &+\frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^{k-1}\|^2 - \frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^k\|^2\Bigg).\end{aligned} \quad (33)$$

Proof. See Appendix F. □

Based on Lemma 3, we give the following theorem.

Theorem 4. *Assume $\ell(\cdot)$ and $R(\cdot)$ are convex and smooth, $\|\nabla^2\ell(\cdot)\| \leq c_3$, and $\|\nabla^2 R(\cdot)\| \leq c_4$. Let*

$$\eta_i^k = \left(c_3 + \lambda c_4/n + \frac{4c_1\sqrt{dpk\ln(1.25/\delta)}}{m_i\epsilon c_w}\right)^{-1}. \quad (34)$$

Define

$$M_3(\epsilon,\delta) = \sum_{i=1}^{n}\frac{4c_w c_1\sqrt{dp\ln(1.25/\delta)}}{m_i\epsilon}, \quad (35)$$

and

$$M_4 = \frac{nc_w^2(c_3+\lambda c_4/n+\rho)+n\beta^2/\rho}{2}. \quad (36)$$

For any $t \geq 1$ and β, we have:

$$\mathbb{E}\left[\sum_{i=1}^{n}\left(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}^*) + \beta\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\|\right)\right] \leq \frac{M_3(\epsilon,\delta)}{\sqrt{t}} + \frac{M_4}{t}. \quad (37)$$

Proof. See Appendix G. □

Theorem 4 also shows an explicit relation between the privacy budget (i.e., ϵ and δ) and the utility of our approach with smoothness, and demonstrates that the result from our algorithm converges to the optimal result at a rate of $O(1/\sqrt{t})$.

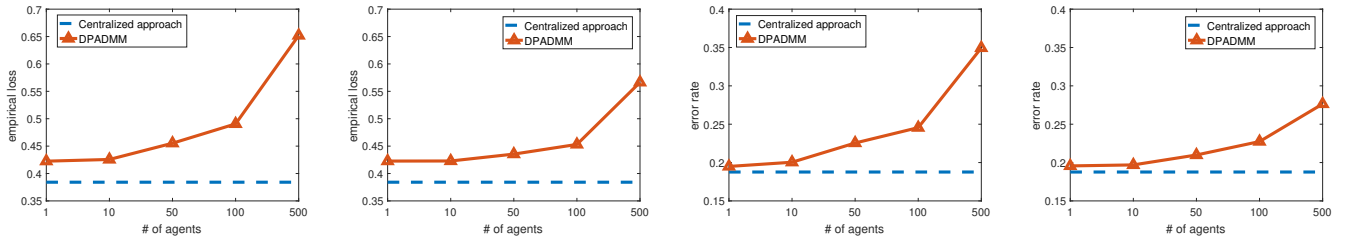
VI. Performance Evaluation

In this section, we evaluate the performance of DP-ADMM with both non-smooth objectives and smooth objectives by considering logistic regression problems with l_1-norm and l_2-norm regularizers, respectively.

Dataset. We evaluate our approach on a real-world dataset: Adult dataset [28] from UCI Machine Learning Repository. Adult dataset includes $48,842$ instances. Each instance has 14 attributes such as age, sex, education, occupation, marital status, and native country, and is associated with a label representing whether the income is above $\$50,000$ or not. Before the simulation, we firstly preprocess the data by removing all the instances with missing values, converting the categorical attributes into binary vectors, normalizing columns to guarantee the maximum value of each column is 1, normalizing rows to enforce their l_2 norm to be less than 1, and converting the labels $\{>50k, <50k\}$ into $\{+1,-1\}$. After this, we obtain $45,222$ entries each with a 104-dimensional feature vector ($d=104$) and a 1-dimensional label belonging to $\{+1,-1\}$ ($p=1$). In each simulation, we sample $40,000$ instances for training, and the remaining $5,222$ instances for testing. In the training process, we divide the training data into n groups randomly, and thus each group contains $40000/n$ data points ($m_i = 40000/n$).

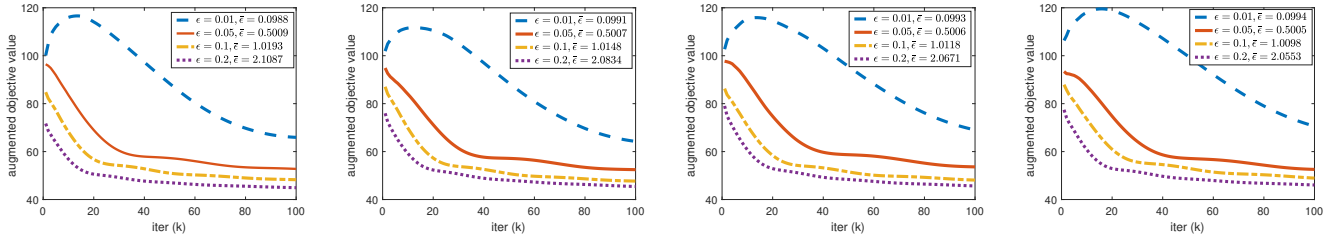
Baseline algorithms. We compare our DP-ADMM (Algorithm 3) with five baseline algorithms: (1) non-private centralized approach, (2) ADMM algorithm (Algorithm 1), (3) ADMM algorithm with PVP (Algorithm 2), (4) ADMM with dual variable perturbation (DVP) in [17], and (5) differentially private stochastic gradient descent (DPSGD) in [20] for distributed settings. We evaluate the accuracy and effectiveness of our approach by comparing it with the five baseline algorithms.

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 the total iteration number $t = 100$ and the penalty parameter $\rho = 0.1$, and choose the optimal regularizer parameter λ/n to be 10^{-6} by 10-cross-validation in non-private setting. In DPSGD, we set the optimal learning



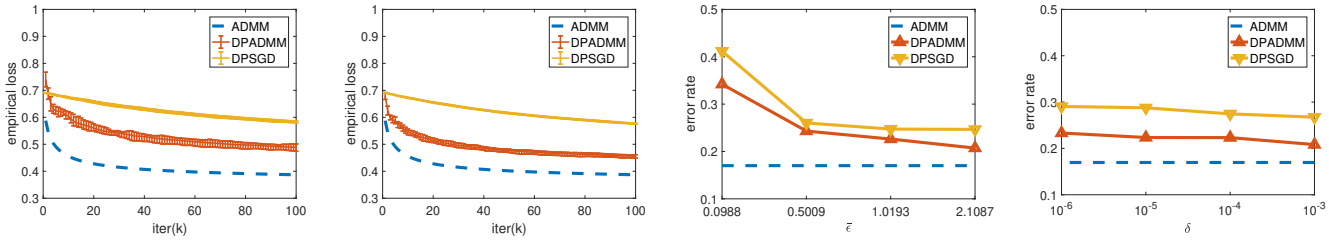
(a) $\epsilon = 0.05, \bar{\epsilon} = 0.5009, \delta = 10^{-3}$ (b) $\epsilon = 0.1, \bar{\epsilon} = 1.0193, \delta = 10^{-3}$ (c) $\epsilon = 0.05, \bar{\epsilon} = 0.5009, \delta = 10^{-3}$ (d) $\epsilon = 0.1, \bar{\epsilon} = 1.0193, \delta = 10^{-3}$

Fig. 1: Impact of distributed data source number on DP-ADMM (l_1-regularized logistic regression).



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

Fig. 2: Convergence properties of DP-ADMM (l_1-regularized logistic regression).



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

Fig. 3: Accuracy comparison in empirical loss and classification error rate (l_1-regularized logistic regression).

TABLE II: Computation Time (100 iterations).

	ADMM	PVP	DVP	DPADMM
$\epsilon = 0.01$	67.242s	102.282s	59.743s	6.937s
$\epsilon = 0.05$	67.242s	78.798s	65.935s	5.322s
$\epsilon = 0.1$	67.242s	79.013s	69.855s	5.218s

rate to be 0.1 and the sampling ratio to be 1. We focus on the settings with strong privacy guarantee and thus we set privacy budget per iteration $\epsilon = \{0.01, 0.05, 0.1, 0.2\}$ and $\delta = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, and use moments accountant method to obtain the corresponding total privacy loss $\bar{\epsilon}$. In each simulation, we run it for 10 times to get averaged result.

Evaluations. We consider logistic regression problem in a distributed setting and evaluate our approach for logistic regression problems with l_1-norm and l_2-norm regularizers respectively, in terms of convergence, accuracy, and computation cost. The loss function of binary logistic regression is defined by (2). The convergence properties are evaluated with respect to the augmented objective value, which measures the loss as well as the constraint penalty and is defined as $\sum_{i=1}^{n} \left(f_i(\bar{\boldsymbol{w}}_i^k) + \rho \|\bar{\boldsymbol{w}}_i^k - \bar{\boldsymbol{w}}^k\| \right)$. We evaluate the accuracy by empirical loss $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{m_i}\ell(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{i,j}, \tilde{\boldsymbol{w}}_i^k)$, and classification error rate. We measure the computation cost using the running time of training.

A. L_1-Regularized Logistic Regression

We obtain the DP-ADMM steps for l_1 regularized logistic regression by:

$$\boldsymbol{w}_i^k = \left(\frac{1}{m_i} \sum_{j=1}^{m_i} \frac{\boldsymbol{b}_{i,j}\boldsymbol{a}_{i,j}}{1+\exp(\boldsymbol{b}_{i,j}\tilde{\boldsymbol{w}}_i^{k-1\intercal}\boldsymbol{a}_{i,j})} - \frac{\lambda}{n}\mathrm{sgn}(\tilde{\boldsymbol{w}}_i^{k-1}) + \boldsymbol{\gamma}_i^{k-1} + \rho\boldsymbol{w}^{k-1} + \tilde{\boldsymbol{w}}_i^{k-1}/\eta_i^k \right) \left(\rho + 1/\eta_i^k\right)^{-1}, \quad (38a)$$

$$\tilde{\boldsymbol{w}}_i^k = \boldsymbol{w}_i^k + \mathcal{MN}_{d,p}(0, \sigma_{i,k}^2\mathbf{I}_d, \sigma_{i,k}^2\mathbf{I}_p), \quad (38b)$$

$$\boldsymbol{w}^k = \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{w}}_i^k - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}_i^{k-1}/\rho, \quad (38c)$$

$$\boldsymbol{\gamma}_i^k = \boldsymbol{\gamma}_i^{k-1} - \rho\left(\tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k\right), \quad (38d)$$

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

Since the l_1 regularized objective function is convex but non-smooth, we apply Theorem 3 to set η_i^k. Since we enforce $\|\ell'(\cdot)\| \leq 1$ by data preprocessing, and we have

$\|R^{'}(\cdot)\| \leq \sqrt{dp}$ ($d = 104$ and $p = 1$), we set $c_1 = 1$ and $c_2 = \sqrt{104}$. We obtain $\boldsymbol{w}^*$ by pre-training and set c_w to be 23. According to Theorem 3, we set η_i^k to be $23\big(2k(1+10^{-6}\sqrt{104}/n)^2 + 1664k\ln(1.25/\delta)/\big(m_i^2\epsilon^2\big)\big)^{-\frac{1}{2}}$.

Since PVP and DVP cannot be applied when the objective function is non-smooth, we only compare our approach with ADMM and DPSGD in this section. We first investigate the performance of our approach with different numbers of distributed data sources and compare it with the centralized approach. Figure 1 shows that the accuracy of our training model would decrease if we consider larger number of data sources. Since the size of local dataset is smaller for larger number of agents, more noise should be introduced to guarantee the same level of differential privacy, thus degrading the performance of the trained model. This is consistent with Theorem 1 that the noise magnitude is scaled by $1/m_i$. In following simulations, we consider the case when the number of agents n equals 100. Figure 2 demonstrates the convergence properties of our approach by showing how the augmented objective value converges for different ϵ and δ. It shows that our approach with larger ϵ and larger δ has better convergence, which is consistent with Theorem 3. Finally, we evaluate the accuracy of our approach by empirical loss and classification error rate by comparing with ADMM and DPSGD. Figure 3 shows our approach outperforms DPSGD due to the faster convergence property, demonstrating the advantage of ADMM framework. In addition, Figure 3 shows the privacy-utility trade-off of our approach. When privacy leakage increases (larger ϵ and larger δ), our approach achieves better utility.

B. L_2-Regularized Logistic Regression

The DP-ADMM steps for l_2 regularized logistic regression are described as follows:

$$\boldsymbol{w}_i^k = \bigg(\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{\boldsymbol{b}_{i,j}\boldsymbol{a}_{i,j}}{1+\exp(\boldsymbol{b}_{i,j}\tilde{\boldsymbol{w}}_i^{k-1\intercal}\boldsymbol{a}_{i,j})} - \frac{\lambda}{n}\tilde{\boldsymbol{w}}_i^{k-1} + \boldsymbol{\gamma}_i^{k-1} + \rho\boldsymbol{w}^{k-1} + \tilde{\boldsymbol{w}}_i^{k-1}/\eta_i^k\bigg)\bigg(\rho + 1/\eta_i^k\bigg)^{-1}, \quad (39a)$$

$$\tilde{\boldsymbol{w}}_i^k = \boldsymbol{w}_i^k + \mathcal{MN}_{d,p}(0, \sigma_{i,k}^2\mathbf{I}_d, \sigma_{i,k}^2\mathbf{I}_p), \quad (39b)$$

$$\boldsymbol{w}^k = \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{w}}_i^k - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}_i^{k-1}/\rho, \quad (39c)$$

$$\boldsymbol{\gamma}_i^k = \boldsymbol{\gamma}_i^{k-1} - \rho\left(\tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k\right). \quad (39d)$$

Here the l_2 regularized objective function is convex and smooth, thus we apply Theorem 4 to set η_i^k. Since we have $\|\nabla^2 R(\cdot)\| \leq 1$, and we enforce $\|\nabla\ell(\cdot)\| \leq 1$ and $\|\nabla^2\ell(\cdot)\| \leq 0.25$ by data preprocessing, thus we set $c_1 = 1$, $c_3 = 0.25$, and $c_4 = 1$. We obtain the optimal solution $\boldsymbol{w}^*$ by pre-training, and set c_w to be 89. According to Theorem 4, we set η_i^k to be $\big(0.25 + 10^{-6} + 2\sqrt{416k\ln(1.25/\delta)}/\big(89m_i\epsilon\big)\big)^{-1}$.

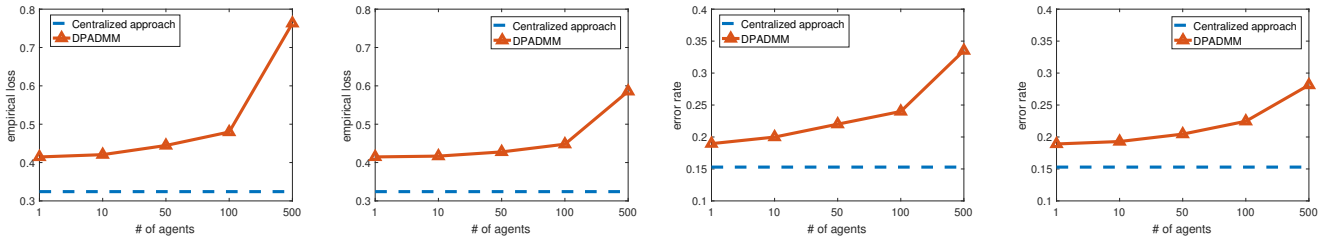
We fist investigate the performance of our approach under the settings with different numbers of distributed data sources and Figure 4 depicts the corresponding accuracy changes (accuracy decreases with increasing number of agents). Since the total data size is fixed, when we consider a larger number of agents, the size of local dataset is smaller, so the training model has lower accuracy due to more added noise for the same level of privacy guarantee. In the following simulations, we focus on the case where the number of agents is 100. Next, we show the convergence properties of our approach. Figure 5 demonstrates that under weaker privacy guarantee (larger ϵ and larger δ), our approach has better convergence, which is consistent with Theorem 4. We evaluate the accuracy of our approach by comparing it with ADMM, PVP, DVP, and DPSGD on empirical loss and classification error rate. Figure 6 shows that our approach outperforms PVP, DVP, and DPSGD. Specifically, ADMM has fast convergence but is sensitive to noise. Thus the methods directly perturbing intermediate results in ADMM (PVP and DVP) have poor performance. Gradient-based method (DPSGD) has good noise-resilience property but converges slowly. Our approach is based on ADMM framework, and combines the approximate augmented Lagrangian function with time-varying Gaussian noise addition to achieve higher utility. Furthermore, the results in Figure 6 also show the utility-privacy trade-off of our approach: larger ϵ and larger δ indicating weaker privacy guarantee would result in better utility. Finally, we show the advantage of our approach in computation cost by running time. Table II gives the comparison and shows that DP-ADMM has much less computation cost than all three ADMM baseline algorithms, which is resulted from the first-order approximation used in our approach enabling updates with closed-form solutions.

VII. Related work

The existing literature related to our work could be categorized by: privacy-preserving empirical risk minimization, privacy-preserving distributed learning, and variants of ADMM.

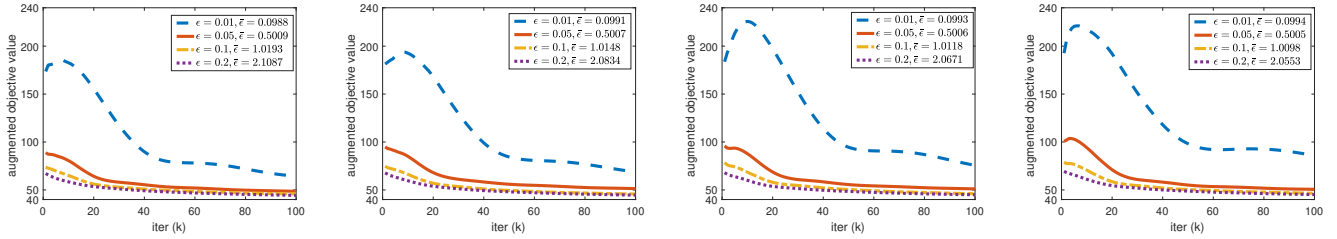
Privacy-preserving empirical risk minimization. There have been tremendous research efforts on privacy-preserving empirical risk minimization [23], [29]–[31]. Most of them focus on a centralized setting where sensitive data is collected and stored centrally, thus the privacy leakage comes from the final released trained model. Chaudhuri et al. [23] propose two perturbation methods: output perturbation and objective perturbation to guarantee ϵ-differential privacy. Bassily et al. [29] provide a systematic investigation of differentially private algorithms for convex empirical risk minimization and propose efficient algorithms with tighter error bound. Wang et al. [30] focus on a more general problem: non-convex problem, and propose a faster algorithm based on a proximal stochastic gradient method. Smith and Thakurta [31] explore the stability of model selection problems, and propose two differentially private algorithms based on perturbation stability and subsampling stability respectively.

Privacy-preserving distributed learning. Preserving privacy in distributed learning is challenging due to frequent information exchange in the iterative process. Recently, much works have been done to develop privacy-preserving distributed learning algorithms. Some of them employ cryptography-based methods in the protocol to hide the private information [32]–[35]. A recent work [34] uses partially homomorphic cryptography in ADMM-based distributed



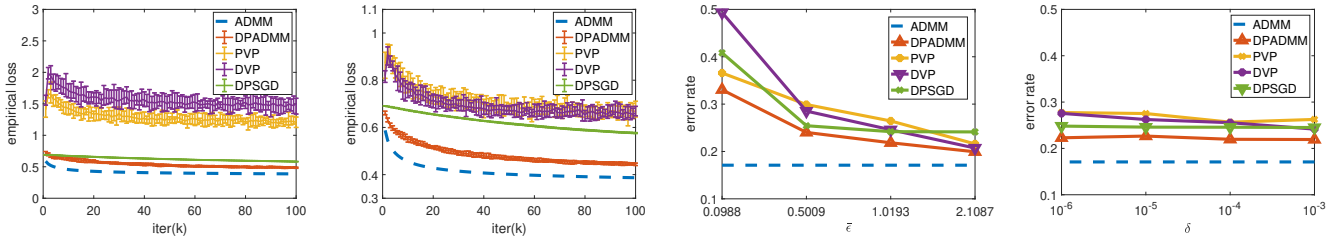
(a) $\epsilon = 0.05, \bar{\epsilon} = 0.5009, \delta = 10^{-3}$ (b) $\epsilon = 0.1, \bar{\epsilon} = 1.0193, \delta = 10^{-3}$ (c) $\epsilon = 0.05, \bar{\epsilon} = 0.5009, \delta = 10^{-3}$ (d) $\epsilon = 0.1, \bar{\epsilon} = 1.0193, \delta = 10^{-3}$

Fig. 4: Impact of distributed data source number on DP-ADMM (l_2-regularized logistic regression).



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

Fig. 5: Convergence properties of DP-ADMM (l_2-regularized logistic regression).



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

Fig. 6: Accuracy comparison in empirical loss and classification error rate (l_2-regularized logistic regression).

learning to preserve data privacy but the proposed approach cannot protect the information leakage of the private user data from the final learned models. In contrast, our approach provides differential privacy in the final trained machine learning models. Among the works on distributed learning with differential privacy, most of them focus on subgradient-based algorithms [36]–[39] and only a few works consider ADMM-based methods [1], [2], [17]–[19]. Zhang and Zhu [17] propose two perturbation methods: primal perturbation and dual perturbation to guarantee dynamic differential privacy in ADMM-based distributed learning. Zhang et al. [1] propose to perturb the penalty parameter of ADMM to guarantee differential privacy. Zhang et al. [2] propose recycled ADMM with differential privacy guarantee where the results from odd iterations could be re-utilized by the even iterations, and thus half of updates incur no privacy leakage. Guo and Gong [18] preserve differential privacy in the asynchronous ADMM algorithm. We design an ADMM-based distributed learning scheme with differential privacy which uses approximate augmented Lagrangian function for all iterations and adaptively changes the variance of added Gaussian noise in each iteration. We also use moments accountant method to analyze the total privacy loss to better estimate the trade-off between the data privacy and utility. We are the first to analyze rigorously the convergence rate and utility performance of ADMM with differential privacy.

Variants of ADMM. Some variants of ADMM have been proposed recently for applicability to more generous problems. Linearized ADMM [25], [26] replaces the quadratic function in the augmented Lagrangian function with a linearized approximation and thus provides a better way to solve subproblems without closed-form solutions. Stochastic ADMM [40], [41] considers stochastic and composite objective functions caused by natural uncertainties in observations. Our DP-ADMM algorithm inherits the features of linearized ADMM and stochastic ADMM, and guarantees strong differential privacy with good utility and low computation cost.

VIII. Conclusion

In this paper, we have 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 designing an approximate augmented Lagrangian function and Gaussian mechanism with time-varying variance, our

novel approach is noise-resilient, convergent and computation-efficient, especially under high privacy guarantee. We have also applied the moments accountant method to analyze the end-to-end privacy loss of the proposed iterative algorithm. The theoretical convergence guarantee and utility bound of our approach are derived. The evaluations on real-world datasets have demonstrated the effectiveness of our approach in the setting under high privacy guarantee.

REFERENCES

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

[2] ——, "Recycled admm: Improve privacy and accuracy with less computation in distributed algorithms," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2018, pp. 959–965.

[3] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "Distributed subgradient methods and quantization effects," in *2008 47th IEEE Conference on Decision and Control*. IEEE, 2008, pp. 4177–4184.

[4] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, p. 48, 2009.

[5] 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.

[6] Q. Ling and A. Ribeiro, "Decentralized linearized alternating direction method of multipliers," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5447–5451.

[7] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014.

[8] R. Zhang and J. Kwok, "Asynchronous distributed admm for consensus optimization," in *International Conference on Machine Learning*, 2014, pp. 1701–1709.

[9] P. Bianchi, W. Hachem, and F. Iutzeler, "A stochastic primal-dual algorithm for distributed asynchronous composite optimization," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 732–736.

[10] E. Wei and A. Ozdaglar, "Distributed alternating direction method of multipliers," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, 2012, pp. 5445–5450.

[11] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel, "D-admm: A communication-efficient distributed algorithm for separable optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2718–2723, 2013.

[12] 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.

[13] 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.

[14] 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.

[15] Z. Huang and Y. Gong, "Differential location privacy for crowdsourced spectrum sensing," in *2017 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2017, pp. 1–9.

[16] Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, "Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 591–606, 2018.

[17] 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.

[18] Y. Guo and Y. Gong, "Practical collaborative learning for crowdsensing in the internet of things with differential privacy," in *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2018, pp. 1–9.

[19] J. Ding, S. M. Errapotu, H. Zhang, Y. Gong, M. Pan, and Z. Han, "Stochastic admm based distributed machine learning with differential privacy," to appear in SecureComm. EAI, 2019.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE, 2014, pp. 464–473.

[30] D. Wang, M. Ye, and J. Xu, "Differentially private empirical risk minimization revisited: Faster and more general," in *Advances in Neural Information Processing Systems*, 2017, pp. 2722–2731.

[31] A. G. Thakurta and A. Smith, "Differentially private feature selection via stability arguments, and the robustness of the lasso," in *Conference on Learning Theory*, 2013, pp. 819–850.

[32] 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.

[33] 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.

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

[35] Y. Gong, Y. Fang, and Y. Guo, "Privacy-preserving collaborative learning for mobile health monitoring," in *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2015, pp. 1–6.

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

[37] 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.

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

[39] 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.

[40] 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.

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

[42] S. Shalev-Shwartz and Y. Singer, "Online learning: Theory, algorithms, and applications," 2007.

APPENDIX A
LEMMA 4 (l_2 SENSITIVITY OF PRIMAL VARIABLE UPDATE IN ALGORITHM 2)

Lemma 4. *Assume the objective function is smooth, $R(\cdot)$ is 1-strongly convex, and $\|\nabla\ell(\cdot)\| \leq c_1$. The l_2 sensitivity of primal variable update in Algorithm 2 is defined by:*

$$\max_{\mathcal{D}_i, \mathcal{D}_i^{'}} \|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\| = \frac{2c_1}{(\lambda/n+\rho)m_i}. \tag{40}$$

Proof. We define:

$$\begin{aligned} G(\boldsymbol{w}_i) &= \mathcal{L}_{\rho,i}(\boldsymbol{w}_i, \boldsymbol{w}^{k-1}, \boldsymbol{\gamma}_i^{k-1}), \\ g(\boldsymbol{w}_i) &= \frac{1}{m_i}\ell(\boldsymbol{a}_{i,m_i}^{'}, \boldsymbol{b}_{i,m_i}^{'}, \boldsymbol{w}_i) - \frac{1}{m_i}\ell(\boldsymbol{a}_{i,m_i}, \boldsymbol{b}_{i,m_i}, \boldsymbol{w}_i). \end{aligned}$$

According to the first step of ADMM, we have:

$$\boldsymbol{w}_{i,\mathcal{D}_i}^k = \operatorname*{argmin}_{\boldsymbol{w}_i} G(\boldsymbol{w}_i), \tag{42a}$$

$$\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k = \operatorname*{argmin}_{\boldsymbol{w}_i} G(\boldsymbol{w}_i) + g(\boldsymbol{w}_i). \tag{42b}$$

Also by assuming the smoothness of the objective function, the functions $G(\cdot)$ and $G(\cdot)+g(\cdot)$ are smooth, thus we have:

$$\nabla G(\boldsymbol{w}_{i,\mathcal{D}_i}^k) = \nabla G(\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k) + \nabla g(\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k) = 0. \tag{43}$$

Since we assume that the regularizer $R(\cdot)$ is 1-strongly convex, then function $G(\cdot)$ is $(\lambda/n+\rho)$-strongly convex. From the Lemma 14 of [42], we have:

$$\big(\nabla G(\boldsymbol{w}_{i,\mathcal{D}_i}^k) - \nabla G(\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k)\big)^\intercal(\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k) \geq (\lambda/n+\rho)\|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\|^2. \tag{44}$$

Combining this with the Cauchy-Schwartz inequality, we can get:

$$\begin{aligned} \|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\| \cdot \|\nabla g(\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k)\| \geq& \big(\nabla G(\boldsymbol{w}_{i,\mathcal{D}_i}^k) - \nabla G(\boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k)\big)^\intercal(\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k) \\ \geq& (\lambda/n+\rho)\|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\|^2. \end{aligned} \tag{45}$$

By dividing both sides of the above inequality by $(\lambda/n+\rho)\|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\|$, we can get:

$$\|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\| \leq \frac{\|\nabla\ell(\boldsymbol{a}_{i,m_i}, \boldsymbol{b}_{i,m_i}, \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k) - \nabla\ell(\boldsymbol{a}_{i,m_i}^{'}, \boldsymbol{b}_{i,m_i}^{'}, \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k)\|}{m_i(\lambda/n+\rho)}. \tag{46}$$

As we assume that $\|\nabla\ell(\cdot)\| \leq c_1$, then we obtain the result:

$$\max \|\boldsymbol{w}_{i,\mathcal{D}_i}^k - \boldsymbol{w}_{i,\mathcal{D}_i^{'}}^k\| = \frac{2c_1}{(\lambda/n+\rho)m_i}. \tag{47}$$

□

APPENDIX B
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 agent i for k-th iteration could be defined by the log moment generating function at τ:

$$\alpha_i^k(\tau) = \ln\left(\mathbb{E}_{\tilde{\boldsymbol{w}}_i^k}\left[\left(\frac{\Pr\left[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i\right]}{\Pr\left[\tilde{\boldsymbol{w}}_i^k|\mathcal{D}_i^{'}\right]}\right)^\tau\right]\right). \tag{48}$$

In the 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 $\mathcal{N}(2c_1/\big(m_i(\rho+1/\eta_i^k)\big),\sigma_{i,k}^2)$. We obtain that $\alpha_i^k(\tau)$ by $\alpha_i^k(\tau) = \ln\big(\max(E_1,E_2)\big)$, 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,

$$\mathbb{E}_{z\sim\mu_0}\left[(\frac{\mu_0(z)}{\mu_1(z)})^\tau\right]=\exp\left(\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}\right), \tag{49a}$$

$$\mathbb{E}_{z\sim\mu_1}\left[(\frac{\mu_1(z)}{\mu_0(z)})^\tau\right]=\exp\left(\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}\right), \tag{49b}$$

we have:

$$\alpha_i^k(\tau)=\frac{\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}. \tag{50}$$

According to Theorem 2 (linear composability) in [20], we have the τ^{th} log moment of the overall privacy loss from i:

$$\alpha_i(\tau)=\sum_{k=1}^{t}\alpha_i^k(\tau)=\frac{t\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}. \tag{51}$$

We aim to prove that our proposed algorithm DP-ADMM (Algorithm 3) achieves $(\bar{\epsilon},\delta)$-differential privacy. According to Theorem 2 (tail bound) in [20], we have:

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

Since $\delta\in(0,1)$, there exists a positive integer τ to make $t\tau(\tau+1)\epsilon^2/\big(4\ln(1.25/\delta)\big)-\tau\bar{\epsilon}<0$. Furthermore, $t\tau(\tau+1)\epsilon^2/\big(4\ln(1.25/\delta)\big)-\tau\bar{\epsilon}$ is a quadratic function w.r.t. τ. Thus, if there is a solution to the above minimization problem, we must have: when $\tau=1$,

$$\frac{t\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}-\tau\bar{\epsilon}=\frac{t\epsilon^2}{2\ln(1.25/\delta)}-\bar{\epsilon}<0. \tag{52}$$

Therefore, we obtain:

$$\frac{t\epsilon^2}{2\ln(1.25/\delta)}<\bar{\epsilon}. \tag{53}$$

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

$$\ln(\delta)=\min_{\tau\in\mathbb{Z}^+}\left(\frac{t\tau(\tau+1)\epsilon^2}{4\ln(1.25/\delta)}-\tau\bar{\epsilon}\right)\geq-\frac{t\epsilon^2}{16\ln(1.25/\delta)}+\frac{\bar{\epsilon}}{2}-\frac{\bar{\epsilon}^2\ln(1.25/\delta)}{t\epsilon^2} \tag{54}$$

From (53) and (54), we obtain:

$$\ln(1/\delta)\leq-\frac{3\bar{\epsilon}}{8}+\frac{\bar{\epsilon}^2\ln(1.25/\delta)}{t\epsilon^2}\leq\frac{\bar{\epsilon}^2\ln(1.25/\delta)}{t\epsilon^2}, \tag{55}$$

which leads to the following inequality:

$$\bar{\epsilon}\geq\sqrt{\frac{t\ln(1/\delta)}{\ln(1.25/\delta)}}\epsilon. \tag{56}$$

Therefore, there exists a constant c_0, the overall privacy loss $\bar{\epsilon}$ satisfies:

$$\bar{\epsilon}=c_0\sqrt{t}\epsilon. \tag{57}$$

□

APPENDIX C
LEMMA 5 USED IN THE PROOF OF LEMMA 2

Lemma 5. *Assume $L(\cdot)$ is a convex differentiable function. $s\geq0$ is a scalar. For any vector $\boldsymbol{x}\in\mathbb{R}^d$ and $\boldsymbol{y}\in\mathbb{R}^d$, we denote their Bregman divergence as $D(\boldsymbol{x},\boldsymbol{y})=h(\boldsymbol{x})-h(\boldsymbol{y})-\langle\nabla h(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\rangle$, where $h(\cdot)$ is a continuously-differentiable real-valued and strictly convex function. If we define:*

$$\boldsymbol{x}^*=\operatorname*{argmin}_{\boldsymbol{x}}L(\boldsymbol{x})+sD(\boldsymbol{x},\boldsymbol{y}), \tag{58}$$

then

$$\big\langle\nabla L(\boldsymbol{x}^*),\boldsymbol{x}^*-\boldsymbol{x}\big\rangle\leq s\big(D(\boldsymbol{x},\boldsymbol{y})-D(\boldsymbol{x},\boldsymbol{x}^*)-D(\boldsymbol{x}^*,\boldsymbol{y})\big). \tag{59}$$

Proof. According to the optimality condition,

$$\big\langle\nabla L(\boldsymbol{x}^*)+s\nabla D(\boldsymbol{x}^*,\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{x}^*\big\rangle\geq0. \tag{60}$$

Then,

$$\begin{aligned}\big\langle \nabla L(\boldsymbol{x}^*), \boldsymbol{x}^* - \boldsymbol{x}\big\rangle \leq & s\big\langle \nabla D(\boldsymbol{x}^*, \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}^*\big\rangle \\ = & s\big\langle \nabla h(\boldsymbol{x}^*) - \nabla h(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}^*\big\rangle \\ = & s\big(D(\boldsymbol{x}, \boldsymbol{y}) - D(\boldsymbol{x}, \boldsymbol{x}^*) - D(\boldsymbol{x}^*, \boldsymbol{y})\big).\end{aligned} \tag{61}$$

□

APPENDIX D
PROOF OF LEMMA 2

Proof. Since we assume that $\ell(\cdot)$ and $R(\cdot)$ are convex, the function $f_i(\cdot)$ is convex. Due to the convexity of $f_i(\cdot)$, we have:

$$f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) \leq \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}), \tilde{\boldsymbol{w}}_i^{k-1} - \boldsymbol{w}_i\big\rangle, \tag{62}$$

which can lead to:

$$\begin{aligned}f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + \big\langle \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i, -\boldsymbol{\gamma}_i^k\big\rangle \leq & \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}), \tilde{\boldsymbol{w}}_i^{k-1} - \boldsymbol{w}_i\big\rangle + \big\langle \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i, -\boldsymbol{\gamma}_i^k\big\rangle \\ = & \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k-1} - \tilde{\boldsymbol{w}}_i^{k}\big\rangle - \big(\rho + 1/\eta_i^k\big)\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle \\ & + \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - \boldsymbol{\gamma}_i^k - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i\big\rangle.\end{aligned} \tag{63}$$

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

$$\begin{aligned}\big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - \boldsymbol{\gamma}_i^k - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i\big\rangle = & \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - \boldsymbol{\gamma}_i^{k-1} + \rho(\tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i\big\rangle \\ & + \big\langle \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i, \rho(\boldsymbol{w}^{k-1} - \boldsymbol{w}^{k})\big\rangle.\end{aligned} \tag{64}$$

By combining (63) and (64), we obtain:

$$\begin{aligned}f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + \big\langle \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i, -\boldsymbol{\gamma}_i^k\big\rangle \leq & \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k-1} - \tilde{\boldsymbol{w}}_i^{k}\big\rangle \\ & + \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - \boldsymbol{\gamma}_i^{k-1} + \rho(\tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i\big\rangle \\ & + \big\langle \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i, \rho(\boldsymbol{w}^{k-1} - \boldsymbol{w}^{k})\big\rangle - \big(\rho + 1/\eta_i^k\big)\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle.\end{aligned} \tag{65}$$

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

$$\begin{aligned}\big\langle \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i, \rho(\boldsymbol{w}^{k-1} - \boldsymbol{w}^{k})\big\rangle = & \frac{\rho}{2}\big(\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2 - \|\boldsymbol{w}_i - \boldsymbol{w}^{k}\|^2\big) + \frac{\rho}{2}\big(\|\tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}^{k}\|^2 - \|\tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}^{k-1}\|^2\big) \\ \leq & \frac{\rho}{2}\big(\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2 - \|\boldsymbol{w}_i - \boldsymbol{w}^{k}\|^2\big) + \frac{\rho}{2}\|\tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}^{k}\|^2 \\ = & \frac{\rho}{2}\big(\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2 - \|\boldsymbol{w}_i - \boldsymbol{w}^{k}\|^2\big) + \frac{1}{2\rho}\|\boldsymbol{\gamma}_i^{k} - \boldsymbol{\gamma}_i^{k-1}\|^2.\end{aligned} \tag{66}$$

According to the Line 4 and 6 of Algorithm 3, $\tilde{\boldsymbol{w}}_i^k$ is equal to the solution to $\min_{\boldsymbol{w}_i} \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}), \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle - \big\langle \boldsymbol{\gamma}_i^{k-1}, \boldsymbol{w}_i - \boldsymbol{w}^{k-1}\big\rangle + \rho\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2/2 + \|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2/(2\eta_i^k) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\boldsymbol{w}_i$. By applying Lemma 5 where $D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$, $s = 1/\eta_i^k$, and $L(\boldsymbol{x}) = \big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}), \boldsymbol{x} - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle - \big\langle \boldsymbol{\gamma}_i^{k-1}, \boldsymbol{x} - \boldsymbol{w}^{k-1}\big\rangle + \rho\|\boldsymbol{x} - \boldsymbol{w}^{k-1}\|^2/2 - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\boldsymbol{w}_i$, we have:

$$\big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - \boldsymbol{\gamma}_i^{k-1} + \rho(\tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i\big\rangle \leq \frac{1}{2\eta_i^k}\big(\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2 - \|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k}\|^2 - \|\tilde{\boldsymbol{w}}_i^{k} - \tilde{\boldsymbol{w}}_i^{k-1}\|^2\big). \tag{67}$$

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

$$\big\langle f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k, \tilde{\boldsymbol{w}}_i^{k-1} - \tilde{\boldsymbol{w}}_i^{k}\big\rangle \leq \frac{\eta_i^k}{2}\big\| f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\big\|^2 + \frac{1}{2\eta_i^k}\|\tilde{\boldsymbol{w}}_i^{k} - \tilde{\boldsymbol{w}}_i^{k-1}\|^2. \tag{68}$$

Combining (65),(66),(67), and (68), we have:

$$\begin{aligned}f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + \big\langle \tilde{\boldsymbol{w}}_i^{k} - \boldsymbol{w}_i, -\boldsymbol{\gamma}_i^k\big\rangle \leq & \frac{\eta_i^k}{2}\|f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\|^2 - \big(\rho + 1/\eta_i^k\big)\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle \\ & + \frac{1}{2\eta_i^k}\big(\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2 - \|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k}\|^2\big) \\ & + \frac{\rho}{2}(\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2 - \|\boldsymbol{w}_i - \boldsymbol{w}^{k}\|^2) + \frac{1}{2\rho}\|\boldsymbol{\gamma}_i^{k} - \boldsymbol{\gamma}_i^{k-1}\|^2.\end{aligned} \tag{69}$$

Next, according to our algorithm where $\boldsymbol{\gamma}_i^k = \boldsymbol{\gamma}_i^{k-1} - \rho(\tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k)$ and $\boldsymbol{w}^k = \frac{1}{n}\sum_{i=1}^n \tilde{\boldsymbol{w}}_i^k - \frac{1}{n}\sum_{i=1}^n \boldsymbol{\gamma}_i^{k-1}/\rho$, we have:

$$\sum_{i=1}^{n} \big\langle \boldsymbol{w}^k - \boldsymbol{w}, \boldsymbol{\gamma}_i^k \big\rangle = 0. \tag{70}$$

And also, we could obtain:

$$\big\langle \boldsymbol{\gamma}_i^k - \boldsymbol{\gamma}_i, \tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k \big\rangle = \frac{1}{\rho}\big\langle \boldsymbol{\gamma}_i^k - \boldsymbol{\gamma}_i, \boldsymbol{\gamma}_i^{k-1} - \boldsymbol{\gamma}_i^k \big\rangle = \frac{1}{2\rho}\big(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^{k-1}\|^2 - \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^k\|^2 - \|\boldsymbol{\gamma}_i^k - \boldsymbol{\gamma}_i^{k-1}\|^2\big). \tag{71}$$

Thus, combining (69), (70) and (71), we obtain the result in the Lemma 2:

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + (\boldsymbol{u}_i^k - \boldsymbol{u}_i)^{\intercal} F(\boldsymbol{u}_i^k)\Big)\\
=&\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + \big\langle -\boldsymbol{\gamma}_i^k, \tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}_i\big\rangle + \big\langle \boldsymbol{\gamma}_i^k, \boldsymbol{w}^k - \boldsymbol{w}\big\rangle + \big\langle \boldsymbol{\gamma}_i^k - \boldsymbol{\gamma}_i, \tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k\big\rangle\Big)\\
\le&\sum_{i=1}^{n}\Big(\frac{\eta_i^k}{2}\|f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\|^2 - \big(\rho + 1/\eta_i^k\big)\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle + \frac{\rho}{2}\big(\|\boldsymbol{w}_i - \boldsymbol{w}^{k-1}\|^2 - \|\boldsymbol{w}_i - \boldsymbol{w}^k\|^2\big)\\
&\quad + \frac{1}{2\eta_i^k}\big(\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\|^2 - \|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^k\|^2\big) + \frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^{k-1}\|^2 - \frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^k\|^2\Big).
\end{aligned} \tag{72}$$

□

APPENDIX E
PROOF OF THEOREM 3

Proof. According to the convexity of $f_i(\cdot)$ and the monotonicity of the operator $F(\cdot)$, and applying Lemma 2, we have:

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i) + (\bar{\boldsymbol{u}}_i^t - \boldsymbol{u}_i)^{\intercal} F(\bar{\boldsymbol{u}}_i^t)\Big)\\
=&\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i) + \big\langle -\bar{\boldsymbol{\gamma}}_i^t, \bar{\boldsymbol{w}}_i^t - \boldsymbol{w}_i\big\rangle + \big\langle \bar{\boldsymbol{\gamma}}_i^t, \bar{\boldsymbol{w}}^t - \boldsymbol{w}\big\rangle + \big\langle \bar{\boldsymbol{\gamma}}_i^t - \boldsymbol{\gamma}_i, \bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\big\rangle\Big)\\
\le&\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + (\boldsymbol{u}_i^k - \boldsymbol{u}_i)^{\intercal} F(\boldsymbol{u}_i^k)\Big)\\
=&\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^{k-1}) - f_i(\boldsymbol{w}_i) + \big\langle -\boldsymbol{\gamma}_i^k, \tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}_i\big\rangle + \big\langle \boldsymbol{\gamma}_i^k, \boldsymbol{w}^k - \boldsymbol{w}\big\rangle + \big\langle \boldsymbol{\gamma}_i^k - \boldsymbol{\gamma}_i, \tilde{\boldsymbol{w}}_i^k - \boldsymbol{w}^k\big\rangle\Big)\\
\le&\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\Big(\frac{\eta_i^k}{2}\big\|f_i^{'}(\tilde{\boldsymbol{w}}_i^k) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\big\|^2 - \big(\rho + 1/\eta_i^k\big)\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\Big)\\
&+ \frac{1}{t}\sum_{i=1}^{n}\Big(\frac{1}{2\eta^t}\big\|\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i^0\big\|^2 + \frac{\rho}{2}\|\boldsymbol{w}_i - \boldsymbol{w}^0\|^2 + \frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^0\|^2\Big).
\end{aligned} \tag{73}$$

Let $(\boldsymbol{w}_i, \boldsymbol{w})$ be the optimal solution $(\boldsymbol{w}_i^*, \boldsymbol{w}^*)$ in the above inequality. We get:

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i^*) + \big\langle -\bar{\boldsymbol{\gamma}}_i^t, \bar{\boldsymbol{w}}_i^t - \boldsymbol{w}_i^*\big\rangle + \big\langle \bar{\boldsymbol{\gamma}}_i^t, \bar{\boldsymbol{w}}^t - \boldsymbol{w}^*\big\rangle + \big\langle \bar{\boldsymbol{\gamma}}_i^t - \boldsymbol{\gamma}_i, \bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\big\rangle\Big)\\
\le&\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\frac{\eta_i^k}{2}\big\|f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k\big\|^2 - \sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\big(\rho + 1/\eta_i^k\big)\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i^* - \tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\\
&+ \frac{1}{t}\sum_{i=1}^{n}\frac{c_w^2}{2\eta_i^t} + \frac{n}{t}\frac{\rho}{2}c_w^2 + \frac{1}{t}\sum_{i=1}^{n}\frac{1}{2\rho}\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^0\|^2.
\end{aligned} \tag{74}$$

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 of the above inequality:

$$\begin{aligned}
&\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}} \sum_{i=1}^{n} \Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i^*) + \big\langle -\bar{\boldsymbol{\gamma}}_i^t, \bar{\boldsymbol{w}}_i^t - \boldsymbol{w}_i^* \big\rangle + \big\langle \bar{\boldsymbol{\gamma}}_i^t, \bar{\boldsymbol{w}}^t - \boldsymbol{w}^* \big\rangle + \big\langle \bar{\boldsymbol{\gamma}}_i^t - \boldsymbol{\gamma}_i, \bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t \big\rangle \Big) \\
= &\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}} \sum_{i=1}^{n} \Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i) - \boldsymbol{\gamma}_i(\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t) \Big) \\
= &\sum_{i=1}^{n} \Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i) - \max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}} \boldsymbol{\gamma}_i(\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t) \Big) \\
= &\sum_{i=1}^{n} \Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i) + \beta(\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\|) \Big).
\end{aligned} \tag{75}$$

And we also get the maximum of the right side:

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \frac{\eta_i^k}{2} \big\| f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k \big\|^2 - \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \big(\rho + 1/\eta_i^k\big) \big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i^* - \tilde{\boldsymbol{w}}_i^{k-1} \big\rangle \\
&+ \frac{1}{t} \sum_{i=1}^{n} \frac{c_w^2}{2\eta_i^t} + \frac{\rho n}{2t} c_w^2 + \max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}} \frac{1}{t} \sum_{i=1}^{n} \frac{1}{2\rho} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^0\|^2 \\
= &\sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \frac{\eta_i^k}{2} \big\| f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k \big\|^2 - \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \big(\rho + 1/\eta_i^k\big) \big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i^* - \tilde{\boldsymbol{w}}_i^{k-1} \big\rangle \\
&+ \frac{1}{t} \sum_{i=1}^{n} \frac{c_w^2}{2\eta_i^t} + \frac{\rho n}{2t} c_w^2 + \frac{n}{t} \frac{\beta^2}{2\rho}.
\end{aligned} \tag{76}$$

Thus, we obtain the inequality:

$$\begin{aligned}
&\sum_{i=1}^{n} \Big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i) + \beta\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\| \Big) \\
\leq &\sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \frac{\eta_i^k}{2} \big\| f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k \big\|^2 - \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \big(\rho + 1/\eta_i^k\big) \big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i^* - \tilde{\boldsymbol{w}}_i^{k-1} \big\rangle \\
&+ \frac{1}{t} \sum_{i=1}^{n} \frac{c_w^2}{2\eta_i^t} + \frac{\rho n}{2t} c_w^2 + \frac{n}{t} \frac{\beta^2}{2\rho}.
\end{aligned} \tag{77}$$

Since we assume $\|\ell^{'}(\cdot)\| \leq c_1$ and $\|R^{'}(\cdot)\| \leq c_2$, we have $\mathbb{E}\big[\big\| f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k \big\|^2\big] = (c_1 + \lambda c_2/n)^2 + 8dpc_1^2 \ln(1.25/\delta)/\big(m_i^2\epsilon^2\big)$. With $\mathbb{E}\big[\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i^* - \tilde{\boldsymbol{w}}_i^{k-1} \big\rangle\big] = 0$ and $\eta_i^k = c_w\big(2k(c_1 + \lambda c_2/n)^2 + 16kdpc_1^2 \ln(1.25/\delta)/\big(m_i^2\epsilon^2\big)\big)^{-\frac{1}{2}}$, by taking expectation of the inequality (77), we obtain:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i=1}^{n} \big(f_i(\bar{\boldsymbol{w}}_i^t) - f_i(\boldsymbol{w}_i^*) + \beta\|\bar{\boldsymbol{w}}_i^t - \bar{\boldsymbol{w}}^t\| \big) \Big] \\
\leq &\sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \mathbb{E}\Big[\frac{\eta_i^k}{2} \| f_i^{'}(\tilde{\boldsymbol{w}}_i^{k-1}) - (\rho + 1/\eta_i^k)\boldsymbol{\xi}_i^k \|^2 \Big] - \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} \big(\rho + 1/\eta_i^k\big) \mathbb{E}\Big[\big\langle \boldsymbol{\xi}_i^k, \boldsymbol{w}_i^* - \tilde{\boldsymbol{w}}_i^{k-1} \big\rangle \Big] \\
&+ \frac{1}{t} \sum_{i=1}^{n} \frac{c_w^2}{2\eta_i^t} + \frac{\rho n}{2t} c_w^2 + \frac{n}{t} \frac{\beta^2}{2\rho},
\end{aligned} \tag{78}$$

which leads to the result in the theorem:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{i=1}^{n}\big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i^*)+\beta\|\bar{\boldsymbol{w}}_i^t-\bar{\boldsymbol{w}}^t\|\big)\Big]\\
=&\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\frac{c_w}{2\sqrt{2k}}\sqrt{(c_1+\lambda c_2/n)^2+\frac{8dpc_1^2\ln(1.25/\delta)}{m_i^2\epsilon^2}}\\
&+\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\frac{c_w\sqrt{2t}}{2}\sqrt{(c_1+\lambda c_2/n)^2+\frac{8dpc_1^2\ln(1.25/\delta)}{m_i^2\epsilon^2}}+\frac{n\rho}{2t}c_w^2+\frac{n\beta^2}{2\rho t}\\
=&\sum_{i=1}^{n}\frac{c_w}{2\sqrt{2t}}\sqrt{(c_1+\lambda c_2/n)^2+\frac{8dpc_1^2\ln(1.25/\delta)}{m_i^2\epsilon^2}}\Big(\sum_{k=1}^{t}\frac{1}{\sqrt{k}}+2\sqrt{t}\Big)+\frac{n\rho}{2t}c_w^2+\frac{n\beta^2}{2\rho t}\\
\leq&\sum_{i=1}^{n}\frac{\sqrt{2}c_w}{\sqrt{t}}\sqrt{(c_1+\lambda c_2/n)^2+\frac{8dpc_1^2\ln(1.25/\delta)}{m_i^2\epsilon^2}}+\frac{n(\rho c_w^2+\beta^2/\rho)}{2t}.
\end{aligned}\tag{79}$$

□

APPENDIX F
PROOF OF LEMMA 3

Proof. As we assume that $\ell(\cdot)$ and $R(\cdot)$ are smooth and convex, $\|\nabla^2\ell(\cdot)\|\leq c_3$, and $\|\nabla^2 R(\cdot)\|\leq c_4$, thus we have $\|\nabla^2 f_i(\cdot)\| = \|\nabla^2\ell(\cdot)+\lambda/n\nabla^2 R(\cdot)\|\leq c_3+\lambda c_4/n$ is bounded. This leads to:

$$\|\nabla f_i(\boldsymbol{x})-\nabla f_i(\boldsymbol{y})\|\leq(c_3+\lambda c_4/n)\|\boldsymbol{x}-\boldsymbol{y}\|.\tag{80}$$

Thus, $f_i(\cdot)$ is $(c_3+\lambda c_4/n)$-Lipschitz smooth. According to the property of Lipschitz smooth, we have:

$$\begin{aligned}
f_i(\tilde{\boldsymbol{w}}_i^k)\leq&f_i(\tilde{\boldsymbol{w}}_i^{k-1})+\big\langle\nabla f_i(\tilde{\boldsymbol{w}}_i^{k-1}),\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle+\frac{c_3+\lambda c_4/n}{2}\|\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\|^2\\
=&f_i(\tilde{\boldsymbol{w}}_i^{k-1})+\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle+\frac{c_3+\lambda c_4/n}{2}\|\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\|^2\\
&+\big\langle\nabla f_i(\tilde{\boldsymbol{w}}_i^{k-1})-(\rho+1/\eta_i^k)\boldsymbol{\xi}_i^k,\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle.
\end{aligned}\tag{81}$$

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

$$f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)\leq\big\langle\nabla f_i(\tilde{\boldsymbol{w}}_i^k),\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i\big\rangle.\tag{82}$$

According to (81) and (82), we have:

$$\begin{aligned}
f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)+\big\langle\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i,-\boldsymbol{\gamma}_i^k\big\rangle\leq&f_i(\tilde{\boldsymbol{w}}_i^{k-1})-f_i(\boldsymbol{w}_i)+\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\\
&+\big\langle\nabla f_i(\tilde{\boldsymbol{w}}_i^{k-1})-(\rho+1/\eta_i^k)\boldsymbol{\xi}_i^k,\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\\
&+\frac{c_3+\lambda c_4/n}{2}\|\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\|^2+\big\langle\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i,-\boldsymbol{\gamma}_i^k\big\rangle,
\end{aligned}\tag{83}$$

which leads to:

$$\begin{aligned}
f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)+\big\langle\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i,-\boldsymbol{\gamma}_i^k\big\rangle\leq&\big\langle\nabla f_i(\tilde{\boldsymbol{w}}_i^{k-1})-\boldsymbol{\gamma}_i^k-(\rho+1/\eta_i^k)\boldsymbol{\xi}_i^k+\rho(\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}^{k-1}),\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i\big\rangle\\
&+\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle+\frac{c_3+\lambda c_4/n}{2}\|\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\|^2\\
&+\big\langle\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i,\rho(\boldsymbol{w}^{k-1}-\boldsymbol{w}^k)\big\rangle-\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle.
\end{aligned}\tag{84}$$

Based on Young's inequality,

$$\big\langle(\rho+1/\eta_i^k)\boldsymbol{\xi}_i^k,\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\leq\frac{1}{2(1/\eta_i^k-(c_3+\lambda c_4/n))}\big\|(\rho+1/\eta_i^k)\boldsymbol{\xi}_i^k\big\|^2+\frac{1/\eta_i^k-(c_3+\lambda c_4/n)}{2}\big\|\tilde{\boldsymbol{w}}_i^k-\tilde{\boldsymbol{w}}_i^{k-1}\big\|^2.\tag{85}$$

Combining (66), (67), (84) and (85), we have:

$$\begin{aligned}
f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)+\big\langle\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i,-\boldsymbol{\gamma}_i^k\big\rangle\leq&\frac{(\rho+1/\eta_i^k)^2}{2(1/\eta_i^k-(c_3+\lambda c_4/n))}\big\|\boldsymbol{\xi}_i^k\big\|^2-\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\\
&+\frac{1}{2\eta_i^k}(\|\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^{k-1}\|^2-\|\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^k\|^2)\\
&+\frac{\rho}{2}(\|\boldsymbol{w}_i-\boldsymbol{w}^{k-1}\|^2-\|\boldsymbol{w}_i-\boldsymbol{w}^k\|^2)+\frac{1}{2\rho}\|\boldsymbol{\gamma}_i^k-\boldsymbol{\gamma}_i^{k-1}\|^2.
\end{aligned}\tag{86}$$

Combining (86), (70) and (71), we get the result as desired:

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)+(\boldsymbol{u}_i^k-\boldsymbol{u}_i)^\intercal F(\boldsymbol{u}_i^k)\Big)\\
=&\frac{1}{n}\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)+\big\langle-\boldsymbol{\gamma}_i^k,\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i\big\rangle+\big\langle\boldsymbol{\gamma}_i^k,\boldsymbol{w}^k-\boldsymbol{w}\big\rangle+\big\langle\boldsymbol{\gamma}_i^k-\boldsymbol{\gamma}_i,\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}^k\big\rangle\Big)\\
\le&\sum_{i=1}^{n}\Big(\frac{\big(\rho+1/\eta_i^k\big)^2}{2(1/\eta_i^k-(c_3+\lambda c_4/n))}\big\|\boldsymbol{\xi}_i^k\big\|^2-\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle+\frac{1}{2\eta_i^k}(\|\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^{k-1}\|^2-\|\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^k\|^2)\\
&+\frac{\rho}{2}(\|\boldsymbol{w}_i-\boldsymbol{w}^{k-1}\|^2-\|\boldsymbol{w}_i-\boldsymbol{w}^k\|^2)+\frac{1}{2\rho}(\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^{k-1}\|^2-\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^k\|^2)\Big).
\end{aligned}\tag{87}$$

□

APPENDIX G
PROOF OF THEOREM 4

Proof. According to the convexity of $f_i(\cdot)$ and the monotonicity of $F(\cdot)$, and applying Lemma 3, we have:

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i)+(\bar{\boldsymbol{u}}_i^t-\boldsymbol{u}_i)^\intercal F(\bar{\boldsymbol{u}}_i^t)\Big)\\
=&\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i)+\big\langle-\bar{\boldsymbol{\gamma}}_i^t,\bar{\boldsymbol{w}}_i^t-\boldsymbol{w}_i\big\rangle+\big\langle\bar{\boldsymbol{\gamma}}_i^t,\bar{\boldsymbol{w}}^t-\boldsymbol{w}\big\rangle+\big\langle\bar{\boldsymbol{\gamma}}_i^t-\boldsymbol{\gamma}_i,\bar{\boldsymbol{w}}_i^t-\bar{\boldsymbol{w}}^t\big\rangle\Big)\\
\le&\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)+(\boldsymbol{u}_i^k-\boldsymbol{u}_i)^\intercal F(\boldsymbol{u}_i^k)\Big)\\
=&\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{n}\Big(f_i(\tilde{\boldsymbol{w}}_i^k)-f_i(\boldsymbol{w}_i)+\big\langle-\boldsymbol{\gamma}_i^k,\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}_i\big\rangle+\big\langle\boldsymbol{\gamma}_i^k,\boldsymbol{w}^k-\boldsymbol{w}\big\rangle+\big\langle\boldsymbol{\gamma}_i^k-\boldsymbol{\gamma}_i,\tilde{\boldsymbol{w}}_i^k-\boldsymbol{w}^k\big\rangle\Big)\\
\le&\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\Big(\frac{\big(\rho+1/\eta_i^k\big)^2}{2(1/\eta_i^k-(c_3+\lambda c_4/n))}\big\|\boldsymbol{\xi}_i^k\big\|^2-\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\Big)\\
&+\frac{1}{t}\sum_{i=1}^{n}\Big(\frac{1}{2\eta_i^t}\|\boldsymbol{w}_i-\tilde{\boldsymbol{w}}_i^0\|^2+\frac{\rho}{2}\|\boldsymbol{w}_i-\boldsymbol{w}^0\|^2+\frac{1}{2\rho}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^0\|^2\Big).
\end{aligned}\tag{88}$$

By letting $(\boldsymbol{w}_i,\boldsymbol{w})$ be the optimal solution $(\boldsymbol{w}_i^*,\boldsymbol{w}^*)$, we have:

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i^*)+\big\langle-\bar{\boldsymbol{\gamma}}_i^t,\bar{\boldsymbol{w}}_i^t-\boldsymbol{w}_i^*\big\rangle+\big\langle\bar{\boldsymbol{\gamma}}_i^t,\bar{\boldsymbol{w}}^t-\boldsymbol{w}^*\big\rangle+\big\langle\bar{\boldsymbol{\gamma}}_i^t-\boldsymbol{\gamma}_i,\bar{\boldsymbol{w}}_i^t-\bar{\boldsymbol{w}}^t\big\rangle\Big)\\
=&\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\Big(\frac{\big(\rho+1/\eta_i^k\big)^2}{2(1/\eta_i^k-(c_3+\lambda c_4/n))}\big\|\boldsymbol{\xi}_i^k\big\|^2-\big(\rho+1/\eta_i^k\big)\big\langle\boldsymbol{\xi}_i^k,\boldsymbol{w}_i^*-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\Big)\\
&+\frac{1}{t}\sum_{i=1}^{n}\frac{c_w^2}{2\eta_i^t}+\frac{\rho n}{2t}c_w^2+\frac{1}{t}\sum_{i=1}^{n}\frac{1}{2\rho}\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^0\|^2.
\end{aligned}\tag{89}$$

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

$$\begin{aligned}
&\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\le\beta\}}\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i^*)+\big\langle-\bar{\boldsymbol{\gamma}}_i^t,\bar{\boldsymbol{w}}_i^t-\boldsymbol{w}_i^*\big\rangle+\big\langle\bar{\boldsymbol{\gamma}}_i^t,\bar{\boldsymbol{w}}^t-\boldsymbol{w}^*\big\rangle+\big\langle\bar{\boldsymbol{\gamma}}_i^t-\boldsymbol{\gamma}_i,\bar{\boldsymbol{w}}_i^t-\bar{\boldsymbol{w}}^t\big\rangle\Big)\\
=&\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\le\beta\}}\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i)-\boldsymbol{\gamma}_i(\bar{\boldsymbol{w}}_i^t-\bar{\boldsymbol{w}}^t)\Big)\\
=&\sum_{i=1}^{n}\Big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i)+\beta\|\bar{\boldsymbol{w}}_i^t-\bar{\boldsymbol{w}}^t\|\Big).
\end{aligned}\tag{90}$$

Since we have $\mathbb{E}\big[\big\langle\boldsymbol{\xi}_i^k,\boldsymbol{w}_i^*-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\big]=0$ and $\mathbb{E}\big[\big\|\boldsymbol{\xi}_i^k\big\|^2\big]=dp\sigma_{i,k}^2=8dp\ln(1.25/\delta)c_1^2/\big(m_i^2\epsilon^2(\rho+1/\eta_i^k)^2\big)$ due to the variance definition, we take the expectation of the (90) and let $\eta_i^k=\big(c_3+\lambda c_4/n+2c_1\sqrt{4dpk\ln(1.25/\delta)}/\big(\epsilon m_i c_w\big)\big)^{-1}$, which leads

to the result:

$$
\begin{aligned}
&\mathbb{E}\bigg[\sum_{i=1}^{n}\big(f_i(\bar{\boldsymbol{w}}_i^t)-f_i(\boldsymbol{w}_i^*)+\beta\big\|\bar{\boldsymbol{w}}_i^t-\bar{\boldsymbol{w}}^t\big\|\big)\bigg]\\
\leq&\mathbb{E}\bigg[\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\frac{\big(\rho+1/\eta_i^k\big)^2}{2(1/\eta_i^k-(c_3+\lambda c_4/n))}\big\|\boldsymbol{\xi}_i^k\big\|^2\bigg]-\sum_{i=1}^{n}\frac{1}{t}\sum_{k=1}^{t}\big(\rho+1/\eta_i^k\big)\mathbb{E}\Big[\big\langle\boldsymbol{\xi}_i^k,\boldsymbol{w}_i^*-\tilde{\boldsymbol{w}}_i^{k-1}\big\rangle\Big]\\
&+\frac{1}{t}\sum_{i=1}^{n}\frac{c_w^2}{2\eta_i^t}+\frac{\rho n}{2t}c_w^2+\max_{\{\boldsymbol{\gamma}_i:\|\boldsymbol{\gamma}_i\|\leq\beta\}}\frac{1}{t}\sum_{i=1}^{n}\frac{1}{2\rho}\big\|\boldsymbol{\gamma}_i-\boldsymbol{\gamma}_i^0\big\|^2\\
=&\sum_{i=1}^{n}\frac{c_w c_1\sqrt{dp\ln(1.25/\delta)}}{m_i\epsilon t}\bigg(\sum_{k=1}^{t}\frac{1}{\sqrt{k}}+2\sqrt{t}\bigg)+\frac{nc_w^2(c_3+\lambda c_4/n)}{2t}+\frac{\rho n}{2t}c_w^2+\frac{n}{t}\frac{\beta^2}{2\rho}\\
\leq&\sum_{i=1}^{n}\frac{4c_w c_1\sqrt{dp\ln(1.25/\delta)}}{m_i\epsilon\sqrt{t}}+\frac{nc_w^2(c_3+\lambda c_4/n)}{2}+\frac{nc_w^2\rho+n\beta^2/\rho}{2}.
\end{aligned}
\tag{91}
$$

□