Local Differential Privacy for Decentralized Online Stochastic Optimization with Guaranteed Optimality and Convergence Speed

Ziqin Chen and Yongqiang Wang, *Senior Member, IEEE*

***Abstract*—The increasing usage of streaming data has raised significant privacy concerns in decentralized optimization and learning applications. To address this issue, differential privacy has emerged as a standard approach for privacy protection in decentralized online optimization. Regrettably, existing differential-privacy solutions for decentralized online optimization face the dilemma of trading optimization accuracy for privacy. In this paper, we propose a local-differential-privacy solution for decentralized online optimization/learning that ensures both optimization accuracy and rigorous differential privacy, even in the infinite time horizon. Compared with our prior results that rely on a decaying coupling strength to gradually eliminate the influence of differential-privacy noises, the proposed approach allows the coupling strength to be time-invariant, which ensures a high convergence speed. Moreover, different from prior results which rely on precise gradient information to ensure optimality, the proposed approach can ensure convergence in mean square to the optimal solution even in the presence of stochastic gradients. We corroborate the effectiveness of our algorithm using multiple benchmark machine-learning applications, including logistic regression on the "mushrooms" dataset and CNN-based image classification on the "MNIST" and "CIFAR-10" datasets.**

***Index Terms*—Decentralized stochastic optimization, online learning, local differential privacy.**

I. Introduction

Decentralized stochastic optimization has become a focal point of research in machine learning, signal processing, and control [1], [2]. Its goal is to enable multiple participating agents to collaboratively learn a global model parameter $\theta^* \in \mathbb{R}^n$ that best fits all agents' local data. Mathematically, the problem of decentralized stochastic optimization can be formulated as

$$\min_{\theta\in\mathbb{R}^n} F(\theta)=\frac{1}{m}\sum_{i=1}^{m} f_i(\theta),\;\; f_i(\theta)=\mathbb{E}_{\xi^i\sim\mathcal{D}_i}\left[l(\theta,\xi^i)\right]. \quad (1)$$

Here, the local objective function $f_i(\theta):\mathbb{R}^n\to\mathbb{R}$ represents the mathematical expectation of agent i's loss function $l(\theta,\xi^i)$, where ξ^i denotes agent i's data-points drawn from distribution $\mathcal{D}_i$. In practice, because the data distribution $\mathcal{D}_i$ is generally unknown, it is impossible to analytically find an optimal solution to problem (1). To address this issue, empirical risk minimization (ERM) is usually employed to approximate the optimal solution to problem (1). In traditional ERM, all data are pre-stored and concurrently available [3], [4]. Nevertheless, in emerging applications such as autonomous vehicles, smart grids, and cloud computing, data are acquired in serial [5]. This necessitates a rethinking of off-the-shelf ERM-based solutions and has catalyzed the development of several decentralized online stochastic optimization/learning algorithms [6]–[9].

However, existing decentralized online algorithms require participating agents to share their intermediate parameters and/or gradient estimates, which can cause serious privacy leakage as these intermediate parameters or gradient estimates may contain sensitive information. In fact, recent studies have demonstrated that an external adversary can precisely recover the raw training data from intercepted gradient estimates [10]–[12], posing serious privacy threats to agents participating in decentralized optimization. To meet the urgent need for privacy protection, various privacy mechanisms have been proposed. One mechanism employs homomorphic encryption [13], [14]. While effective, this mechanism is computationally intensive. Hardware-based solutions like trusted hardware enclaves have also been explored, which, however, are server-dependent and cannot prevent data providers from inferring users' data in decentralized training [15]. Recent results have investigated exploiting time or spatially correlated noises for privacy protection [16]–[21]. Although these approaches ensure optimization accuracy by canceling out injected noises, they require each agent to have at least one neighbor not colluding with potential adversaries, which is often difficult to satisfy in many multi-agent applications.

As differential privacy (DP) is evolving as a gold standard for privacy protection [22], [23], plenty of efforts have been reported to incorporate DP in decentralized optimization [24]–[43]. One main difficulty in enabling DP in decentralized optimization lies in the fact that the conventional DP framework relies on a "centralized" data aggregator to collect raw data and inject noises, which may not be available in fully decentralized multi-agent networks [32], [33]. In addition, existing approaches are usually subject to a fundamental tradeoff between optimization accuracy and privacy [25]–[31], [35]–[40], which is undesirable in accuracy-sensitive applications. To tackle this dilemma, our prior result in [34] achieves accurate convergence and privacy protection simultaneously in decentralized offline optimization by introducing a weakening factor into inter-agent iteration to mitigate the impact of DP noises. However, this decaying coupling strength reduces the speed of algorithmic convergence. In addition, this approach requires all data to be static and predetermined, and hence,

The work was supported in part by the National Science Foundation under Grants ECCS-1912702, CCF-2106293, CCF-2215088, CNS-2219487, CCF-2334449, and CNS-2422312. (Corresponding author: Yongqiang Wang)

The authors are with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634 USA (e-mail: yongqiw@clemson.edu).

is inapplicable to online optimization/learning applications where data are acquired in serial. Moreover, by employing the local-differential-privacy (LDP) framework [44]–[46] to eliminate the requirement of a data aggregator in the conventional DP framework, we recently proposed an algorithm that can achieve both LDP and optimality in decentralized online learning [47]. However, this approach also hinges on the incorporation of a weakening factor, which significantly slows down algorithmic convergence.

In this paper, we propose an online approach that can ensure both LDP and optimality in decentralized stochastic optimization, even in the infinite time horizon. To facilitate convergence analysis, we first prove that the decentralized stochastic optimization problem has the same optimal solution as a newly formulated ERM problem under online data acquisition. Then, we propose a decentralized online optimization algorithm and prove that it can ensure convergence in mean square to the optimal solution to the ERM problem, and hence, can ensure convergence in mean square to the optimal solution to the original decentralized stochastic optimization problem. It is worth noting that our approach can ensure a finite cumulative privacy budget in the infinite time horizon without incorporating any weakening factors in inter-agent iterations, which is crucial in [34], [47] to simultaneously ensure optimality and DP in decentralized optimization. The avoidance of such a weakening factor enables us to achieve a higher convergence speed, as analytically proved in our theoretical analysis (Sec. IV-B) and numerically confirmed in our experimental results (Sec. VI).

The main contributions are summarized as follows:

- We prove that the optimal solution to problem (1) is the same as that to a newly formulated ERM problem under serial data acquisition. The result is also true under persistent DP noises, which, to the best of our knowledge, has not been reported before.
- In addition to ensuring convergence in mean square to the exact optimal solution, we also prove that our algorithm can simultaneously ensure a finite cumulative privacy budget, even when the number of iterations tends to infinity. This contrasts sharply with most existing DP solutions for decentralized optimization in [25]–[31], [35]–[40], [42], [43], where the privacy budget grows to infinity as time tends to infinity, implying diminishing DP protection in the infinite time horizon.
- Unlike existing results in [34], [47] which rely on a weakening factor in coupling strength to ensure both DP and optimality, our approach avoids weakening the coupling strength. Not only does avoiding a weakening factor simplify algorithm design by reducing the number of design parameters, but it also permits the coupling strength to be persistent, which ensures a high convergence speed (see Sec. IV-B for comparison results).
- Different from most existing results, which employ the traditional DP framework and (implicitly) rely on a "centralized" data aggregator to collect data and inject noises (see, e.g., [24]–[41]), our approach employs the LDP framework, which eliminates the need for any trusted data aggregators. Our approach also differs from existing LDP results for parameter-server-assisted federated learning (see, e.g., [48]–[51]), which are not applicable to the fully decentralized setting. Moreover, our implementation of ϵ^i-LDP provides a stricter privacy guarantee than the (ϵ^i,δ^i)-LDP framework used in [46].
- We evaluate the effectiveness of our algorithm by using machine-learning experiments, including logistic regression on the "mushrooms" dataset and CNN-based image classification on the "MNIST" and "CIFAR-10" datasets. Our experimental results confirm that the proposed approach is superior to existing counterparts in terms of both learning/test accuracies and convergence speed.

The rest of this paper is organized as follows. Sec. II formulates the problem and introduces necessary notations for later use. Sec. III presents the proposed algorithm. Sec. IV analyzes the optimization accuracy and convergence speed. Sec. V establishes the differential-privacy guarantee. Sec. VI provides experimental results. Sec. VII concludes the paper.

Notations: We use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space and $\mathbb{N}(\mathbb{N}^+)$ to denote the set of non-negative (positive) integers. We let $\otimes$ denote the Kronecker product. We write $\mathbf{1}_n$ and $\mathbf{0}_n$ for n-dimensional all-one and all-zero column vectors, respectively; in both cases we suppress the dimension when clear from the context. We use I_n to denote the n-dimensional identity matrix. We let W^T denote the transpose of a matrix W. We use $\langle\cdot,\cdot\rangle$ to denote the inner product of two vectors and $\|\cdot\|$ for the Euclidean norm of a vector. We write $\text{col}\{\theta_1,\cdots,\theta_m\}$ for the stacked column vector of $\theta_1,\cdots,\theta_m$. The notation $\lceil a\rceil$ refers to the smallest integer no less than a and $\lfloor a\rfloor$ represents the largest integer no greater than a. We use $\text{Lap}(\nu_i)$ to denote the Laplace distribution with a parameter $\nu_i>0$, featuring a probability density function $\frac{1}{2\nu_i}e^{\frac{-|x|}{\nu_i}}$. $\text{Lap}(\nu_i)$ has a mean of zero and a variance of $2\nu_i^2$. We abbreviate *independent and identically distributed* by *i.i.d.*

II. Preliminaries and Problem Formulation

In this section, we present some background information of LDP and the problem formulation.

A. Local differential privacy

Local differential privacy addresses the scenario where no trusted data aggregator exists to gather data and execute the privacy mechanism. As such, privacy protection for LDP is enforced at the agent level.

Before providing the definition of LDP, we first introduce the concept of adjacency on the local dataset of agent i:

Definition 1. *(Adjacency). For any $T\in\mathbb{N}^+$ and any agent $i\in[m]$, given two local datasets $\mathcal{S}^i=\{\xi_1^i,\xi_2^i,\cdots,\xi_T^i\}$ and $\mathcal{S}'^i=\{\xi_1'^i,\xi_2'^i,\cdots,\xi_T'^i\}$, $\mathcal{S}^i$ is said to be adjacent to $\mathcal{S}'^i$ if there exists a time instant $k\in\{1,\cdots,T\}$ such that $\xi_k^i\neq\xi_k'^i$ while $\xi_t^i=\xi_t'^i$ for all $t\in[1,T]$ and $t\neq k$.*

According to Definition 1, two local datasets $\mathcal{S}_t^i$ and $\mathcal{S}_t'^i$ are adjacent if they differ in only one element while all other elements are the same. We denote this adjacency relationship as $\text{Adj}(\mathcal{S}^i,\mathcal{S}'^i)$. With this notation, we present the definition of LDP:

Definition 2. *(LDP) Let $\mathcal{A}_i(\mathcal{S}^i, \theta^{-i})$ be an implementation of a decentralized algorithm by agent i, which takes agent i's dataset $\mathcal{S}^i$ and all received information θ^{-i} as input. Then, agent i's implementation $\mathcal{A}_i$ is said to be ϵ^i-LDP if for any adjacent datasets $\mathcal{S}^i$ and $\mathcal{S'}^i$, and the set of all possible observations $\mathcal{O}^i$, the following inequality holds:*

$$\mathbb{P}[\mathcal{A}_i(\mathcal{S}^i, \theta^{-i}) \in \mathcal{O}^i] \le e^{\epsilon^i}\mathbb{P}[\mathcal{A}_i(\mathcal{S'}^i, \theta^{-i}) \in \mathcal{O}^i]. \quad (2)$$

The definition of LDP implies that if datasets $\mathcal{S}^i$ and $\mathcal{S'}^i$ are adjacent, i.e., differ by a single data (record), their corresponding observations under $\mathcal{A}_i$ are very close in distribution. This ensures that any third party cannot infer agent i's private data from shared messages. The ratio of probabilities is bounded by e^{ϵ^i}, where ϵ^i is known as the *privacy budget* or privacy loss. A smaller ϵ^i implies a stronger privacy protection.

For each agent i's implementation of LDP, the total privacy budget ϵ^i should be preset. Every data access will consume some budget ϵ_t^i. Therefore, to maintain LDP throughout T iterations, the cumulative privacy budget must not exceed the preset budget, i.e., $\sum_{t=1}^{T}\epsilon_t^i \le \epsilon^i$.

To achieve ϵ_t^i-LDP at time t, each agent i adds Laplace noises to its shared messages. The amount of noise added is determined by sensitivity, which quantifies the maximum impact that a single data-point change can have on the output of agent i's implementation at time t:

Definition 3. *(Sensitivity) The sensitivity of each agent i's implementation $\mathcal{A}_i$ at time t is defined as*

$$\Delta_{t+1}^i = \max_{Adj(\mathcal{S}_t^i, \mathcal{S'}_t^i)} \|\mathcal{A}_i(\mathcal{S}_t^i, \theta_t^{-i}) - \mathcal{A}_i(\mathcal{S'}_t^i, \theta_t^{-i})\|_1, \quad (3)$$

where $\mathcal{S}_t^i$ represents agent i's dataset and θ_t^{-i} represents all received information acquired by agent i at time t.

Based on the concept of sensitivity, we introduce the following lemma that delineates a sufficient condition for ϵ^i-LDP over any time period T:

Lemma 1. *At time $t \in \mathbb{N}$, if agent i injects into each of its transmitted messages a noise vector ϑ_t^i consisting of n independent Laplace noises $Lap(\nu_t^i)$ with parameter ν_t^i such that $\sum_{t=1}^{T}\frac{\Delta_t^i}{\nu_t^i} \le \epsilon^i$, then agent i's implementation $\mathcal{A}_i$ is ϵ^i locally differentially private from time $t = 1$ to $t = T$.*

Proof. The lemma can be obtained following the same line of reasoning of Lemma 2 in [24]. □

Lemma 1 indicates that the cumulative privacy budget increases with an increase in the number of iterations T. In fact, in many existing DP solutions for decentralized optimization [25]–[30], [35]–[40], [42], [43], the cumulative privacy budget is allowed to grow to infinity as T tends to infinity. Based on the definition of DP, this implies that privacy protection will eventually be lost.

In this paper, to achieve rigorous DP protection, we ensure the cumulative privacy budget to be finite even when T tends to infinity.

Remark 1. In the LDP definition, each agent i treats all received information θ^{-i}, including the network topology and all messages received from neighboring agents, as external information that does not influence its DP design. Therefore, agents independently choose their privacy budgets ϵ^i and corresponding DP noises based on their practical needs, irrespective of other agents' actions. This differs from the centralized DP framework in existing algorithms [32], [33], which requires agents to mutually trust each other to cooperatively determine the DP-noise needed to guarantee a universal global privacy budget ϵ.

B. Decentralized online stochastic optimization

We consider a network of m agents that cooperatively learn an optimal solution θ^* to the stochastic optimization problem (1) with data acquired in serial. More specifically, at time t, each agent i acquires a data-point ξ_t^i, which is drawn from distribution $\mathcal{D}_i$. Each data-point ξ_t^i is associated with a loss function $l(\theta, \xi_t^i)$, which is a random realization of the local objective function $f_i(\theta)$. The agents cooperate to minimize the average objective function $F(\theta) = \frac{1}{m}\sum_{i=1}^{m} f_i(\theta)$. We assume that the m agents interact on an undirected and connected graph $\mathcal{G} = ([m], \mathcal{E})$, where $[m] = \{1, \cdots, m\}$ denotes the set of agents and $\mathcal{E} \subseteq [m] \times [m]$ denotes the edge set. The neighboring set of agent i is denoted as $\mathcal{N}_i = \{j : (i, j) \in \mathcal{E}\}$. We define a weight matrix as $W = \{w_{ij}\} \in \mathbb{R}^{m\times m}$, where w_{ij} gives the weight of edge $(i, j) \in \mathcal{E}$ with the convention that $w_{ij} > 0$ if $(i, j) \in \mathcal{E}$, and $w_{ij} = 0$ otherwise. We define $w_{ii} = -\sum_{j\in\mathcal{N}_i} w_{ij}$.

To facilitate subsequent analysis, we make the following standard assumptions:

Assumption 1. *The gradients of local objective functions $f_i(\theta)$ are uniformly bounded, i.e., there exists some $D > 0$ such that we have $\|\nabla f_i(\theta)\| \le D$ for all $i \in [m]$ and $\theta \in \mathbb{R}^n$.*

Assumption 1 is standard in nonconvex optimization (see, e.g., [52]–[56]). In fact, in many machine learning applications, the technique of gradient clipping is used to make the norm of gradients no larger than some threshold value [57], [58], which will make our bounded-gradient assumption hold automatically.

Assumption 2. *We assume that the random data-points $\{\xi_k^i\}$ of agent i are i.i.d across different time instants $k \in [0, t]$. In addition, (i) $\mathbb{E}[\nabla l(\theta, \xi_k^i)] = \nabla f_i(\theta)$; (ii) $\mathbb{E}[\|\nabla l(\theta, \xi_k^i) - \nabla f_i(\theta)\|^2] \le \kappa^2$; and (iii) $\|\nabla l(\theta_1, \xi_k^i) - \nabla l(\theta_2, \xi_k^i)\| \le L\|\theta_1 - \theta_2\|$ for any $\theta_1, \theta_2 \in \mathbb{R}^n$.*

Assumption 2 does not require data-points to be *i.i.d.* among different agents. Moreover, Assumptions 2-(i)-(iii) are standard for convergence analysis in federated learning (see, e.g., [59], [60]) and decentralized stochastic optimization (see, e.g., [9], [10], [61]–[63]). In addition, Assumption 2-(iii) ensures the smoothness of loss functions. A large number of loss functions satisfy this assumption, with typical examples including the widely used cross-entropy loss and its variants [64].

Assumption 3. *The weight matrix $W = \{w_{ij}\} \in \mathbb{R}^{m\times m}$ is symmetric and satisfies $\mathbf{1}^T W = \mathbf{0}^T$ and $W\mathbf{1} = \mathbf{0}$. The eigenvalues of W satisfy (after arranged in an increasing order) $-1 < \delta_m \le \cdots \le \delta_2 < \delta_1 = 0$.*

Assumption 3 is standard for achieving average consensus among agents in fully decentralized optimization/learning (see discussions in, e.g., [8], [9], [28], [29], [42]). Moreover, the work [65] has shown that decentralized average consensus (under constant weights w_{ij}) is achievable if and only if w_{ij} satisfies $0 < w_{ij} < \frac{2}{\max_{\{i,j\}\in\mathcal{E}}(d_i+d_j)}$, where d_i and d_j represent the number of neighbors of agent i and agent j, respectively. Since d_i and d_j are always no less than 1, we have that the weights w_{ij} are always less than 1 (see Section 4.1 in [65] for more detailed discussions).

Given that agent i is not aware of data distribution $\mathcal{D}_i$ in problem (1), to solve for (1), it is natural to minimize the following online empirical objective function:

$$\min\nolimits_{\theta\in\mathbb{R}^n} F_t(\theta)=\frac{1}{m}\sum_{i=1}^{m} f_t^i(\theta),\ \ f_t^i(\theta)=\frac{1}{t+1}\sum_{k=0}^{t} l(\theta,\xi_k^i). \tag{4}$$

Remark 2. The online ERM is inspired by the classic ERM formulation [66], but has a remarkable difference. Specifically, in the decentralized ERM formulation [66], a static empirical objective function $\tilde{F}(\theta)=\frac{1}{m}\sum_{i=1}^{m}\tilde{f}_i(\theta)$ is minimized, where $\tilde{f}_i(\theta) = \frac{1}{n_i}\sum_{s=1}^{n_i} l(\theta,\xi_s^i)$ approximates the expected loss function using a static dataset $\mathcal{S}^i = \{\xi_1^i,\cdots,\xi_{n_i}^i\}$. Obviously, this dataset $\mathcal{S}^i$ is available to agent i before algorithm implementation, and hence, $\tilde{f}_i(\theta)$ and $\tilde{F}(\theta)$ are static. In contrast, our ERM formulation involves dynamic datasets for each agent that grow over time, leading to time-varying local empirical objective functions $f_t^i(\theta)$. The dynamic and evolving data landscape brings challenges in ensuring both algorithmic convergence and optimality in algorithm designs.

Now, we prove that the optimal solution to the proposed online ERM problem converges in mean square to the optimal solution to the stochastic optimization problem in (1):

Lemma 2. *Denote θ_t^* and θ^* as the optimal solution to the online ERM problem* (4) *at time t and the optimal solution to the original stochastic optimization problem* (1)*, respectively. Under Assumption 2, if $F(\theta)$ is μ-strongly convex, the following inequality always holds:*

$$\mathbb{E}\big[\|\theta_t^*-\theta^*\|^2\big]\leq\frac{4\kappa^2}{\mu^2(t+1)},\ \forall t\in\mathbb{N}. \tag{5}$$

Proof. Given the relationship $F_t(\theta_t^*)\leq F_t(\theta^*)$, we have

$$F(\theta_t^*)-F(\theta^*)\leq(F(\theta_t^*)-F_t(\theta_t^*))-(F(\theta^*)-F_t(\theta^*)). \tag{6}$$

The mean value theorem implies

$$\begin{aligned}&(F(\theta_t^*)-F_t(\theta_t^*))-(F(\theta^*)-F_t(\theta^*))\\&=\langle\nabla F(\chi)-\nabla F_t(\chi),\theta_t^*-\theta^*\rangle\\&\leq\|\nabla F(\chi)-\nabla F_t(\chi)\|\|\theta_t^*-\theta^*\|,\end{aligned} \tag{7}$$

where the variable χ is given by $\chi=\alpha\theta_t^*+(1-\alpha)\theta^*$ for some constant $\alpha\in(0,1)$.

Defining $\nabla F(\chi)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\nabla l(\chi,\xi^i)]$ leads to

$$\begin{aligned}&\mathbb{E}[\|\nabla F_t(\chi)-\nabla F(\chi)\|]=\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_t^i(\chi)-\nabla F(\chi)\right\|\right]\\&\leq\frac{1}{m}\sum_{i=1}^{m}\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}\left[\|\nabla l(\chi,\xi_k^i)-\mathbb{E}[\nabla l(\chi,\xi_k^i)]\|\right].\end{aligned} \tag{8}$$

Since the data-points $\{\xi_k^i\}$ for agent i are *i.i.d.* across different time instants, we use the Lyapunov inequality $E[\|X\|]\leq(E[\|X\|^p])^{\frac{1}{p}}$ for any $p\geq 1$ and Assumption 2-(ii) to obtain

$$\begin{aligned}&\sum_{k=0}^{t}\mathbb{E}\left[\|\nabla l(\chi,\xi_k^i)-\mathbb{E}[\nabla l(\chi,\xi_k^i)]\|\right]\\&\leq\sqrt{\mathbb{E}\left[\left(\sum_{k=0}^{t}\|\nabla l(\chi,\xi_k^i)-\mathbb{E}[\nabla l(\chi,\xi_k^i)]\|\right)^2\right]}\\&=\sqrt{\mathbb{E}\left[\sum_{k=0}^{t}\left\|\nabla l(\chi,\xi_k^i)-\nabla f_i(\chi)\right\|^2\right]}\leq\kappa\sqrt{t+1}.\end{aligned} \tag{9}$$

Incorporating (9) into (8) yields $\mathbb{E}\left[\|\nabla F_t(\chi)-\nabla F(\chi)\|\right]\leq\frac{\kappa}{\sqrt{t+1}}$. By using (6) and (7), we have

$$\mathbb{E}\left[F(\theta_t^*)-F(\theta^*)\right]\leq\frac{\kappa}{\sqrt{t+1}}\mathbb{E}\left[\|\theta_t^*-\theta^*\|\right]. \tag{10}$$

Given that $F(\theta)$ is strongly convex, we have $\frac{\mu}{2}\|\theta_t^*-\theta^*\|^2\leq F(\theta_t^*)-F(\theta^*)$. By using (10), we obtain

$$\frac{\mu}{2}\mathbb{E}\left[\|\theta_t^*-\theta^*\|^2\right]\leq\frac{\kappa}{\sqrt{t+1}}\mathbb{E}\left[\|\theta_t^*-\theta^*\|\right],$$

which implies $\mathbb{E}\left[\|\theta_t^*-\theta^*\|\right]\leq\frac{2\kappa}{\mu}(t+1)^{-\frac{1}{2}}$ and inequality (5). □

Lemma 2 ensures that the optimal solution to the online ERM problem in (4) converges in mean square to the optimal solution to the stochastic optimization problem in (1). With this understanding, we aim to develop a decentralized online optimization algorithm that generates a sequence $\{\theta_t^i\}$ to track the optimal solution to (4) under LDP constraints. Based on the results in Lemma 2, this optimal solution will also converge in mean square to the optimal solution to (1) even under LDP constraints.

III. Locally Differentially Private Decentralized Online Learning Algorithm Design

We propose Algorithm 1 to solve stochastic optimization problem (1) while ensuring rigorous ϵ^i-LDP. The injected DP noise satisfies the following assumption:

Assumption 4. *For each agent $i\in[m]$ and at any time $t\geq 0$, the DP noise ϑ_t^i remains independent across iterations and satisfies $\mathbb{E}[\vartheta_t^i]=0$ and $\mathbb{E}[\|\vartheta_t^i\|^2]=(\sigma_t^i)^2$. The variance is given by $\sigma_t^i=\frac{\sigma_0^i}{(t+1)^{\varsigma^i}}$ with $\sigma_0^i>0$ and $\varsigma^i\in(\frac{1}{2},1)$. Moreover, the following inequality always holds:*

$$\max_{i\in[m]}\{\varsigma^i\}<v<1, \tag{11}$$

where the constant v is the decaying rate of the stepsize λ_t.

Our algorithm can ensure accurate convergence despite the presence of persistent DP noises. This is fundamentally different from existing DP solutions for decentralized optimization that patch DP noises with a given existing decentralized optimization/learning algorithm (see, e.g., [25]–[31], [35]–[40]), which do not fully exploit the flexibilities in the design of optimization algorithm or noise-injection mechanism.

Algorithm 1 Locally differentially private decentralized online optimization for agent $i \in [m]$

```
Input: Random initialization $\theta_0^i \in \mathbb{R}^n$; stepsize $\lambda_t = \frac{\lambda_0}{(t+1)^v}$ with $\lambda_0 > 0$ and $v \in (\frac{1}{2}, 1)$; and DP-noise variance $\sigma_t^i = \frac{\sigma_0^i}{(t+1)^{\varsigma^i}}$ with $\sigma_0^i > 0$ and $\varsigma^i \in (\frac{1}{2}, 1)$.
for $t = 0, 1, \cdots, T-1$ do
    Acquire a new data-point $\xi_t^i \sim \mathcal{D}_i$.
    Compute the gradient $\nabla f_t^i(\theta_t^i) = \frac{1}{t+1}\sum_{k=0}^{t} \nabla l(\theta_t^i, \xi_k^i)$ by using all available data up to time $t$, i.e., $\xi_k^i \in \mathcal{S}_t^i,\ k \in [0, t]$ and the current parameter $\theta_t^i$.
    Receive neighbors' parameters $y_t^j = \theta_t^j + \vartheta_t^j,\ j \in \mathcal{N}_i$.
    $\theta_{t+1}^i = \theta_t^i + \sum_{j \in \mathcal{N}_i} w_{ij}(y_t^j - \theta_t^i) - \lambda_t \nabla f_t^i(\theta_t^i)$.
    Add DP noises $\vartheta_{t+1}^i$ to $\theta_{t+1}^i$ and then send the obscured parameter $y_{t+1}^i = \theta_{t+1}^i + \vartheta_{t+1}^i$ to its neighbors.
end for
```

One key reason for our algorithm to be robust to DP noises is the use of all data available at time t to compute the gradient. This strategy can enhance optimization accuracy compared with existing online algorithms that only use the single data-point of the current time instant in each iteration (see, e.g., [35]–[40]). The advantage of our proposed strategy is also clearly demonstrated in our experimental results (see Fig. 3 for details). Moreover, this strategy also helps achieving rigorous ϵ^i-LDP with a finite cumulative privacy budget in the infinite time horizon (see Eq. (28)), which is unattainable in most existing DP solutions for decentralized optimization [25]–[31], [35]–[40]. One drawback of this strategy is that it increases the computational complexity compared with the strategy of using one data-point per iteration. However, for continuous loss functions, our prior work [47] shows that the increased computational complexity can be made independent of the iteration number by using interpolation.

Another key reason for our algorithm's achieving of both accurate convergence and rigorous differential privacy is the co-design of the stepsize and DP-noise injection mechanism. By judiciously designing the decaying rate of stepsize (v) and the decaying rate of DP-noise variances (ς^i, see Assumption 4 for details), we can ensure a reduced sensitivity of our algorithm, which is key to ensure a finite cumulative privacy budget in the infinite time horizon. This is different from existing DP approaches for decentralized online optimization/learning [35]–[40], which employ the same decaying rate for stepsizes and DP-noise variances, leading to either the loss of accurate convergence (when the decaying rate is low) or a diminishing privacy protection as iteration proceeds (as the cumulative privacy budget will explode as the number of iterations tends to infinity when the decaying rate is high).

Remark 3. Our algorithm let each agent add DP noise to its local parameter before sharing it with neighboring agents. It has been shown that sharing local parameters directly can leak sensitive information of local training datasets $\mathcal{S}_t^i$. For example, in [61], it has been shown that sharing intermediate parameters allows an adversary to precisely recover the raw data ξ_t^i. Moreover, our shared parameter $y_t^j = \theta_t^j + \vartheta_t^j$ has the same dimension as the optimization parameter θ_t^j, which does not incur extra communication overhead in each iteration compared with traditional distributed optimization algorithms that do not consider privacy [6]–[9].

Remark 4. Note that the stepsize λ_t in our algorithm can be hard-coded in each agent's program prior to implementation, and hence, it does not necessitate any adjustment or coordination among agents during algorithmic implementation.

Remark 5. Our recent works on decentralized batch (offline) optimization [34] and decentralized online learning [47] employ a weakening factor on inter-agent iteration to attenuate the influence of DP noises, and hence, enable both optimality and differential privacy. However, this weakening factor leads to decaying coupling strength, which in turn reduces the speed of algorithmic convergence. As rigorously proven in Sec. IV-B and illustrated in experimental results in Fig. 1-Fig. 3, by avoiding using the weakening factor, Algorithm 1 ensures a higher convergence speed compared with our prior results in [47] while ensuring the same strength of privacy protection.

IV. Optimization Accuracy and Convergence Speed Analysis

A. Optimization accuracy analysis

In this subsection, we prove that the intermediate parameter of Algorithm 1 converges in mean square to the optimal solution to problem (1) when the global objective function is strongly convex. For general convex global objective functions, we prove that the objective function value converges in mean to the minimal objective function value. For nonconvex global objective functions, we prove that the gradient value of the objective function converges in mean square to zero. We first give a preliminary result.

Lemma 3. *Denote θ_t^* as the optimal solution to the ERM problem* (4) *at time t $\in \mathbb{N}$. Under Assumption 2, if the objective function is strongly convex, then the optimal solution to the online ERM problem* (4) *satisfies*

$$\mathbb{E}\left[\|\theta_{t+1}^* - \theta_t^*\|^2\right] \leq \frac{16(\kappa^2 + D^2)}{(t+1)^2}\left(\frac{2}{\mu^2} + \frac{1}{L^2}\right). \tag{12}$$

Proof. The lemma can be obtained following the same line of reasoning of Lemma 1 in [47]. □

Lemma 3 establishes the mean square convergence of the optimal solution to problem (4). Using Lemma 3, we further characterize the tracking error between the intermediate parameter θ_t^i of Algorithm 1 and the optimal solution θ_t^* to the ERM problem (4).

Theorem 1. *Under Assumptions 1-4, if the global objective function is strongly convex and the initial value of the stepsize satisfies $\lambda_0 \in (0, \frac{-\delta_2 \mu}{2\mu^2 + 16L^2}]$, then we have:*

$$\mathbb{E}[\|\theta_t^i - \theta_t^*\|^2] \leq \mathcal{O}(t^{-\beta}),\ \forall t \in \mathbb{N}^+, \tag{13}$$

with $\beta = \min\{\frac{3-3v}{2}, 2\varsigma - \frac{v+1}{2}\}$ and $\varsigma = \min_{i \in [m]}\{\varsigma^i\}$.

Proof. See Appendix A. □

Theorem 1 characterizes the deviation between the intermediate parameter θ_t^i and the optimal solution θ_t^* to (4).

Besides providing insights into tuning algorithms to adapt to spatiotemporal fluctuations, it also enables us to evaluate the accuracy of intermediate parameters in solving the original decentralized stochastic optimization problem (1):

Theorem 2. *Denote θ^* as the optimal solution to the original stochastic optimization problem* (1). *Under the conditions in Theorem 1, the parameters θ_t^i generated by Algorithm 1 will converge in mean square to θ^*, i.e.,*

$$\mathbb{E}[\|\theta_t^i - \theta^*\|^2] \leq \mathcal{O}(t^{-\beta}), \tag{14}$$

with $\beta = \min\{\frac{3-3\upsilon}{2}, 2\varsigma - \frac{\upsilon+1}{2}\}$ and $\varsigma = \min_{i\in[m]}\{\varsigma^i\}$.

Proof. Incorporating (5) from Lemma 2 and (13) from Theorem 1 into the triangle inequality $\|\theta_t^i - \theta^*\|^2 \leq 2\|\theta_t^i - \theta_t^*\|^2 + 2\|\theta_t^* - \theta^*\|^2$, we arrive at (14). □

Theorem 2 quantifies the expected difference between the intermediate parameter and the optimal solution to problem (1). Compared with existing results for decentralized optimization/learning [35]–[39] that focus on characterizing some regret value with respect to the solution to an approximated problem of (1), e.g., problem (4), our results directly characterize the tracking error with respect to the optimal solution to the actual stochastic optimization problem (1). Not only does our approach provide a direct characterization of the tracking performance with respect to the actual optimal solution, it also provides a new perspective to solve decentralized stochastic optimization problems.

Next, we establish the convergence of Algorithm 1 under general convex objective functions.

Theorem 3. *Under Assumptions 1-4, if the global objective function is convex, then we have*

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[F(\theta_t^i) - F(\theta^*)] \leq \mathcal{O}(T^{-(1-\upsilon)}). \tag{15}$$

Proof. See Appendix B. □

Theorem 3 shows that the objective function value $F(\theta_t^i)$ converges in mean to the minimal objective function value $F(\theta^*)$ of problem (1). This result is stronger and more precise than the convergence result presented in [47], which only characterizes the expected distance between the instantaneous empirical objective function value $F_t(\theta_t^i)$ and the minimal empirical objective function value $F_t(\theta_t^*)$.

Theorem 4. *Under Assumptions 1-4, if the global objective function is nonconvex and the initial value of the stepsize satisfies $\lambda_0 \in (0, \frac{-\delta_2}{2(L^2+D^2)}]$, then we have*

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\theta_t^i)\|^2] \leq \mathcal{O}(T^{-(1-\upsilon)}). \tag{16}$$

Proof. See Appendix C. □

Theorem 4 shows that the gradient value $\|\nabla F(\theta_t^i)\|$ of the objective function converges in mean square to zero. This result demonstrates the effectiveness of our Algorithm 1 even for nonconvex objective functions.

B. Discussion on convergence speed

In this subsection, we compare the convergence speed of Algorithm 1 with that of our previous decentralized online learning algorithm in [47], which ensures convergence by using a weakening factor γ_t to mitigate the influence of DP noises. The detailed algorithm in [47] is given as follows:

$$\theta_{t+1}^i = \Pi_\Theta\left[\theta_t^i + \gamma_t \sum_{j\in\mathcal{N}_i} w_{ij}(\theta_t^j + \check{\vartheta}_t^j - \theta_t^i) - \check{\lambda}_t \nabla f_t^i(\theta_t^i)\right], \tag{17}$$

where the parameters are given by $\gamma_t = \frac{\gamma_0}{(t+1)^u}$, $\check{\lambda}_t = \frac{\check{\lambda}_0}{(t+1)^{\check{\upsilon}}}$, and $\mathbb{E}[\|\check{\vartheta}_t^i\|^2] = (\check{\sigma}_t^i)^2$ with $\check{\sigma}_t^i = \check{\sigma}_0^i(t+1)^{\check{\varsigma}^i}$. Under the condition $\check{\upsilon} > u > \check{\varsigma} + \frac{1}{2}$ with $\check{\varsigma} = \max_{i\in[m]}\{\check{\varsigma}^i\} \in (0, \frac{1}{2})$, [47] obtains a convergence speed of $\mathcal{O}(t^{-\check{\beta}})$ with $\check{\beta} = \min\{1 - \check{\upsilon}, 2(u - \check{\varsigma}) - 1\}$ for strongly convex objective functions and $\mathcal{O}(t^{-\check{\beta}})$ with $\check{\beta} = \frac{1-\check{\upsilon}}{2}$ for general convex objective functions (note that [47] does not consider the nonconvex case).

In our Algorithm 1, by avoiding using the weakening factor γ_t, we can keep the inter-agent coupling strength to be persistent, which enables us to achieve faster convergence. More specifically, noting $\upsilon > \varsigma$ with $\varsigma = \max_{i\in[m]}\{\varsigma^i\} \in (\frac{1}{2}, 1)$, we have that the convergence speed of our Algorithm 1 is $\mathcal{O}(t^{-\beta})$ with $\beta = \min\{1 - \upsilon + \frac{1-\upsilon}{2}, 2\varsigma - 1 + \frac{1-\upsilon}{2}\}$ and $\varsigma = \min_{i\in[m]}\{\varsigma^i\}$ for strongly convex objective functions and $\mathcal{O}(t^{-\beta})$ with $\beta = 1 - \upsilon$ for general convex objective functions. Therefore,

(i) Under the same decaying rates of stepsizes and DP noises, i.e., $\check{\upsilon} = \upsilon$ and $u - \check{\varsigma} = \varsigma$, the convergence speed of our Algorithm 1 outpaces that of algorithm (17) by a factor of $\mathcal{O}(t^{\frac{1-\upsilon}{2}})$ for strongly convex objective functions. This improvement is also substantiated by our experimental results in Fig. 1.

(ii) By observing the prerequisites $\check{\upsilon} > u > \check{\varsigma} + \frac{1}{2}$ in algorithm (17) and $\upsilon > \varsigma$ in Algorithm 1, our Algorithm 1 allows for more slowly decaying stepsizes than algorithm (17), i.e., $\upsilon < \check{\upsilon}$, which enables Algorithm 1 to acquire a higher convergence speed than algorithm (17) even in the presence of general convex objective functions. In this case, the convergence speed of Algorithm 1 outpaces that of algorithm (17) by a factor of $\mathcal{O}(t^{\frac{1-\upsilon+(\check{\upsilon}-\upsilon)}{2}})$. This improvement is also confirmed by our experimental results in Fig. 2 and Fig. 3, which show that Algorithm 1 attains higher training and test accuracies than algorithm (17).

V. Local differential privacy analysis

This section establishes that Algorithm 1 can ensure rigorous ϵ^i-LDP for each agent, even in the infinite time horizon. We first introduce the following preliminary result:

Lemma 4. *Denote $\{\upsilon_t\}$ as a nonnegative sequence. If there exists a sequence $\beta_t = \frac{\beta_0}{(t+1)^s}$ with some $\beta_0 > 0$ and $s > 0$ such that $\upsilon_{t+1} \leq (1-\alpha)\upsilon_t + \beta_t$ holds for all $\alpha \in (0,1)$, then we always have $\upsilon_t \leq C\beta_t$ for all $t \in \mathbb{N}$, where the constant C is given by $C = (\frac{4s}{e\ln(\frac{2}{2-\alpha})})^s(\frac{\upsilon_0(1-\alpha)}{\beta_0} + \frac{2}{\alpha})$.*

Proof. The lemma can be obtained following the same line of reasoning of Lemma 11 in [67]. □

Without loss of generality, we consider adjacent datasets $\mathcal{S}_T^i$ and $\mathcal{S}_T'^i$ that differ at the k-th element, i.e., ξ_k^i in $\mathcal{S}_T^i$ and $\xi_k'^i$ in $\mathcal{S}_T'^i$, where T denotes the total number of iterations. For the sake of clarity, the parameters learned from $\mathcal{S}_t^i$ and $\mathcal{S}_t'^i$ are denoted as θ_{t+1}^i and $\theta_{t+1}'^i$, respectively. In addition, we introduce the following assumption, which is standard in existing DP solutions for decentralized optimization and learning (see, e.g., [24], [28], [35]–[40], [43]):

Assumption 5. *There exists some positive constant d such that $\|\nabla l(\theta,\xi^i)\|_1 \leq d$ holds for all $\theta \in \mathbb{R}^n$ and $i \in [m]$.*

Theorem 5. *Let Assumptions 2-5 hold. If the nonnegative stepsize sequence λ_t satisfies the condition in Theorem 1, and each component of ϑ_t^i follows the Laplace distribution Lap(ν_t^i) with $(\sigma_t^i)^2 = 2(\nu_t^i)^2$ in line with Assumption 4, then θ_t^i in Algorithm 1 converges in mean square to the optimal solution to the original stochastic optimization problem* (1). *Furthermore,*
(1) For any finite number of iterations T, agent i's implementation is LDP with a cumulative privacy budget bounded by $\epsilon^i \leq \sum_{t=1}^{T} \frac{2\sqrt{2}d\varrho_t(t+1)^{\varsigma^i}}{\sigma_0^i}$, where the parameter ϱ_t is given by $\varrho_t = \sum_{p=1}^{t-1}(1-\bar{w})^{t-p}\lambda_{p-1}+\lambda_{t-1}$ and the constant $\bar{w}$ is given by $\bar{w} = \min\{|w_{ii}|\},\ i \in [m]$.
(2) The cumulative privacy budget is finite even when $T \to \infty$.

Proof. The convergence result follows directly from Theorem 2.

(1) To prove the statement on privacy, we begin by analyzing the sensitivity of agent i's implementation under Algorithm 1. From Definition 3, it is evident that $\theta_t^j + \vartheta_t^j = \theta_t'^j + \vartheta_t'^j$ is valid for all $t \geq 0$ and $j \in \mathcal{N}_i$. Given that only the k-th data-point differs between $\mathcal{S}_T^i$ and $\mathcal{S}_T'^i$, when $t < k$, we have $\theta_t^i = \theta_t'^i$. When $t \geq k$, since the difference in loss functions kicks in at time k, i.e., $\nabla l(\theta,\xi_k^i) \neq \nabla l(\theta,\xi_k'^i)$, we have $\theta_t^i \neq \theta_t'^i$ from time k up to t. Hence, for agent i's implementation of Algorithm 1, we have

$$\begin{aligned}\|\theta_{t+1}^i - \theta_{t+1}'^i\|_1 \leq \big\|&(1+w_{ii})(\theta_t^i - \theta_t'^i) \\ &- \frac{\lambda_t}{t+1}\sum_{p=k}^{t}(\nabla l(\theta_t^i,\xi_p^i) - \nabla l(\theta_t'^i,\xi_p'^i))\big\|_1,\end{aligned} \tag{18}$$

for all $t \geq k$ and any $k \geq 0$, where we have used the definition $w_{ii} = -\sum_{j\in\mathcal{N}_i} w_{ij}$.

Letting $\bar{w} = \min\{|w_{ii}|\}$ for all $i \in [m]$ and using the definition $\theta_{t+1}^i = \mathcal{A}_i(\mathcal{S}_t^i,\theta_t^{-i})$, inequality (18) satisfies

$$\begin{aligned}&\|\mathcal{A}_i(\mathcal{S}_t^i,\theta_t^{-i}) - \mathcal{A}_i(\mathcal{S}_t'^i,\theta_t^{-i})\|_1 \\ &\leq (1-\bar{w})\max_{\text{Adj}(\mathcal{S}_{t-1}^i,\mathcal{S}_{t-1}'^i)}\|\mathcal{A}_i(\mathcal{S}_{t-1}^i,\theta_{t-1}^{-i}) - \mathcal{A}_i(\mathcal{S}_{t-1}'^i,\theta_{t-1}^{-i})\|_1 \\ &\quad + \frac{\lambda_t}{t+1}\textstyle\sum_{p=k}^{t}\|\nabla l(\theta_t^i,\xi_p^i) - \nabla l(\theta_t'^i,\xi_p'^i)\|_1 \\ &\leq (1-\bar{w})\Delta_t^i + \frac{\lambda_t}{t+1}\textstyle\sum_{p=0}^{t}\|\nabla l(\theta_t^i,\xi_p^i) - \nabla l(\theta_t'^i,\xi_p'^i)\|_1,\end{aligned} \tag{19}$$

where we have used $\sum_{p=0}^{k-1}\nabla l(\theta_t^i,\xi_p^i) = \sum_{p=0}^{k-1}\nabla l(\theta_t'^i,\xi_p'^i)$ in the second inequality.

Taking the maximum on adjacent datasets $\mathcal{S}_t^i$ and $\mathcal{S}_t'^i$ on both sides of (19) and using (3), we obtain

$$\begin{aligned}&\max_{\text{Adj}(\mathcal{S}_t^i,\mathcal{S}_t'^i)}\|\mathcal{A}_i(\mathcal{S}_t^i,\theta_t^{-i}) - \mathcal{A}_i(\mathcal{S}_t'^i,\theta_t^{-i})\|_1 = \Delta_{t+1}^i \\ &\leq (1-\bar{w})\Delta_t^i + \frac{\lambda_t}{t+1}\sum_{p=0}^{t}\|\nabla l(\theta_t^i,\xi_p^i) - \nabla l(\theta_t'^i,\xi_p'^i)\|_1.\end{aligned} \tag{20}$$

Note that we have used the fact that Δ_t^i in (20) is a positive constant independent of Adj$(\mathcal{S}_t^i,\mathcal{S}_t'^i)$. Therefore, the sensitivity Δ_{t+1}^i satisfies

$$\Delta_{t+1}^i \leq (1-\bar{w})\Delta_t^i + \frac{\lambda_t}{t+1}\sum_{p=0}^{t}\|\nabla l(\theta_t^i,\xi_p^i) - \nabla l(\theta_t'^i,\xi_p'^i)\|_1. \tag{21}$$

By iterating (21) from $t=1$ to $t=T$ and using $\Delta_0^i = 0$ and Assumption 5, we arrive at

$$\Delta_t^i \leq 2d\left(\sum_{p=1}^{t-1}(1-\bar{w})^{t-p}\lambda_{p-1} + \lambda_{t-1}\right). \tag{22}$$

Therefore, for agent i, the cumulative privacy budget ϵ^i over T iterations is bounded by $\sum_{t=1}^{T}\frac{2\sqrt{2}d\varrho_t(t+1)^{\varsigma^i}}{\sigma_0^i}$, where ϱ_t is defined in the theorem statement.

(2) By leveraging inequality (21) and the fact $\xi_p^i = \xi_p'^i$ for all $p \neq k$, we have

$$\begin{aligned}\Delta_{t+1}^i \leq{}& (1-\bar{w})\Delta_t^i + \frac{\lambda_t}{t+1}\|\nabla l(\theta_t^i,\xi_k^i) - \nabla l(\theta_t'^i,\xi_k'^i)\|_1 \\ &+ \frac{\lambda_t}{t+1}\sum_{p=0,\ p\neq k}^{t}\|\nabla l(\theta_t^i,\xi_p^i) - \nabla l(\theta_t'^i,\xi_p^i)\|_1,\end{aligned} \tag{23}$$

for all $t \geq k$ and any $k \geq 0$. The Lipschitz condition in Assumption 2-(iii) implies that for the same data-points ξ_p^i, we can rewrite (23) as follows:

$$\Delta_{t+1}^i \leq \left(1 - \bar{w} + \sqrt{n}L\frac{\lambda_t t}{t+1}\right)\Delta_t^i + \frac{2\lambda_t d}{t+1}. \tag{24}$$

For the stepsize $\lambda_t = \frac{\lambda_0}{(t+1)^v}$, there always exist some $T_0 \geq 0$ and some constant $C_2 > 0$ satisfying $C_2\bar{w} < 1$ such that

$$\bar{w} - \sqrt{n}L\frac{\lambda_t t}{t+1} = \bar{w} - \frac{\sqrt{n}L\lambda_0 t}{(t+1)^{v+1}} \geq C_2\bar{w}, \tag{25}$$

holds for all $t \geq T_0$. Combining (24) and (25) yields the following inequality for all $t \geq T_0$:

$$\Delta_{t+1}^i \leq (1 - C_2\bar{w})\,\Delta_t^i + \frac{2\lambda_0 d}{(t+1)^{1+v}}. \tag{26}$$

Furthermore, let $C_3 = \max\{(\frac{4(1+v)}{e\ln(\frac{2}{2-C_2\bar{w}})})^{1+v}(\frac{\Delta_0^i(1-C_2\bar{w})}{2\lambda_0 d} + \frac{2}{C_2\bar{w}}), \max_{0\leq t<T_0, i\in[m]}\{\frac{\Delta_t^i(t+1)}{2\lambda_t d}\}\}$. Using Lemma 4, we have that the sensitivity of agent i's implementation of Algorithm 1 satisfies

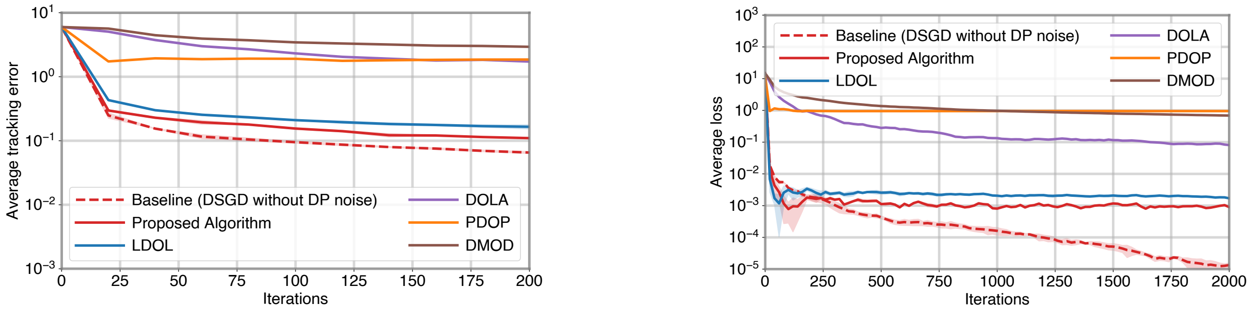
$$\Delta_t^i \leq C_3\frac{2\lambda_0 d}{(t+1)^{1+v}}, \tag{27}$$

for all $t \geq 0$. Hence, by using Lemma 2, we arrive at

$$\sum_{t=1}^{T}\frac{\Delta_t^i}{\nu_t^i} \leq \sum_{t=1}^{T}\frac{2\sqrt{2}\lambda_0 C_3 d}{\sigma_0^i(t+1)^{1+v-\varsigma^i}}, \tag{28}$$

implying a finite cumulative privacy budget ϵ^i even when T tends to infinity given that $v - \varsigma^i$ is always positive. □

Theorem 5 shows that our Algorithm 1 can preserve rigorous ϵ^i-LDP for the entire iteration process, even when the number of iterations T tends to infinity. It effectively solves the problem in existing DP solutions for distributed optimization and learning [25]–[31], [35]–[40] that the cumulative privacy



(a) Average tracking error

(b) Average loss

Fig. 1. Comparison of online logistic regression results by using the "mushrooms" dataset.

budget grows to infinite as time tends to infinity. Moreover, we ensure a finite cumulative privacy budget without using any weakening factors in [34], [47], which enables achieving faster convergence (see detailed discussions in Sec. IV-B).

Remark 6. A key reason for Algorithm 1 to ensure a finite cumulative privacy budget in the infinite time horizon is that our algorithm design leads to diminishing sensitivity. Specifically, Lemma 1 implies that when $\sum_{t=1}^{\infty}\frac{\Delta_t^i}{\nu_t^i}\le\epsilon^i$ is satisfied (where Δ_t^i is the sensitivity and ν_t^i is the parameter of DP-noise variances), agent i's implementation of an iterative algorithm is ϵ^i-locally differentially private in the infinite time horizon. According to Eq. (27), our algorithm design ensures that the sensitivity Δ_t^i (on the order of $\mathcal{O}(t^{-(1+v)})$) decays faster than the DP-noise variance ν_t^i (on the order of $\mathcal{O}(t^{-\varsigma^i})$). More specifically, our design ensures $\sum_{t=1}^{\infty}\frac{\Delta_t^i}{\nu_t^i}\le\sum_{t=1}^{\infty}\mathcal{O}(t^{-(1+v-\varsigma^i)})<\infty$ by requiring the design parameters to satisfy $1+v-\varsigma^i>1$. Therefore, we can ensure that the cumulative privacy budget is always finite.

In fact, our algorithm's achievement of both ϵ^i-LDP and accurate convergence does not come for free, but instead, incurs expense in convergence speed. We use the convergence speed and the cumulative privacy budget under a nonconvex $F(x)$ as an example to quantify this tradeoff:

Corollary 1. *For any given cumulative privacy budget $\epsilon^i>0$, $i\in[m]$, the convergence speed of Algorithm 1 is $\mathcal{O}\left(\frac{T^{-(1-v)}}{\min_{i\in[m]}\{(\epsilon^i)^2\}}\right)$.*

Proof. See Appendix D. □

Corollary 1 implies that a stronger privacy protection (corresponding to a smaller cumulative privacy budget ϵ^i) leads to a lower convergence speed.

Remark 7. Compared with the commonly used centralized DP framework which only allows one agent to change its data in the adjacency definition, our local model of DP allows all agents to change their data in the adjacency definition, and hence, has a sensitivity that is larger than the one in the conventional centralized DP framework. Hence, when adopted to the conventional centralized DP framework, the reduced sensitivity implies a reduced noise level. This reduction in needed noise leads to an increased convergence speed for Algorithm 1, as can be seen from the derivation of Eq. (98).

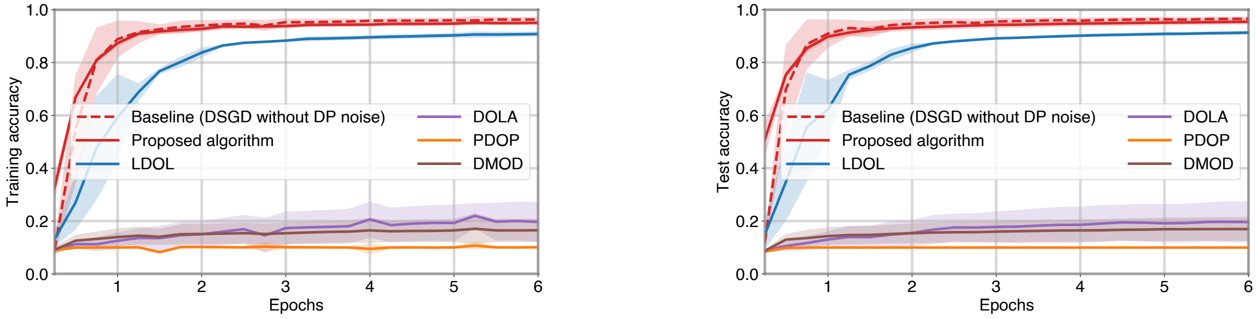
VI. Numerical Experiments

In this section, we performed machine-learning experiments to compare Algorithm 1 with the locally differentially private decentralized online learning algorithm (LDOL) in [47] (i.e., algorithm (17)). We also compared Algorithm 1 with other DP solutions for decentralized learning/optimization, including the decentralized online learning algorithm (DOLA) in [35], the decentralized offline optimization algorithm (PDOP) in [24] (which uses exponentially decaying stepsizes and DP noises to ensure a finite cumulative privacy budget), and the decentralized online mirror descent algorithm (DMOD) in [39]. For DOLA, PDOP, and DMOD, we set their privacy budgets equal to the maximum privacy budget (corresponding to the weakest protection) across all agents in our Algorithm 1. Moreover, we allowed DOLA, PDOP, DMOD, and LDOL to adopt the same gradient-computation strategy as ours in the "mushrooms" and the "MNIST" experiments for comparison. In all experiments, we used decentralized stochastic gradient descent (DSGD) in [9] without DP noises as a baseline for comparison. In addition, we considered ten agents connected in a circle, where each agent can only communicate with its two immediate neighbors. For the matrix W, we set $w_{ij}=0.3$ if agents i and j are neighbors, and $w_{ij}=0$ otherwise.

A. Logistic regression using the "mushrooms" dataset

In the first experiment, we evaluated the effectiveness of Algorithm 1 using a logistic regression classification task on the "mushrooms" dataset [68]. In this case, the loss function in problem (1) is given by $l(\theta,\xi^i)=\frac{1}{N^i}\sum_{s=1}^{N^i}(1-b_s^i)(a_s^i)^T\theta-\log(s((a_s^i)^T\theta))+\frac{r^i}{2}\|\theta\|^2$, where N^i represents the number of samples per iteration and r^i denotes a positive regularization parameter that is inversely proportional to N^i. $s(a)$ is the sigmoid function defined as $s(a)=\frac{1}{1+e^{-a}}$.

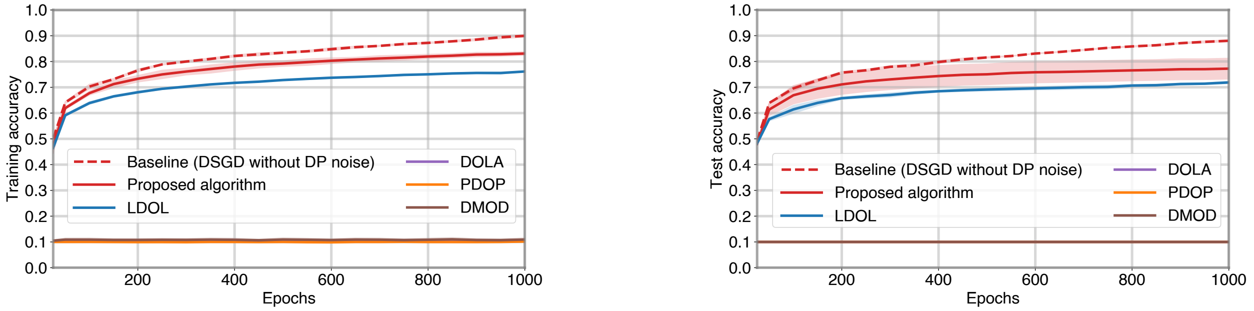
In each iteration, we randomly selected 20 samples and decentralized them to 10 agents. In each iteration, the DP noise variance and the stepsize were configured as $\nu_t^i=\frac{0.1}{(t+1)^{\varsigma^i}}$ with $\varsigma^i=0.5+0.01i$ and $\lambda_t=\frac{1}{(t+1)^{0.71}}$, respectively. The optimal solution θ^* was obtained using a noise-free and centralized gradient descent algorithm. In our comparison, we used the same stepsizes for the baseline (i.e., DSGD without DP noises) and LDOL, where the weakening factor was set as $\gamma_t=\frac{1}{(t+1)^{0.7}}$, in line with the guidelines provided in [47]. For other algorithms, we selected near-optimal stepsizes, ensuring that doubling stepsizes would lead to non-convergent behaviors.



(a) Training accuracy

(b) Test accuracy

Fig. 2. Comparison of CNN classification results by using the "MNIST" dataset.



(a) Training accuracy

(b) Test accuracy

Fig. 3. Comparison of CNN classification results by using the "CIFAR-10" dataset.

Fig. 1-(a) shows the evolution of average tracking errors, while Fig. 1-(b) depicts the evolution of average objective function values. These results confirm that Algorithm 1 has better optimization accuracy and convergence speed compared with existing results. Moreover, our Algorithm 1 only slightly reduces the convergence speed compared with the baseline (i.e., DSGD without DP noises), implying the robustness of our algorithm to DP noises.

B. Neural-network training using the "MNIST" dataset

In our second experiment, we executed decentralized online training of a convolutional neural network (CNN) using the "MNIST" dataset [69]. Each agent trained 32 randomly selected images per iteration. The DP noise was injected with parameters $\nu_t^i = \frac{0.01\times\sqrt{2}}{(t+1)^{\varsigma^i}}$, where $\varsigma^i = 0.5+0.01i$. The stepsize was set as $\lambda_t = \frac{1}{(t+1)^v}$ with $v = 0.68$.

We compared Algorithm 1 with DOLA, PDOP, DMOD, and LDOL under the same CNN model. The baseline (i.e., DSGD without DP noises) used the same stepsize as our algorithm, while DOLA, PDOP, and DMOD followed their recommended default stepsizes in [35], [24], and [39], respectively. The LDOL algorithm in [47] used the slowest allowable stepsize $\lambda_t = \frac{1}{(t+1)^{0.71}}$ to satisfy the condition $v > u$, where $u = 0.7$ is from its weakening factor $\gamma_t = \frac{1}{(t+1)^u}$, in line with recommendations from [47].

Fig. 2 reveals that DOLA and DMOD are incapable of effectively training the CNN model under DP noise injections, even when they use the same gradient-computation strategy as ours. This ineffectiveness is due to their use of identical decaying rates for stepsizes and DP-noise variances. Specifically, under this constraint, rapid decay in stepsizes leads to a low convergence speed, even with the DP noise decaying rapidly. Conversely, slow decay in stepsizes can accelerate convergence, but the corresponding slow decay in DP noises also damages convergence. Similarly, PDOP fails to effectively train the CNN model under DP noise injections because it employs exponentially decaying stepsizes and DP-noise variances to ensure a finite cumulative privacy budget. Such rapidly decaying stepsizes prevent the model from adapting effectively to the data, leading to poor training performance.

C. Neural-network training using the "CIFAR-10" dataset

In the third experiment, we evaluated Algorithm 1 by training a CNN model using the "CIFAR-10" dataset [70], which provides a greater diversity and complexity than the "MNIST" dataset. The DP noise variance was set as $\nu_t^i = \frac{0.05\times\sqrt{2}}{(t+1)^{\varsigma^i}}$, where $\varsigma^i = 0.5 + 0.01i$. The stepsize was configured as $\lambda_t = \frac{1}{(t+1)^v}$ with $v = 0.7$. All other parameters are the same as those employed in the previous experiment on the "MNIST" dataset.

The results are summarized in Fig. 3, which once again confirms the advantage of our proposed algorithm over existing counterparts.

VII. Conclusions

In this study, we have introduced a decentralized online optimization algorithm that ensures both rigorous local differential privacy and optimization accuracy for decentralized stochastic optimization. More specifically, we have proved that our algorithm guarantees convergence in mean square to the optimal solution to the stochastic optimization problem under streaming data, which differs from most existing results that only characterize the regret function, an indirect measure of optimization accuracy. Simultaneously, we have proved that our algorithm ensures a finite privacy budget over an infinite time horizon. This stands in stark contrast to most existing DP solutions for decentralized optimization that have to sacrifice optimization accuracy for privacy. In addition, our algorithm does not use any decaying factors to gradually decay inter-agent coupling strength, which is crucial in existing DP solutions to guarantee optimality and a finite privacy budget, but which unavoidably compromises the speed of convergence. Experimental results on benchmark datasets have been provided to validate the advantages of our algorithm over existing counterparts.

Appendix

For notational simplicity, we add a bar over a letter to denote the average of all agents, e.g., $\bar{\theta}_t = \frac{1}{m}\sum_{i=1}^{m}\theta_t^i$, and use bold font to represent stacked vectors of all agents, e.g., $\boldsymbol{\theta}_t = \text{col}\{\theta_t^1, \cdots, \theta_t^m\}$. For the convenience of derivation, we define $\vartheta_t^{wi} \triangleq \sum_{j\in\mathcal{N}_i} w_{ij}\vartheta_t^j$, $\tilde{\boldsymbol{\theta}}_t \triangleq \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*$, $\check{\boldsymbol{\theta}}_t \triangleq \boldsymbol{\theta}_t - \bar{\boldsymbol{\theta}}_t$, $\check{\boldsymbol{\vartheta}}_t^w \triangleq \boldsymbol{\vartheta}_t^w - \bar{\boldsymbol{\vartheta}}_t^w$, $\nabla \boldsymbol{f}_t(\boldsymbol{\theta}_t) \triangleq \text{col}\{\nabla f_t^1(\theta_t^1), \cdots, \nabla f_t^m(\theta_t^m)\}$, $F_t(\boldsymbol{\theta}_t) \triangleq \frac{1}{m}\sum_{i=1}^{m}\nabla f_t^i(\theta_t^i)$, $\nabla \boldsymbol{f}(\boldsymbol{\theta}_t) \triangleq \text{col}\{\nabla f_1(\theta_t^1), \cdots, \nabla f_m(\theta_t^m)\}$, and $F(\boldsymbol{\theta}_t) \triangleq \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\theta_t^i)$.

A. *Proof of Theorem 1*

We write Line 6 in Algorithm 1 in a compact form $\boldsymbol{\theta}_{t+1} = (I_{mn} + W\otimes I_n)\boldsymbol{\theta}_t + \boldsymbol{\vartheta}_t^w - \lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)$, which implies

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t^*\|^2 \leq \|(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \boldsymbol{\vartheta}_t^w - \lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2 \\ + 2\big\langle\tilde{\boldsymbol{\theta}}_t, (W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \boldsymbol{\vartheta}_t^w - \lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\big\rangle + \|\tilde{\boldsymbol{\theta}}_t\|^2,\end{aligned} \tag{29}$$

where in the derivation we have used $W\mathbf{1}_m = \mathbf{0}_m$.

Assumption 4 ensures $\mathbb{E}[\|\boldsymbol{\vartheta}_t^w\|^2] = \|\boldsymbol{\sigma}_t\|^2$, which further implies that the first term on the right hand side of (29) satisfies

$$\begin{aligned}&\mathbb{E}[\|(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \boldsymbol{\vartheta}_t^w - \lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2] \\ &\leq 6\|\boldsymbol{\sigma}_t\|^2 + \tfrac{3}{2}\mathbb{E}[\|(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t\|^2] + 6\lambda_t^2\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2].\end{aligned} \tag{30}$$

Based on the definition of $\boldsymbol{\vartheta}_t^w$ and the fact $\mathbb{E}[\vartheta_t^j] = 0$ from Assumption 4, we have $\mathbb{E}[\boldsymbol{\vartheta}_t^w] = \mathbf{0}$. Furthermore, since $\tilde{\boldsymbol{\theta}}_t$ and $\boldsymbol{\vartheta}_t^w$ are independent, we have $\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \boldsymbol{\vartheta}_t^w\rangle] = 0$, which further implies the following equality:

$$\begin{aligned}&2\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, (W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \boldsymbol{\vartheta}_t^w - \lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\rangle] \\ &= 2\mathbb{E}[\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t - \langle\tilde{\boldsymbol{\theta}}_t, \lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\rangle].\end{aligned} \tag{31}$$

Substituting (30) and (31) into (29) leads to

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t^*\|^2] \leq \mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t\|^2] + 2\mathbb{E}[\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t] \\ - 2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\rangle] + \tfrac{3}{2}\mathbb{E}[\|(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t\|^2] \\ + 6\|\boldsymbol{\sigma}_t\|^2 + 6\lambda_t^2\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2].\end{aligned} \tag{32}$$

To further characterize (32), we decompose the third term on the right hand side of (32) as follows:

$$\begin{aligned}2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\rangle] = 2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}(\bar{\boldsymbol{\theta}}_t)\rangle] \\ + 2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}(\boldsymbol{\theta}_t) - \nabla\boldsymbol{f}(\bar{\boldsymbol{\theta}}_t)\rangle] \\ + 2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t) - \nabla\boldsymbol{f}(\boldsymbol{\theta}_t)\rangle].\end{aligned} \tag{33}$$

Next, we analyze each item on the right hand side of (33):

(a) Using the strong convexity of $F(\theta)$, the definition $m\bar{\theta}_t = \sum_{i=1}^{m}\theta_t^i$, and the relation $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$, we obtain

$$\begin{aligned}&2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}(\bar{\boldsymbol{\theta}}_t)\rangle] \\ &\geq 2\lambda_t\mathbb{E}[mF(\bar{\theta}_t) - mF(\theta_t^*) + \frac{\mu}{2}m\|\bar{\theta}_t - \theta_t^*\|^2] \\ &\geq 2\lambda_t\mathbb{E}[mF(\bar{\theta}_t) - mF(\theta_t^*) + \frac{\mu}{4}\|\tilde{\boldsymbol{\theta}}_t\|^2 - \frac{\mu}{2}\|\check{\boldsymbol{\theta}}_t\|^2].\end{aligned} \tag{34}$$

(b) The Young's inequality and Assumption 2-(iii) imply

$$\begin{aligned}&2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}(\boldsymbol{\theta}_t) - \nabla\boldsymbol{f}(\bar{\boldsymbol{\theta}}_t)\rangle] \\ &\geq -\lambda_t\left(\frac{\mu}{8}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t\|^2] + \frac{8L^2}{\mu}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2]\right).\end{aligned}$$

(c) The Young's inequality also implies

$$\begin{aligned}&2\lambda_t\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t) - \nabla\boldsymbol{f}(\boldsymbol{\theta}_t)\rangle \\ &\geq -\frac{\lambda_t\mu}{8}\|\tilde{\boldsymbol{\theta}}_t\|^2 - \frac{8\lambda_t}{\mu}\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t) - \nabla\boldsymbol{f}(\boldsymbol{\theta}_t)\|^2.\end{aligned} \tag{35}$$

Based on the definitions of $\boldsymbol{f}_t$ and $\boldsymbol{f}$ and Assumption 2-(ii), the last term on the right hand side of (35) satisfies

$$\begin{aligned}&\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t) - \nabla\boldsymbol{f}(\boldsymbol{\theta}_t)\|^2] \\ &= \sum_{i=1}^{m}\mathbb{E}\left[\left\|\frac{1}{t+1}\sum_{k=0}^{t}\nabla l(\theta_t^i, \xi_k^i) - \mathbb{E}[\nabla l(\theta_t^i, \xi^i)]\right\|^2\right] \\ &\leq \frac{1}{(t+1)^2}\sum_{i=1}^{m}\sum_{k=0}^{t}\mathbb{E}[\|\nabla l(\theta_t^i, \xi_k^i) - \nabla f_i(\theta_t^i)\|^2] \leq \frac{\kappa^2 m}{t+1}.\end{aligned} \tag{36}$$

Incorporating (34)-(36) into (33), one yields

$$\begin{aligned}2\lambda_t\mathbb{E}[\langle\tilde{\boldsymbol{\theta}}_t, \nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\rangle] \geq 2m\lambda_t\mathbb{E}[(F(\bar{\theta}_t) - F(\theta_t^*))] \\ + \frac{\lambda_t\mu}{4}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t\|^2] - \lambda_t\left(\mu + \frac{8L^2}{\mu}\right)\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2] - \frac{8m\kappa^2\lambda_t}{\mu(t+1)}.\end{aligned} \tag{37}$$

We incorporate (37) into (32) to obtain an upper bound on $\mathbb{E}\left[\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t^*\|^2\right]$. Then, by using the relation $\|\tilde{\boldsymbol{\theta}}_{t+1}\|^2 \leq a_t\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t^*\|^2 + b_t\|\boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^*\|^2$ for all $a_t, b_t > 1$ and $(a_t - 1)(b_t - 1) = 1$, we can obtain the following inequality:

$$\begin{aligned}\mathbb{E}\big[\|\tilde{\boldsymbol{\theta}}_{t+1}\|^2\big] \leq a_t\mathbb{E}\Big[&\|\tilde{\boldsymbol{\theta}}_t\|^2 + 2\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t \\ &- 2m\lambda_t\big(F(\bar{\theta}_t) - F(\theta_t^*)\big) - \frac{\lambda_t\mu}{4}\|\tilde{\boldsymbol{\theta}}_t\|^2 + \lambda_t\Big(\mu + \frac{8L^2}{\mu}\Big)\|\check{\boldsymbol{\theta}}_t\|^2 \\ &+ \frac{3}{2}\|(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t\|^2 + 6\|\boldsymbol{\sigma}_t\|^2 + 6\lambda_t^2\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2\Big] \\ &+ \frac{8m\lambda_t\kappa^2 a_t}{\mu(t+1)} + b_t\mathbb{E}\big[\|\boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^*\|^2\big].\end{aligned} \tag{38}$$

Next, to further simplify inequality (38), we first prove that the sum of following three terms in (38) is negative:

$$2\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \frac{3}{2}\|(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t\|^2 + \lambda_t\Big(\mu + \frac{8L^2}{\mu}\Big)\|\check{\boldsymbol{\theta}}_t\|^2 \leq 0. \tag{39}$$

Since the eigenvalues of W satisfy $\delta_i \in (-1, 0]$, we have $\delta_i + \delta_i^2 \leq 0$ for all $i \in [m]$, which further implies

$$\frac{3}{2}\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \frac{3}{2}\|(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t\|^2 \leq 0. \tag{40}$$

Then, we prove $\frac{1}{2}\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \lambda_t(\mu + \frac{8L^2}{\mu})\|\check{\boldsymbol{\theta}}_t\|^2 \leq 0$.

Using the definitions of $\check{\boldsymbol{\theta}}_t$ and $\tilde{\boldsymbol{\theta}}_t$, one obtains

$$\begin{aligned}\check{\boldsymbol{\theta}}_t &= \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* - (\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t^*)\\ &= \tilde{\boldsymbol{\theta}}_t - \Big(\Big(\frac{\mathbf{1}_m\mathbf{1}_m^T}{m}\otimes I_n\Big)\boldsymbol{\theta}_t - \mathbf{1}_m\Big(\frac{\mathbf{1}_m^T\mathbf{1}_m}{m}\Big)\otimes\theta_t^*\Big).\end{aligned}$$

The relationship $\mathbf{1}_m(\frac{\mathbf{1}_m^T\mathbf{1}_m}{m})\otimes\theta_t^* = (\frac{\mathbf{1}_m\mathbf{1}_m^T}{m}\otimes I_n)(\mathbf{1}_m\otimes\theta_t^*)$ implies $\check{\boldsymbol{\theta}}_t = (I_{mn} - \frac{\mathbf{1}_m\mathbf{1}_m^T}{m}\otimes I_n)\tilde{\boldsymbol{\theta}}_t$, which further leads to

$$\tilde{\boldsymbol{\theta}}_t = \check{\boldsymbol{\theta}}_t + \Big(\frac{\mathbf{1}_m\mathbf{1}_m^T}{m}\otimes I_n\Big)\tilde{\boldsymbol{\theta}}_t.$$

By using the relationships $\check{\boldsymbol{\theta}}_t^T(W\otimes I_n)\check{\boldsymbol{\theta}}_t \le \delta_2\|\check{\boldsymbol{\theta}}_t\|^2$, $\mathbf{1}^T W = \mathbf{0}^T$, and $W\mathbf{1} = \mathbf{0}$, we obtain

$$\frac{1}{2}\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t = \frac{1}{2}\check{\boldsymbol{\theta}}_t^T(W\otimes I_n)\check{\boldsymbol{\theta}}_t \le \frac{1}{2}\delta_2\|\check{\boldsymbol{\theta}}_t\|^2.$$

Given that the stepsize satisfies $\lambda_t \le \lambda_0 \le \frac{-\delta_2\mu}{2(\mu^2+8L^2)}$ from the statement of Theorem 1, we can obtain

$$\frac{1}{2}\tilde{\boldsymbol{\theta}}_t^T(W\otimes I_n)\tilde{\boldsymbol{\theta}}_t + \lambda_t\Big(\mu + \frac{8L^2}{\mu}\Big)\|\check{\boldsymbol{\theta}}_t\|^2 \le 0. \qquad (41)$$

Combining (40) and (41) leads to inequality (39).

Hence, by omitting the three terms in (39) from (38), we can simplify (38) as

$$\begin{aligned}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_{t+1}\|^2] &\le -2ma_t\lambda_t\mathbb{E}[F(\bar{\theta}_t) - F(\theta_t^*)]\\ &+ a_t\left(1 - \frac{\mu\lambda_t}{4}\right)\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t\|^2] + 6a_t\|\boldsymbol{\sigma}_t\|^2 + \frac{8m\lambda_t\kappa^2 a_t}{\mu(t+1)}\\ &+ 6a_t\lambda_t^2\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2] + b_t\mathbb{E}[\|\boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^*\|^2]. \qquad (42)\end{aligned}$$

We proceed to characterize the first term on the right hand side of (42). Using the relation $\mathbb{E}[F(\bar{\theta}_t) - F(\theta^*)] \ge 0$ yields

$$\begin{aligned}\mathbb{E}[F(\bar{\theta}_t) - F(\theta_t^*)] &= \mathbb{E}[F(\bar{\theta}_t) - F(\theta^*)] + \mathbb{E}[F(\theta^*) - F(\theta_t^*)]\\ &\ge \mathbb{E}[F(\theta^*) - F(\theta_t^*)] \ge \frac{-2\kappa^2}{\mu(t+1)}, \qquad (43)\end{aligned}$$

where we have used $\mathbb{E}[\|\theta_t^* - \theta^*\|] \le \frac{2\kappa}{\mu\sqrt{t+1}}$ from Lemma 2 and Eq. (10) in the last inequality.

By incorporating (43) into (42) and letting $a_t = 1 + \frac{\mu\lambda_t}{8}$ and $b_t = 1 + \frac{8}{\lambda_t\mu}$, we can rewrite inequality (42) as follows:

$$\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_{t+1}\|^2] \le \left(1 - \frac{\mu\lambda_t}{8}\right)\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t\|^2] + \Phi_t, \qquad (44)$$

where the term Φ_t is given by

$$\begin{aligned}\Phi_t &= \left(1 + \frac{8}{\lambda_t\mu}\right)\mathbb{E}[\|\boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^*\|^2] + \left(1 + \frac{\mu\lambda_t}{8}\right)\frac{12m\kappa^2\lambda_t}{\mu(t+1)}\\ &+ 6\left(1 + \frac{\mu\lambda_t}{8}\right)\left(\lambda_t^2\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2] + \|\boldsymbol{\sigma}_t\|^2\right). \qquad (45)\end{aligned}$$

By iterating (44) from 0 to t, one yields

$$\begin{aligned}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_{t+1}\|^2] &\le \prod_{p=0}^{t}\left(1 - \frac{\mu\lambda_p}{8}\right)\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_0\|^2]\\ &+ \sum_{p=1}^{t}\prod_{q=p}^{t}\left(1 - \frac{\mu\lambda_q}{8}\right)\Phi_{p-1} + \Phi_t. \qquad (46)\end{aligned}$$

Since $\ln(1-u) \le -u$ holds for all $u \in (0,1)$, we always have $\prod_{p=0}^{t}(1 - \frac{\mu\lambda_p}{8}) \le e^{-\frac{1}{8}\mu\sum_{p=0}^{t}\lambda_p}$. Hence, inequality (46) can be rewritten as follows:

$$\begin{aligned}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_{t+1}\|^2] &\le e^{-\frac{1}{8}\mu\sum_{p=0}^{t}\lambda_p}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_0\|^2]\\ &+ \sum_{p=1}^{t}\Phi_{p-1}e^{-\frac{1}{8}\mu\sum_{q=p}^{t}\lambda_q} + \Phi_t. \qquad (47)\end{aligned}$$

We now analyze the first term on the right hand side of (47).

Since $\frac{\lambda_0}{(p+1)^v} \ge \frac{\lambda_0}{(t+1)^v}$ holds for all $t \ge p$ and $(t+1)^v \le 2^v t^v$ holds for all $t > 0$, we have

$$\sum_{p=0}^{t}\lambda_p = \sum_{p=0}^{t}\frac{\lambda_0}{(p+1)^v} \ge \frac{\lambda_0}{(t+1)^v}(t+1) \ge \frac{\lambda_0}{2^v t^{v-1}}, \qquad (48)$$

which further implies $e^{\frac{\mu}{8}\sum_{p=0}^{t}\lambda_p} \ge e^{\frac{\mu}{8}\frac{\lambda_0}{2^v t^{v-1}}}$. Using Taylor expansion $e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!}$, we have that for any $n_0 \in \mathbb{N}^+$, $e^x \ge \frac{x^{n_0}}{n_0!}$ holds. By setting $n_0 = \lceil\frac{1}{1-v}\rceil$, we have $(1-v)n_0 \ge 1$, which further implies

$$e^{\frac{1}{8}\mu\sum_{p=0}^{t}\lambda_p} \ge \frac{1}{(\frac{1}{1-v}+1)!}\Big(\frac{\mu\lambda_0}{8\times 2^v}\Big)^{\frac{1}{1-v}}t. \qquad (49)$$

By substituting (49) into the first term on the right hand side of (47), we arrive at

$$e^{-\frac{\mu}{8}\sum_{p=0}^{t}\lambda_p}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_0\|^2] \le c_1 t^{-1}, \qquad (50)$$

where $c_1 = (\frac{1}{1-v}+1)!\big(\frac{\mu\lambda_0}{8\times 2^v}\big)^{\frac{1}{v-1}}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_0\|^2]$.

We proceed to analyze the second and third terms on the right hand side of (47). By introducing a variable $\alpha\in(v,1)$, we have $e^{-\frac{\mu}{8}\sum_{q=p}^{t}\lambda_q} \le e^{-\frac{\mu}{8}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q}$ for all $p\in[1,\lceil t-t^\alpha\rceil]$ and $e^{-\frac{\mu}{8}\sum_{q=\lceil t-t^\alpha\rceil+1}^{t}\lambda_q} < 1$. Therefore, the second and third terms on the right hand side of (47) satisfy

$$\begin{aligned}&\sum_{p=1}^{t}\Phi_{p-1}e^{-\frac{1}{8}\mu\sum_{q=p}^{t}\lambda_q} + \Phi_t\\ &< \sum_{p=1}^{\lceil t-t^\alpha\rceil}\Phi_{p-1}e^{-\frac{1}{8}\mu\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} + \sum_{p=\lceil t-t^\alpha\rceil+1}^{t}\Phi_{p-1} + \Phi_t\\ &= \sum_{p=0}^{\lfloor t-t^\alpha\rfloor}\Phi_p e^{-\frac{\mu}{8}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} + \sum_{p=\lceil t-t^\alpha\rceil}^{t}\Phi_p. \qquad (51)\end{aligned}$$

To proceed, we need to compute an upper bound on Φ_t. To this end, we first establish the following relations:

(a) By using inequality (12), we have $\mathbb{E}\big[\|\boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^*\|^2\big] \le \frac{c_{01}}{(t+1)^2}$ with $c_{01} = 16m(\kappa^2 + D^2)(\frac{2}{\mu^2} + \frac{1}{L^2})$.

(b) By utilizing the definitions $\varsigma = \min_{i\in[m]}\{\varsigma^i\}$ and $\sigma^+ = \max_{i\in[m]}\{\sigma_0^i\}$, we have $\|\boldsymbol{\sigma}_t\|^2 \le \frac{m(\sigma^+)^2}{(t+1)^{2\varsigma}}$.

(c) Using the definition $\nabla f_i(\theta_t^*) = \mathbb{E}[\nabla l(\theta_t^*, \xi^i)]$, Assumption 1, and Assumption 2-(ii), we obtain

$$\begin{aligned}&\mathbb{E}[\|\nabla l(\theta_t^*, \xi^i)\|^2]\\ &\le 2\mathbb{E}[\|\nabla l(\theta_t^*, \xi^i) - \nabla f_i(\theta_t^*)\|^2] + 2\mathbb{E}[\|\nabla f_i(\theta_t^*)\|^2]\\ &\le 2(\kappa^2 + D^2). \qquad (52)\end{aligned}$$

Given $\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2] = \sum_{i=1}^{m}\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla l(\theta_t^i, \xi_k^i)\|^2]$, we have $\lambda_t^2\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|^2] \le 2m\lambda_t^2(\kappa^2 + D^2)$.

Substituting the above results in (a)-(c) into (45) and using the relation $\lambda_t \leq \lambda_0$, we obtain

$$\begin{aligned}\Phi_t \leq\; & c_{01}(t+1)^{-2} + c_{02}(t+1)^{-2+v} + c_{03}(t+1)^{-2\varsigma} \\ & + c_{04}(t+1)^{-2v} + c_{05}(t+1)^{-1-v},\end{aligned} \tag{53}$$

where $c_{0i},\ i = 1,2,3,4,5$ are given by $c_{01} = 16m(\kappa^2 + D^2)(\frac{2}{\mu^2} + \frac{1}{L^2})$, $c_{02} = \frac{8}{\lambda_0\mu}c_{01}$, $c_{03} = 6m(\sigma^+)^2(1+\frac{\lambda_0\mu}{8})$, $c_{04} = 12m(\kappa^2 + D^2)(1+\frac{\lambda_0\mu}{8})\lambda_0^2$, and $c_{05} = (1+\frac{\mu\lambda_0}{8})\frac{12m\lambda_0\kappa^2}{\mu}$, respectively.

Using inequality (53), we can now analyze the first term on the right hand side of (51). To this end, we first characterize the term $e^{-\frac{\mu}{8}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q}$ in (51).

Given that the inequality $\frac{1}{(q+1)^v} \geq \frac{1}{(t+1)^v}$ is valid for all $q \in [\lceil t-t^\alpha\rceil, t]$, we have

$$\begin{aligned}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q = \sum_{q=\lceil t-t^\alpha\rceil}^{t}\frac{\lambda_0}{(q+1)^v} &\geq \frac{\lambda_0}{(t+1)^v}(t-\lceil t-t^\alpha\rceil+1), \\ &\geq \frac{\lambda_0 t^\alpha}{(t+1)^v} \geq \frac{\lambda_0 t^{\alpha-v}}{2^v},\end{aligned} \tag{54}$$

where we have used the relations $\lceil t-t^\alpha\rceil \leq t-t^\alpha+1$ and $(t+1)^v \leq 2^v t^v$ in the derivation. Inequality (54) further implies $e^{\frac{\mu}{8}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} \geq e^{\frac{\mu}{8}\frac{\lambda_0 t^{\alpha-v}}{2^v}}$. Using an argument similar to the derivation of (49), we define $n_0 = \lceil\frac{1}{\alpha-v}\rceil$ (i.e., $(\alpha-v)n_0 \geq 1$) for the Taylor expansion and obtain the following inequality:

$$e^{\frac{\mu}{8}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} \geq \frac{1}{(\frac{1}{\alpha-v}+1)!}\Big(\frac{\mu\lambda_0}{8\times 2^v}\Big)^{\frac{1}{\alpha-v}} t. \tag{55}$$

Incorporating (55) into the first term on the right hand side of (51) leads to

$$\textstyle\sum_{p=0}^{\lfloor t-t^\alpha\rfloor}\Phi_p e^{-\frac{\mu}{8}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} < \Big(\Phi_0 + \sum_{p=1}^{\infty}\Phi_p\Big)c' t^{-1}, \tag{56}$$

where the constant c' is given by $c' = (\frac{\alpha-v+1}{\alpha-v})!\left(\frac{\mu\lambda_0}{8\times 2^v}\right)^{\frac{1}{v-\alpha}}$.

By setting $t = 0$ in (53), we can derive $\Phi_0 = \sum_{i=1}^{5}c_{0i}$. We further compute an upper bound on $\sum_{p=1}^{\infty}\Phi_p$ in (56). By using (53), one yields

$$\begin{aligned}\sum_{p=1}^{\infty}\Phi_p \leq & \sum_{p=1}^{\infty}\frac{c_{01}}{(p+1)^2} + \sum_{p=1}^{\infty}\frac{c_{02}}{(p+1)^{2-v}} + \sum_{p=1}^{\infty}\frac{c_{03}}{(p+1)^{2\varsigma}} \\ & + \sum_{p=1}^{\infty}\frac{c_{04}}{(p+1)^{2v}} + \sum_{p=1}^{\infty}\frac{c_{05}}{(p+1)^{1+v}}.\end{aligned} \tag{57}$$

All items on the right hand side of (57) can be simplified. For example, the third item can be bounded as follows:

$$\sum_{p=1}^{\infty}(p+1)^{-2\varsigma} \leq \int_1^\infty \frac{1}{x^{2\varsigma}}dx \leq \frac{1}{2\varsigma-1}. \tag{58}$$

Applying the same argument to the other items on the right hand side of (57) yields

$$\begin{cases}\displaystyle\sum_{p=1}^{\infty}\Big(\frac{1}{(p+1)^2}+\frac{1}{(p+1)^{2-v}}\Big) \leq 1+\frac{1}{1-v}, \\ \displaystyle\sum_{p=1}^{\infty}\Big(\frac{1}{(p+1)^{2v}}+\frac{1}{(p+1)^{1+v}}\Big) \leq \frac{1}{2v-1}+\frac{1}{v}.\end{cases} \tag{59}$$

Substituting inequalities (58)-(59) into (57), we have

$$\sum_{p=1}^{\infty}\Phi_p \leq c_{01} + \frac{c_{02}}{1-v} + \frac{c_{03}}{2\varsigma-1} + \frac{c_{04}}{2v-1} + \frac{c_{05}}{v}. \tag{60}$$

Incorporating $\Phi_0 = \sum_{i=1}^{5}c_{0i}$ and (60) into (56) yields that the first term on the right hand side of (51) is bounded by

$$\sum_{p=0}^{\lfloor t-t^\alpha\rfloor}\Phi_p e^{-\frac{1}{8}\mu\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} < c_2 t^{-1}, \tag{61}$$

where c_2 is given by $c_2 = (\sum_{i=1}^{5}c_{0i} + c_{01} + \frac{c_{02}}{1-v} + \frac{c_{03}}{2\varsigma-1} + \frac{c_{04}}{2v-1} + \frac{c_{05}}{v})c'$ with $c_{0i}, i = 1,2,3,4,5$ given in (53) and c' given in (56).

We proceed to characterize the second term on the right hand side of (51). By using (53), we obtain

$$\begin{aligned}\sum_{p=\lceil t-t^\alpha\rceil}^{t}\Phi_p \leq \sum_{p=\lceil t-t^\alpha\rceil}^{t} & \Big(\frac{c_{01}}{(p+1)^2} + \frac{c_{02}}{(p+1)^{2-v}} \\ & + \frac{c_{03}}{(p+1)^{2\varsigma}} + \frac{c_{04}}{(p+1)^{2v}} + \frac{c_{05}}{(p+1)^{1+v}}\Big).\end{aligned} \tag{62}$$

All items on the right hand side of (62) can be simplified. We compute an upper bound on the third item as an example.

Given $\frac{1}{(p+1)^{2\varsigma}} \leq \frac{1}{(\lceil t-t^\alpha\rceil+1)^{2\varsigma}}$ valid for all $p \in [\lceil t-t^\alpha\rceil, t]$, we have

$$\sum_{p=\lceil t-t^\alpha\rceil}^{t}\frac{1}{(p+1)^{2\varsigma}} \leq \frac{1}{(\lceil t-t^\alpha\rceil+1)^{2\varsigma}}(t-\lceil t-t^\alpha\rceil+1). \tag{63}$$

To simplify the expression on the right hand side of (63), we first prove the following inequality:

$$\lceil t-t^\alpha\rceil + 1 \geq t(1-\alpha),\ \forall t \in \mathbb{N}. \tag{64}$$

Considering the relation $\lceil t-t^\alpha\rceil + 1 - t(1-\alpha) \geq t - t^\alpha + 1 - t(1-\alpha) = \alpha t - t^\alpha + 1$, we construct a function $f(t) = \alpha t - t^\alpha + 1 : \mathbb{N} \to \mathbb{R}$. The derivative of this function is $f'(t) = \alpha - \alpha t^{\alpha-1}$. With $f(t=0) = 1$, $f(t=1) = \alpha$, and $f'(t) \geq 0$ for all $t \geq 1$, $f(t) \geq 0$ holds for all $t \in \mathbb{N}$. This implies that (64) always holds for all $t \in \mathbb{N}$. Using both (64) and the relation $t^\alpha + 1 \leq 2t^\alpha$, we can rewrite (63) as follows:

$$\sum_{p=\lceil t-t^\alpha\rceil}^{t}\frac{1}{(p+1)^{2\varsigma}} \leq \frac{t^\alpha+1}{t^{2\varsigma}(1-\alpha)^{2\varsigma}} \leq \frac{2t^{\alpha-2\varsigma}}{(1-\alpha)^{2\varsigma}}. \tag{65}$$

Applying the same argument to the other items on the right hand side of (62) yields

$$\begin{cases}\displaystyle\sum_{p=\lceil t-t^\alpha\rceil}^{t}\Big(\frac{1}{(p+1)^2}+\frac{1}{(p+1)^{2-v}}\Big) \leq \frac{2t^{\alpha-2}}{(1-\alpha)^2}+\frac{2t^{\alpha-2+v}}{(1-\alpha)^{2-v}}, \\ \displaystyle\sum_{p=\lceil t-t^\alpha\rceil}^{t}\Big(\frac{1}{(p+1)^{2v}}+\frac{1}{(p+1)^{1+v}}\Big) \leq \frac{2t^{\alpha-2v}}{(1-\alpha)^{2v}}+\frac{2t^{\alpha-v-1}}{(1-\alpha)^{1+v}}.\end{cases} \tag{66}$$

Substituting (65) and (66) into (62), we have that the second term on the right hand side of (51) is bounded by

$$\begin{aligned}\sum_{p=\lceil t-t^\alpha\rceil}^{t}\Phi_p \leq\; & c_3 t^{\alpha-2} + c_4 t^{\alpha-2+v} + c_5 t^{\alpha-2\varsigma} \\ & + c_6 t^{\alpha-2v} + c_7 t^{\alpha-v-1},\end{aligned} \tag{67}$$

with $c_3 = \frac{2c_{01}}{(1-\alpha)^2}$, $c_4 = \frac{2c_{02}}{(1-\alpha)^{2-v}}$, $c_5 = \frac{2c_{03}}{(1-\alpha)^{2\varsigma}}$, $c_6 = \frac{2c_{04}}{(1-\alpha)^{2v}}$, and $c_7 = \frac{2c_{05}}{(1-\alpha)^{v+1}}$.

We incorporate (61) and (67) into (51) and further substitute (50) and (51) into (47) to arrive at

$$\begin{aligned}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_{t+1}\|^2] &< (c_1+c_2)t^{-1} + c_3 t^{\alpha-2} + c_4 t^{\alpha-2+v} \\ &\quad + c_5 t^{\alpha-2\varsigma} + c_6 t^{\alpha-2v} + c_7 t^{\alpha-v-1},\ t>0,\end{aligned} \tag{68}$$

where c_1 is given in (50), c_2 is given in (61), and c_3 to c_7 are given in (67).

We further analyze an upper bound on $\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_1\|^2]$. By incorporating $\Phi_0 = \sum_{i=1}^{5} c_{0i}$ into (44), we derive $\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_1\|^2] \leq c_0$, with $c_0 = (1-\frac{\mu\lambda_0}{8})\mathbb{E}\left[\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^*\|^2\right] + \sum_{i=1}^{5} c_{0i}$. Combining this relationship with (68), we arrive at

$$\begin{aligned}\mathbb{E}[\|\tilde{\boldsymbol{\theta}}_t\|^2] &< (c_0+c_1+c_2)t^{-1} + c_3 t^{\alpha-2} + c_4 t^{\alpha-2+v} \\ &\quad + c_5 t^{\alpha-2\varsigma} + c_6 t^{\alpha-2v} + c_7 t^{\alpha-v-1} \leq \mathcal{O}(t^{-\beta}),\end{aligned} \tag{69}$$

where β is given by $\beta = \min\{2-v-\alpha, 2\varsigma-\alpha\}$ due to the relation $\min\{1, 2-\alpha, 2-v-\alpha, 2\varsigma-\alpha, 2v-\alpha, v+1-\alpha\} = \min\{2-v-\alpha, 2\varsigma-\alpha\}$. Since α can selected within the interval $(v,1)$, we let $\alpha = \frac{v+1}{2}$ and further obtain (13) in Theorem 1.

B. Proof of Theorem 3

To facilitate the derivation, we define $\hat{\theta}_t \triangleq \bar{\theta}_t - \theta^*$ and $\sigma_t \triangleq \frac{\sigma^+}{(t+1)^{\varsigma}}$ with $\sigma^+ = \max_{i\in[m]}\{\sigma^i\}$ and $\varsigma = \min_{i\in[m]}\{\varsigma^i\}$.

Recalling the definitions of $F_t(\boldsymbol{\theta}_t)$ and $\hat{\theta}_t$, we use Line 6 in Algorithm 1 and Assumption 4 to obtain

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_{t+1}\|^2] &= \mathbb{E}[\|\hat{\theta}_t\|^2] + \mathbb{E}[\|\bar{\theta}_t^w\|^2] + \lambda_t^2\mathbb{E}[\|\nabla F_t(\boldsymbol{\theta}_t)\|^2] \\ &\quad - 2\lambda_t\mathbb{E}[\langle\hat{\theta}_t, \nabla F_t(\boldsymbol{\theta}_t)\rangle].\end{aligned} \tag{70}$$

By using Assumption 1 and Assumption 2-(ii), we have that the third term on the right hand side of (70) satisfies

$$\begin{aligned}\mathbb{E}[\|\nabla F_t(\boldsymbol{\theta}_t)\|^2] &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\nabla f_t^i(\theta_t^i) - \nabla f_i(\theta_t^i) + \nabla f_i(\theta_t^i)\|^2] \\ &\leq \frac{\kappa^2}{t+1} + D^2.\end{aligned} \tag{71}$$

We proceed to characterize the last term on the right hand side of (70) by using the following decomposition:

$$\begin{aligned}2\mathbb{E}[\langle\hat{\theta}_t, \nabla F_t(\boldsymbol{\theta}_t)\rangle] &= 2\mathbb{E}[\langle\hat{\theta}_t, \nabla F(\boldsymbol{\theta}_t)\rangle] \\ &\quad + 2\mathbb{E}[\langle\hat{\theta}_t, \nabla F_t(\boldsymbol{\theta}_t) - \nabla F(\boldsymbol{\theta}_t)\rangle].\end{aligned} \tag{72}$$

The first term on the right hand side of (72) satisfies

$$\begin{aligned}2\mathbb{E}[\langle\hat{\theta}_t, \nabla F(\boldsymbol{\theta}_t)\rangle] &= 2\mathbb{E}[\langle\hat{\theta}_t, \nabla F(\bar{\theta}_t) - \nabla F(\theta^*)\rangle] \\ &\quad + 2\mathbb{E}[\langle\hat{\theta}_t, \nabla F(\boldsymbol{\theta}_t) - \nabla F(\bar{\theta}_t)\rangle].\end{aligned} \tag{73}$$

Since the Lipschitz property in Assumption 2-(iii) implies $\mathbb{E}[\|\nabla F(\boldsymbol{\theta}_t) - \nabla F(\bar{\theta}_t)\|^2] \leq \frac{L^2}{m}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2]$ and the convexity of $F(\theta)$ implies $\mathbb{E}[\langle\hat{\theta}_t, \nabla F(\bar{\theta}_t) - \nabla F(\theta^*)\rangle] \geq \mathbb{E}[F(\bar{\theta}_t) - F(\theta^*)]$, equality (73) can be rewritten as

$$\begin{aligned}2\mathbb{E}[\langle\hat{\theta}_t, \nabla F(\boldsymbol{\theta}_t)\rangle] &\geq 2\mathbb{E}[F(\bar{\theta}_t) - F(\theta^*)] \\ &\quad - \frac{1}{(t+1)^a}\mathbb{E}[\|\hat{\theta}_t\|^2] - (t+1)^a\frac{L^2}{m}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2],\end{aligned} \tag{74}$$

where the constant a satisfies $a \in (\frac{1}{2}, v)$.

The second term on the right hand side of (72) satisfies

$$\begin{aligned}&2\mathbb{E}[\langle\hat{\theta}_t, \nabla F_t(\boldsymbol{\theta}_t) - \nabla F(\boldsymbol{\theta}_t)\rangle] \\ &\geq -\frac{1}{(t+1)^a}\mathbb{E}[\|\hat{\theta}_t\|^2] - (t+1)^a\frac{\kappa^2}{t+1}.\end{aligned} \tag{75}$$

Substituting (74) and (75) into (72) and further Substituting (71) and (72) into (70), we obtain

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_{t+1}\|^2] &\leq \left(1 + \frac{2\lambda_t}{(t+1)^a}\right)\mathbb{E}[\|\hat{\theta}_t\|^2] + \sigma_t^2 \\ &\quad - 2\lambda_t\mathbb{E}[F(\bar{\theta}_t) - F(\theta^*)] + (t+1)^a\frac{L^2\lambda_t}{m}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2] \\ &\quad + \lambda_t(t+1)^a\frac{\kappa^2}{t+1} + \frac{\kappa^2\lambda_t^2}{t+1} + D^2\lambda_t^2.\end{aligned} \tag{76}$$

We proceed to characterize the fourth term on the right hand side of (76). According to Algorithm 1, we have $\check{\boldsymbol{\theta}}_{t+1} = (I_{mn} + W\otimes I_n)\check{\boldsymbol{\theta}}_t + \check{\boldsymbol{\vartheta}}_t^w - \lambda_t\left(\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t) - \mathbf{1}_m\otimes\nabla F_t(\boldsymbol{\theta}_t)\right)$. By using Assumption 3 and the Young's inequality, we have

$$\begin{aligned}\mathbb{E}[\|\check{\boldsymbol{\theta}}_{t+1}\|^2] &\leq (1+\delta_2)\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2] + 2m\sigma_t^2 \\ &\quad + 8\left(1 + \frac{1}{\delta_2}\right)m(\kappa^2 + D^2)\lambda_t^2,\end{aligned} \tag{77}$$

where in the derivation we have used relationships $\mathbb{E}[\|\nabla f_i(\theta_t^i)\|] \leq D$ and $\mathbb{E}[\|\nabla f_t^i(\theta_t^i) - \nabla f_i(\theta_t^i)\|^2] \leq \frac{\kappa^2}{t+1} \leq \kappa^2$.

Given that the decaying rate of the stepsize (v) is higher than the decaying rate of the DP-noise variance (ς), inequality (77) can be rewritten as

$$\mathbb{E}[\|\check{\boldsymbol{\theta}}_{t+1}\|^2] \leq (1+\delta_2)\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2] + \frac{c_1}{(t+1)^{2\varsigma}}, \tag{78}$$

with $c_1 = 2m(\sigma^+)^2 + 8\left(1 + \frac{1}{\delta_2}\right)m(\kappa^2 + D^2)\lambda_0^2$.

Applying Lemma 4 to (78) with $c = -\delta_2 \in (0,1)$ yields

$$\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2] \leq \frac{c_2}{(t+1)^{2\varsigma}}, \tag{79}$$

where c_2 is given by $c_2 = (\frac{8\varsigma}{e\ln(\frac{2}{2+\delta_2})})^{2\varsigma}(\frac{\mathbb{E}[\|\check{\boldsymbol{\theta}}_0\|^2](1+\delta_2)}{c_1} - \frac{2}{\delta_2})$.

Substituting (79) into (76) and summing both sides of (76) from $t=0$ to $t=T$ to obtain

$$\begin{aligned}\sum_{t=0}^{T}\mathbb{E}[\|\hat{\theta}_{t+1}\|^2] &\leq \sum_{t=0}^{T}\left(1 + \frac{2\lambda_t}{(t+1)^a}\right)\mathbb{E}[\|\hat{\theta}_t\|^2] \\ &\quad - 2\sum_{t=0}^{T}\lambda_t\mathbb{E}[F(\bar{\theta}_t) - F(\theta^*)] + \sum_{t=0}^{T}\Phi_t,\end{aligned} \tag{80}$$

with $\Phi_t = \sigma_t^2 + \frac{c_2L^2\lambda_t}{m(t+1)^{2\varsigma-a}} + \frac{\kappa^2\lambda_t}{(t+1)^{1-a}} + \frac{\kappa^2\lambda_t^2}{t+1} + D^2\lambda_t^2$.

By using the relationship $\sum_{t=0}^{T}\mathbb{E}[\|\hat{\theta}_t\|^2] = \mathbb{E}[\|\hat{\theta}_0\|^2] + \sum_{t=0}^{T}\mathbb{E}[\|\hat{\theta}_{t+1}\|^2] - \mathbb{E}[\|\hat{\theta}_{T+1}\|^2]$, we can rewrite (80) as follows:

$$\begin{aligned}2\sum_{t=0}^{T}\lambda_t\mathbb{E}[F(\bar{\theta}_t) - F(\theta^*)] &\leq (1+2\lambda_0)\mathbb{E}[\|\hat{\theta}_0\|^2] \\ &\quad + \sum_{t=1}^{T}\left(\frac{2\lambda_t}{(t+1)^a}\right)\mathbb{E}[\|\hat{\theta}_t\|^2] + \sum_{t=0}^{T}\Phi_t,\end{aligned} \tag{81}$$

where in the derivation we have omitted the negative term $-\mathbb{E}[\|\hat{\theta}_{T+1}\|^2]$.

We proceed to characterize the second term on the right hand side of (81). By iterating (76) from 0 to t and omitting the negative term $-2\lambda_t\mathbb{E}[F(\bar{\theta}_t)-F(\theta^*)]$ in (76), we obtain

$$\mathbb{E}[\|\hat{\theta}_{t+1}\|^2] \leq \prod_{k=0}^{t}\left(1+\frac{2\lambda_0}{(t+1)^{v+a}}\right)\Bigg[\mathbb{E}[\|\hat{\theta}_0\|^2] + \sum_{k=0}^{t}\Bigg(\frac{(\sigma^+)^2}{(t+1)^{2\varsigma}}+\frac{L^2c_2\lambda_0}{m(t+1)^{2\varsigma+v-a}} + \frac{\lambda_0\kappa^2}{(t+1)^{1+v-a}}+\frac{\kappa^2\lambda_0^2}{(t+1)^{1+2v}}+\frac{D^2\lambda_0^2}{(t+1)^{2v}}\Bigg)\Bigg]. \tag{82}$$

Since $\ln(1+u)\leq u$ holds for all $u>0$, we always have $\prod_{k=0}^{t}\left(1+\frac{2\lambda_0}{(t+1)^{2v+a}}\right)\leq e^{2\lambda_0\sum_{k=0}^{t}\frac{1}{(t+1)^{2v+a}}}$. Moreover, all items on the right hand side of (82) can be simplified. For example, the second term can be bounded as follows:

$$\begin{aligned}\sum_{k=0}^{t}\frac{(\sigma^+)^2}{(t+1)^{2\varsigma}} &\leq (\sigma^+)^2+\sum_{k=1}^{\infty}\frac{(\sigma^+)^2}{(t+1)^{2\varsigma}}\\ &\leq (\sigma^+)^2+(\sigma^+)^2\int_1^\infty\frac{1}{x^{2\varsigma}}dx\leq\frac{2\varsigma(\sigma^+)^2}{2\varsigma-1}.\end{aligned} \tag{83}$$

Applying the same argument to the other items on the right hand side of (82), we arrive at

$$\mathbb{E}[\|\hat{\theta}_{t+1}\|^2]\leq e^{\frac{2\lambda_0(v+a)}{v+a-1}}\Bigg[\mathbb{E}[\|\hat{\theta}_0\|^2]+\frac{2\varsigma(\sigma^+)^2}{2\varsigma-1} + \frac{c_2L^2\lambda_0(2\varsigma+v-a)}{m(2\varsigma+v-a-1)}+\frac{\lambda_0\kappa^2(1+v-a)}{v-a} + \frac{\kappa^2\lambda_0^2(2v+1)}{2v}+\frac{2vD^2\lambda_0^2}{2v-1}\Bigg]\triangleq c_3, \tag{84}$$

for all $t\geq 0$. Then, substituting (84) into the second term on the right hand side of (81), we obtain

$$\sum_{t=1}^{T}\left(\frac{2\lambda_t}{(t+1)^a}\right)\mathbb{E}[\|\hat{\theta}_t\|^2]\leq\int_1^\infty\frac{2c_3\lambda_0}{x^{a+v}}dx\leq\frac{2c_3\lambda_0}{a+v-1}. \tag{85}$$

Using an argument similar to the derivation of (83) yields that the third term on the right hand side of (81) satisfies

$$\sum_{t=0}^{T}\Phi_t\leq\frac{2\varsigma(\sigma^+)^2}{2\varsigma-1}+\frac{c_2L^2\lambda_0(2\varsigma+v-a)}{m(2\varsigma+v-a-1)}+\frac{\lambda_0\kappa^2(1+v-a)}{v-a} + \frac{\kappa^2\lambda_0^2(2v+1)}{2v}+\frac{2vD^2\lambda_0^2}{2v-1}\triangleq c_4. \tag{86}$$

Substituting (85) and (86) into (81), we have

$$\begin{aligned}&\sum_{t=0}^{T}\lambda_t\mathbb{E}[F(\bar{\theta}_t)-F(\theta^*)]\\ &=\sum_{t=0}^{T}\lambda_t\mathbb{E}[F(\theta_t^i)-F(\bar{\theta}_t)]+\sum_{t=0}^{T}\lambda_t\mathbb{E}[F(\bar{\theta}_t)-F(\theta^*)]\\ &\leq\sum_{t=0}^{T}\lambda_tD\mathbb{E}[\|\theta_t^i-\bar{\theta}_t\|]+c_5\leq\sum_{t=0}^{T}\frac{\sqrt{c_2}D\lambda_0}{(t+1)^{\varsigma+v}}+c_5\leq c_6,\end{aligned} \tag{87}$$

with $c_5=\frac{1+2\lambda_0}{2}\mathbb{E}[\|\hat{\theta}_0\|^2]+\frac{c_3\lambda_0}{a+v-1}+\frac{c_4}{2}$ and $c_6=\frac{(\varsigma+v)\sqrt{c_2}D\lambda_0}{\varsigma+v-1}+c_5$, where we have used (79) in the second inequality and used an argument similar to the derivation of (83) in the last inequality. Since inequality (87) ensures $\lambda_T\sum_{t=0}^{T}\mathbb{E}[F(\bar{\theta}_t)-F(\theta^*)]\leq c_6$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[F(\bar{\theta}_t)-F(\theta^*)]\leq\frac{c_6}{\lambda_0(T+1)^{1-v}}. \tag{88}$$

C. Proof of Theorem 4

The Lipschitz property in Assumption 2-(iii) implies

$$F(\bar{\theta}_{t+1})\leq F(\bar{\theta}_t)+\langle\nabla F(\bar{\theta}_t),\bar{\theta}_{t+1}-\bar{\theta}_t\rangle+\frac{L^2}{2}\|\bar{\theta}_{t+1}-\bar{\theta}_t\|^2. \tag{89}$$

We characterize the second term on the right hand side of (89). Using the definition of $F_t(\boldsymbol{\theta}_t)$ and Line 6 in Algorithm 1, we have $\bar{\theta}_{t+1}-\bar{\theta}_t=\bar{\vartheta}_t^w-\lambda_t\nabla F_t(\boldsymbol{\theta}_t)$, which implies

$$\begin{aligned}&\mathbb{E}[\langle\nabla F(\bar{\theta}_t),\bar{\theta}_{t+1}-\bar{\theta}_t\rangle]=\mathbb{E}[\langle\nabla F(\bar{\theta}_t),\bar{\vartheta}_t^w-\lambda_t\nabla F_t(\boldsymbol{\theta}_t)\rangle]\\ &=-\lambda_t\mathbb{E}[\|\nabla F(\bar{\theta}_t)\|^2]+\lambda_t\mathbb{E}[\langle\nabla F(\bar{\theta}_t),\nabla F(\bar{\theta}_t)-\nabla F_t(\boldsymbol{\theta}_t)\rangle]\\ &\leq-\frac{\lambda_t}{2}\mathbb{E}[\|\nabla F(\bar{\theta}_t)\|^2]+\frac{\lambda_t}{2}\mathbb{E}[\|\nabla F(\bar{\theta}_t)-\nabla F_t(\boldsymbol{\theta}_t)\|^2].\end{aligned} \tag{90}$$

The second term on the right hand side of (90) satisfies

$$\begin{aligned}\mathbb{E}[\|\nabla F(\bar{\theta}_t)-\nabla F_t(\boldsymbol{\theta}_t)\|^2]&\leq2\mathbb{E}[\|\nabla F(\bar{\theta}_t)-\nabla F(\boldsymbol{\theta}_t)\|^2]\\ &\quad+2\mathbb{E}[\|\nabla F(\boldsymbol{\theta}_t)-\nabla F_t(\boldsymbol{\theta}_t)\|^2].\end{aligned} \tag{91}$$

Since Assumption 2-(iii) implies $\mathbb{E}[\|\nabla F(\bar{\theta}_t)-\nabla F(\boldsymbol{\theta}_t)\|^2]\leq\frac{L^2}{m}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2]$ and Assumption 2-(ii) implies $\mathbb{E}[\|\nabla F(\boldsymbol{\theta}_t)-\nabla F_t(\boldsymbol{\theta}_t)\|^2]\leq\frac{\kappa^2}{t+1}$, we have

$$\mathbb{E}[\|\nabla F(\bar{\theta}_t)-\nabla F_t(\boldsymbol{\theta}_t)\|^2]\leq\frac{2L^2}{m}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2]+\frac{2\kappa^2}{t+1}. \tag{92}$$

Substituting (92) into (90), we arrive at

$$\begin{aligned}&\mathbb{E}[\langle\nabla F(\bar{\theta}_t),\bar{\theta}_{t+1}-\bar{\theta}_t\rangle]\\ &\leq-\frac{\lambda_t}{2}\mathbb{E}[\|\nabla F(\bar{\theta}_t)\|^2]+\frac{L^2\lambda_t}{m}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2]+\frac{\kappa^2\lambda_t}{t+1}.\end{aligned} \tag{93}$$

We proceed to estimate an upper bound on the last term on the right hand side of (89). By using again the relation $\bar{\theta}_{t+1}-\bar{\theta}_t=\bar{\vartheta}_t^w-\lambda_t\nabla F_t(\boldsymbol{\theta}_t)$ and an argument similar to the derivation of (71), we have

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_{t+1}-\bar{\theta}_t\|^2]&=\mathbb{E}[\|\bar{\vartheta}_t^w\|^2]+\lambda_t^2\mathbb{E}[\|\nabla F_t(\boldsymbol{\theta}_t)\|^2]\\ &\leq\sigma_t^2+D^2\lambda_t^2+\frac{\kappa^2\lambda_t^2}{t+1}.\end{aligned} \tag{94}$$

Substituting (93) and (94) into (89) yields

$$\begin{aligned}\frac{\lambda_t}{2}\mathbb{E}[\|\nabla F(\bar{\theta}_t)\|^2]&\leq\mathbb{E}[F(\bar{\theta}_t)-F(\bar{\theta}_{t+1})]+\frac{L^2\lambda_t}{m}\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2]\\ &\quad+\frac{\kappa^2\lambda_t}{t+1}+\sigma_t^2+D^2\lambda_t^2+\frac{\kappa^2\lambda_t^2}{t+1}.\end{aligned} \tag{95}$$

Following an argument similar to the derivation of (77), we have

$$\begin{aligned}\mathbb{E}[\|\check{\boldsymbol{\theta}}_{t+1}\|^2]&\leq\left(1+\frac{\delta_2}{2}\right)\mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2]+2m\sigma_t^2\\ &\quad+8\left(1+\frac{2}{\delta_2}\right)m\left(\kappa^2+D^2\right)\lambda_t^2.\end{aligned} \tag{96}$$

By using $\mathbb{E}[\|\nabla F(\bar{\theta}_t)\|^2]\geq\frac{1}{2}\mathbb{E}[\|\nabla F(\theta_t^i)\|^2]-\mathbb{E}[\|\nabla F(\theta_t^i)-\nabla F(\bar{\theta}_t)\|^2]$ and (96), we sum both sides of (95) from 0 to T

to obtain

$$\textstyle\sum_{t=0}^{T} \frac{\lambda_t}{4} \mathbb{E}[\|\nabla F(\theta_t^i)\|^2] + \sum_{t=0}^{T} \mathbb{E}[\|\check{\boldsymbol{\theta}}_{t+1}\|^2] \leq \mathbb{E}[F(\bar{\theta}_0) - F(\bar{\theta}_{T+1})] \\ + \textstyle\sum_{t=0}^{T} \left(1 + \frac{\delta_2}{2} + \frac{3L^2\lambda_t}{2m}\right) \mathbb{E}[\|\check{\boldsymbol{\theta}}_t\|^2] + (m+1) \sum_{t=0}^{T} \sigma_t^2 \\ + \textstyle\sum_{t=0}^{T} \frac{\kappa^2 \lambda_t}{t+1} + 2D^2 \left(1 + 4m\left(1 - \frac{2}{\delta_2}\right)\right) \sum_{t=0}^{T} \lambda_t^2 \\ + 2\kappa^2 \left(1 + 4m\left(1 - \frac{2}{\delta_2}\right)\right) \textstyle\sum_{t=0}^{T} \frac{\lambda_t^2}{t+1}. \quad (97)$$

By applying $\sum_{t=0}^{T} \frac{1}{(t+1)^r} \leq 1 + \int_1^\infty \frac{1}{x^r} dx \leq \frac{r}{r-1}$ valid for any $r > 1$ to (97), we obtain $\frac{\lambda_T}{4} \sum_{t=0}^{T} \mathbb{E}[\|\nabla F(\theta_t^i)\|^2] \leq c_1$, where c_1 is given by $c_1 = \mathbb{E}[F(\bar{\theta}_0) - F(\theta^*)] + \mathbb{E}[\|\check{\boldsymbol{\theta}}_0\|^2] + \frac{2\varsigma(\sigma^+)^2(m+1)}{2\varsigma-1} + \frac{\kappa^2\lambda_0(1+v)}{v} + 2D^2(1 + 4m(1 - \frac{2}{\delta_2}))\frac{\lambda_0^2(2v)}{2v-1} + 2\kappa^2(1 + 4m(1 - \frac{2}{\delta_2}))\frac{\lambda_0^2(2v)}{2v}$. Therefore, we arrive at

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla F(\theta_t^i)\|^2] \leq \frac{c_1}{\lambda_0 (T+1)^{1-v}}. \quad (98)$$

D. Proof of Corollary 1

We first characterize inequality (28). By using the relationship $\sum_{t=1}^{T} \frac{1}{(t+1)^r} \leq \int_0^T \frac{1}{(x+1)^r} dx = \frac{1}{1-r}((T+1)^{1-r} - 1)$ valid for all $r \in (0,1)$, inequality (28) satisfies $\epsilon^i \leq \frac{2\sqrt{2}\lambda_0 C_3 d}{\sigma_0^i (v - \max_{i\in[m]}\{\varsigma^i\})}$. Therefore, for any given cumulative privacy budget $\epsilon^i > 0$, we have $\sigma_0^i = \frac{2\sqrt{2}\lambda_0 C_3 d}{(v - \max_{i\in[m]}\{\varsigma^i\})\epsilon^i}$, in which v and ς^i are predetermined parameters satisfying Assumption 4. It is clear that a small ϵ^i result in a larger σ_0^i.

Next, we analyze the convergence speed of Algorithm 1 when $F(x)$ is nonconvex. Based on (98), we have

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla F(\theta_t^i)\|^2] \leq \frac{C}{(T+1)^{1-v}}, \quad (99)$$

with $C = \frac{1}{\lambda_0}(\mathbb{E}[F(\bar{\theta}_0) - F(\theta^*)] + \mathbb{E}[\|\check{\boldsymbol{\theta}}_0\|^2] + \frac{2\varsigma(\sigma^+)^2(m+1)}{2\varsigma-1} + \frac{\kappa^2\lambda_0(1+v)}{v} + 2D^2(1 + 4m(1 - \frac{2}{\delta_2}))\frac{\lambda_0^2(2v)}{2v-1} + 2\kappa^2(1 + 4m(1 - \frac{2}{\delta_2}))\frac{\lambda_0^2(2v)}{2v}) = \mathcal{O}((\sigma^+)^2)$.

Given that v is independent of σ_0^i, the accurate convergence of Algorithm 1 remains attainable even if ϵ^i tends to zero. However, $C = \mathcal{O}((\sigma^+)^2)$ is a positive constant that is positively correlated with DP-noise parameter $(\sigma^+)^2$. Given that $\sigma^+ = \max_{i\in[m]}\{\sigma_0^i\}$ is inversely proportional to ϵ^i, we can obtain the following inequality based on (99):

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla F(\theta_t^i)\|^2] \leq \mathcal{O}\left(\frac{T^{-(1-v)}}{\min_{i\in[m]}\{(\epsilon^i)^2\}}\right). \quad (100)$$

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[2] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A survey of optimization methods from a machine learning perspective," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3668–3681, 2020.

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

[4] T. Zhang and Q. Zhu, "Dynamic differential privacy for ADMM-based distributed classification learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 1, pp. 172–187, 2016.

[5] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Found. Trends Mach. Learn.*, vol. 4, no. 2, pp. 107–194, 2012.

[6] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction using mini-batches," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 165–202, 2012.

[7] S. Lee, A. Nedić, and M. Raginsky, "Stochastic dual averaging for decentralized online optimization on time-varying communication graphs," *IEEE Trans. Autom. Control*, vol. 62, no. 12, pp. 6407–6414, 2017.

[8] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 714–725, 2017.

[9] S. Pu, A. Olshevsky, and I. C. Paschalidis, "A sharp estimate on the transient time of distributed stochastic gradient descent," *IEEE Trans. Autom. Control*, vol. 67, no. 11, pp. 5900–5915, 2021.

[10] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Secur.*, pp. 587–601, 2017.

[11] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 17–31, 2019.

[12] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Comput. Surv.*, vol. 54, no. 11s, pp. 1–37, 2022.

[13] C. Zhang, M. Ahmad, and Y. Wang, "ADMM based privacy-preserving decentralized optimization," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 3, pp. 565–580, 2018.

[14] C. N. Hadjicostis and A. D. Domínguez-García, "Privacy-preserving distributed averaging via homomorphically encrypted ratio consensus," *IEEE Trans. Autom. Control*, vol. 65, no. 9, pp. 3887–3894, 2020.

[15] F. Tramer and D. Boneh, "Slalom: Fast, verifiable and private execution of neural networks in trusted hardware," in *Int. Conf. Learn. Represent.*, pp. 1–19, 2018.

[16] F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi, "Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2483–2493, 2012.

[17] Y. Wang and T. Başar, "Quantization enabled privacy protection in decentralized stochastic optimization," *IEEE Trans. Autom. Control*, vol. 68, no. 7, pp. 4038–4052, 2022.

[18] Y. Lou, L. Yu, S. Wang, and P. Yi, "Privacy preservation in distributed subgradient optimization algorithms," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2154–2165, 2017.

[19] Y. Wang, "Privacy-preserving average consensus via state decomposition," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4711–4716, 2019.

[20] C. Altafini, "A system-theoretic framework for privacy preservation in continuous-time multiagent dynamics," *Automatica*, vol. 122, p. 109253, 2020.

[21] H. Wang, K. Liu, D. Han, S. Chai, and Y. Xia, "Privacy-preserving distributed online stochastic optimization with time-varying distributions," *IEEE Trans. Control Netw. Syst.*, vol. 10, no. 2, pp. 1069–1082, 2022.

[22] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.

[23] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur.*, pp. 308–318, 2016.

[24] Z. Huang, S. Mitra, and N. Vaidya, "Differentially private distributed optimization," in *Proc. 16th Int. Conf. Distrib. Comput. Netw.*, pp. 1–10, 2015.

[25] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private distributed convex optimization via functional perturbation," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 1, pp. 395–408, 2016.

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

[27] M. T. Hale and M. Egerstedt, "Cloud-enabled differentially private multiagent optimization with constraints," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 4, pp. 1693–1706, 2018.

[28] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, and Y. Gong, "DP-ADMM: ADMM-based distributed learning with differential privacy," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1002–1012, 2020.

[29] X. Chen, L. Huang, L. He, S. Dey, and L. Shi, "A differentially private method for distributed optimization in directed networks via state decomposition," *IEEE Trans. Control Netw. Syst.*, vol. 10, no. 4, pp. 2165–2177, 2023.

[30] C. Liu, K. H. Johansson, and Y. Shi, "Distributed empirical risk minimization with differential privacy," *Automatica*, vol. 162, p. 111514, 2024.

[31] L. Huang, J. Wu, D. Shi, S. Dey, and L. Shi, “Differential privacy in distributed optimization with gradient tracking,” *IEEE Trans. Autom. Control*, vol. 69, no. 9, pp. 5727–5742, 2024.

[32] Y. Lin, K. Liu, D. Han, and Y. Xia, “Statistical privacy-preserving online distributed nash equilibrium tracking in aggregative games,” *IEEE Trans. Autom. Control*, vol. 69, no. 1, pp. 323–330, 2023.

[33] Y. Wang and A. Nedić, “Differentially-private distributed algorithms for aggregative games with guaranteed convergence,” *IEEE Trans. Autom. Control*, vol. 69, no. 8, pp. 5168–5183, 2024.

[34] Y. Wang and A. Nedić, “Tailoring gradient methods for differentially private distributed optimization,” *IEEE Trans. Autom. Control*, vol. 69, no. 2, pp. 872–887, 2023.

[35] C. Li, P. Zhou, L. Xiong, Q. Wang, and T. Wang, “Differentially private distributed online learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, pp. 1440–1453, 2018.

[36] J. Zhu, C. Xu, J. Guan, and D. O. Wu, “Differentially private distributed online algorithms over time-varying directed networks,” *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 1, pp. 4–17, 2018.

[37] Q. Lü, K. Zhang, S. Deng, Y. Li, H. Li, S. Gao, and Y. Chen, “Privacy-preserving decentralized dual averaging for online optimization over directed networks,” *IEEE Trans. Ind. Cyber-Phys. Syst.*, vol. 1, pp. 79–91, 2023.

[38] Y. Xiong, J. Xu, K. You, J. Liu, and L. Wu, “Privacy-preserving distributed online optimization over unbalanced digraphs via subgradient rescaling,” *IEEE Trans. Control Netw. Syst.*, vol. 7, no. 3, pp. 1366–1378, 2020.

[39] M. Yuan, J. Lei, and Y. Hong, “Differentially private distributed online mirror descent algorithm,” *Neurocomputing*, vol. 551, p. 126531, 2023.

[40] Z. Zhao, J. Yang, W. Gao, Y. Wang, and M. Wei, “Differentially private distributed online optimization via push-sum one-point bandit dual averaging,” *Neurocomputing*, vol. 572, p. 127184, 2024.

[41] J. Wang and J.-F. Zhang, “Differentially private distributed stochastic optimization with time-varying sample sizes,” *IEEE Trans. Autom. Control*, vol. 69, no. 9, pp. 6341–6348, 2024.

[42] X. Zhang, M. M. Khalili, and M. Liu, “Improving the privacy and accuracy of ADMM-based distributed algorithms,” in *Int. Conf. Mach. Learn.*, pp. 5796–5805, PMLR, 2018.

[43] T. Ding, S. Zhu, J. He, C. Chen, and X. Guan, “Differentially private distributed optimization via state and direction perturbation in multiagent systems,” *IEEE Trans. Autom. Control*, vol. 67, no. 2, pp. 722–737, 2021.

[44] E. Garcelon, V. Perchet, C. Pike-Burke, and M. Pirotta, “Local differential privacy for regret minimization in reinforcement learning,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 10561–10573, 2021.

[45] Z. Chen and Y. Wang, “Locally differentially private decentralized stochastic bilevel optimization with guaranteed convergence accuracy,” in *Proc. Int. Conf. Mach. Learn.,*, pp. 1–51, 2024.

[46] B. Li and Y. Chi, “Convergence and privacy of decentralized nonconvex optimization with gradient clipping and communication compression,” *arXiv preprint arXiv:2305.09896*, 2023.

[47] Z. Chen and Y. Wang, “Locally differentially private distributed online learning with guaranteed optimality,” *IEEE Trans. Autom. Control (Early access)*, 2024.

[48] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei, “LDP-Fed: federated learning with local differential privacy,” in *Proc. 3rd ACM Int. Workshop Edge Syst. Anal. Netw.*, pp. 61–66, 2020.

[49] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, “Federated learning with differential privacy: Algorithms and performance analysis,” *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3454–3469, 2020.

[50] M. Kim, O. Günlü, and R. F. Schaefer, “Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication,” in *Proc. 2021 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2650–2654, IEEE, 2021.

[51] N. Lang, E. Sofer, T. Shaked, and N. Shlezinger, “Joint privacy enhancement and quantization in federated learning,” *IEEE Trans. Signal Process.*, vol. 71, pp. 295–310, 2023.

[52] J. Zeng and W. Yin, “On nonconvex decentralized gradient descent,” *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2834–2848, 2018.

[53] P. Bianchi and J. Jakubowicz, “Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization,” *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, 2012.

[54] D. Jakovetić, J. Xavier, and J. M. Moura, “Fast distributed gradient methods,” *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[55] T. Tatarenko and B. Touri, “Non-convex distributed optimization,” *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 3744–3757, 2017.

[56] H.-T. Wai, J. Lafond, A. Scaglione, and E. Moulines, “Decentralized frank–wolfe algorithm for convex and nonconvex problems,” *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 5522–5537, 2017.

[57] A. K. Menon, A. S. Rawat, S. J. Reddi, and S. Kumar, “Can gradient clipping mitigate label noise?,” in *Int. Conf. Learn. Represent.*, pp. 1–26, 2020.

[58] X. Chen, S. Z. Wu, and M. Hong, “Understanding gradient clipping in private sgd: A geometric perspective,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 13773–13782, 2020.

[59] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of FedAvg on Non-IID data,” in *Int. Conf. Learn. Represent.*, pp. 1–26, 2019.

[60] H. Yang, M. Fang, and J. Liu, “Achieving linear speedup with partial worker participation in Non-IID federated learning,” in *Int. Conf. Learn. Represent.*, pp. 1–23, 2021.

[61] Y. Wang and H. V. Poor, “Decentralized stochastic optimization with inherent privacy protection,” *IEEE Trans. Autom. Control*, vol. 68, no. 4, pp. 2293–2308, 2022.

[62] T. Homem-de Mello, “On rates of convergence for stochastic optimization problems under non–independent and identically distributed sampling,” *SIAM J. Optim.*, vol. 19, no. 2, pp. 524–551, 2008.

[63] B. Bullins, K. Patel, O. Shamir, N. Srebro, and B. E. Woodworth, “A stochastic newton algorithm for distributed convex optimization,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 26818–26830, 2021.

[64] L. Berrada, A. Zisserman, and P. Mudigonda, “Smooth loss functions for deep top-k classification,” in *Int. Conf. Learn. Represent.*, pp. 1–25, 2018.

[65] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” *Syst. Control Lett.*, vol. 53, no. 1, pp. 65–78, 2004.

[66] C. Liu, K. H. Johansson, and Y. Shi, “Private stochastic dual averaging for decentralized empirical risk minimization,” *IFAC-PapersOnLine*, vol. 55, no. 13, pp. 43–48, 2022.

[67] Z. Chen and Y. Wang, “Locally differentially private gradient tracking for distributed online learning over directed graphs,” *IEEE Trans. Autom. Control (Early access)*, 2024.

[68] D. Dua, C. Graff, *et al.*, *UCI machine learning repository*. School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2007. Available: http://archive.ics.uci.edu/ml.

[69] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” in *Proc. IEEE*, vol. 86, pp. 2278–2324, IEEE, 1998.

[70] A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*. Master’s thesis, Univ. Toronto, Canada, 2009.

Ziqin Chen received the Ph.D. degree in Automation from the University of Science and Technology of China, Hefei, China, in 2020. She was a postdoctoral researcher at Tongji University, China, from 2020 to 2022.

She is currently a postdoctoral researcher in the Department of Electrical and Computer Engineering at Clemson University, USA. Her research interests include differential privacy, decentralized optimization/learning, and game theory.

Yongqiang Wang (Senior Member, IEEE) was born in Shandong, China. He received the dual B.S. degrees in electrical engineering and automation and computer science and technology from Xi’an Jiaotong University, Xi’an, China, in 2004, and the M.Sc. and Ph.D. degrees in control science and engineering from Tsinghua University, Beijing, China, in 2009. From 2007 to 2008, he was with the University of Duisburg-Essen, Duisburg, Germany, as a Visiting Student. He was a Project Scientist with the University of California, Santa Barbara, CA, USA before joining Clemson University, SC, USA, where he is currently an Associate Professor. His current research interests include decentralized control, optimization, and learning, with an emphasis on privacy and security.

Prof. Wang currently serves as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL and IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS.