Locally Differentially Private Gradient Tracking for Distributed Online Learning over Directed Graphs

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

***Abstract*—Distributed online learning has been proven extremely effective in solving large-scale machine learning problems over streaming data. However, information sharing between learners in distributed learning also raises concerns about the potential leakage of individual learners' sensitive data. To mitigate this risk, differential privacy, which is widely regarded as the "gold standard" for privacy protection, has been widely employed in many existing results on distributed online learning. However, these results often face a fundamental tradeoff between learning accuracy and privacy. In this paper, we propose a locally differentially private gradient-tracking-based distributed online learning algorithm that successfully circumvents this tradeoff. We prove that the proposed algorithm converges in mean square to the exact optimal solution while ensuring rigorous local differential privacy, with the cumulative privacy budget guaranteed to be finite even when the number of iterations tends to infinity. The algorithm is applicable even when the communication graph among learners is directed. To the best of our knowledge, this is the first result that simultaneously ensures learning accuracy and rigorous local differential privacy in distributed online learning over directed graphs. We evaluate our algorithm's performance by using multiple benchmark machine-learning applications, including logistic regression of the "mushrooms" dataset and CNN-based image classification of the "MNIST" and "CIFAR-10" datasets, respectively. The experimental results confirm that the proposed algorithm outperforms existing counterparts in both training and test accuracies.**

***Index Terms*—Decentralized online learning, local differential privacy, directed graph, gradient tracking.**

I. Introduction

Machine learning is rapidly reshaping the landscape of various engineering domains, ranging from wireless sensor networks [1], autonomous driving [2] to image classification [3]. Different from the conventional centralized learning scheme, where all data are stored on one device, distributed learning enables multiple participating learners to cooperatively learn a common optimal solution while each participating learner only trains on its own local dataset. Hence, compared with centralized learning, distributed learning provides inherent advantages in scalability and privacy, and thereby has garnered increased attention over the past decade [4]–[7].

In existing distributed learning approaches, the most commonly used algorithm is distributed stochastic gradient descent (DSGD) [8]. While DSGD is communication-efficient and simple to implement, it suffers from slow convergence when data are heterogeneous among learners [9], [10]. To mitigate the issue brought by data heterogeneity, gradient-tracking-based distributed optimization algorithms have emerged [11]–[15], which replace the local gradient in every learner's update in DSGD with an estimated global gradient. Besides the classical gradient-tracking approach which requires balanced network topologies, this approach has also been extended to the case with general directed network topologies in both others' works [16]–[21] and our prior work [22]. All aforementioned gradient-tracking-based algorithms consider a fixed and static objective function, which, in machine learning, amounts to requiring all training data to be available beforehand. However, in numerous real-world applications, the data are sequentially acquired [23], which prompts the investigation of online gradient-tracking-based algorithms [24]–[26].

Moreover, in existing online gradient-tracking-based algorithms, repeated message exchanges are required among neighboring learners, which poses significant privacy threats to individual learners' sensitive datasets. As shown in [27], [28], even though raw data are not shared during distributed training, external adversaries could infer individuals' sensitive information from shared messages. To address privacy concerns in distributed learning/optimization, various approaches have been proposed. For example, partially homomorphic cryptography has been widely considered in distributed optimization [29]. But this approach incurs a high communication and computation cost. Another approach involves the injection of spatially- or temporally-correlated noises to obfuscate information shared in distributed optimization [30]–[34]. However, to protect the privacy of a learner, this approach requires this learner to have at least one trustworthy neighbor, which is undesirable for fully distributed applications. As the de facto standard for privacy protection, differential privacy (DP) is gaining increased traction, and has been employed in many distributed learning and optimization algorithms [35]–[47].

Most existing DP solutions for distributed learning and optimization only consider undirected or balanced graph topologies [35]–[37], [40]–[42]. Recently, results have emerged on DP design for distributed learning under directed graph topologies, for both online learning [43]–[45] and conventional (offline) learning where data are predetermined [46], [47]. However, existing results on differentially private distributed online learning over directed graphs usually build on the combination of push-sum [43] based or eigenvector-estimation [44], [45] based and DSGD, which limits their achievable convergence speeds. To the best of our knowledge, no results have been reported on DP design for gradient-tracking-based online learning over directed graphs.

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.

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

Another limitation with existing DP solutions [35]–[37], [40]–[46] for distributed learning and optimization applications is that these solutions are subject to a fundamental tradeoff between privacy and learning accuracy. Recently, our work [47] for distributed offline optimization successfully circumvents this tradeoff and achieves both optimality and privacy. Nevertheless, the gradient-tracking-based approach therein relies on incorporating two weakening factors (two decaying sequences that are multiplied on coupling weights to make the weights decay with time) into inter-agent interaction to mitigate the impact of DP noises, which consequently slows down algorithmic convergence. Furthermore, this approach is designed for an offline setting, where all data are predetermined. Our recent work [48] proposed a local differential privacy (LDP) approach for distributed online learning that can ensure learning accuracy and privacy simultaneously. However, this approach requires undirected graphs. In addition, it also hinges on a weakening factor, which significantly decreases the speed of algorithm convergence.

In this work, we introduce an LDP approach for distributed online learning over directed graphs that ensures both learning accuracy and rigorous LDP (with the privacy budget guaranteed to be finite even when the number of iterations tends to infinity). Specifically, we first modify the conventional architecture of gradient tracking to ensure learning accuracy despite the presence of DP noises. This modification is crucial because DP noises will accumulate in the estimate of the global gradient in conventional gradient-tracking algorithms. In fact, in the presence of DP noises, the variance of accumulated noises will grow to infinity in conventional gradient-tracking-based distributed optimization, which significantly affects learning accuracy, as confirmed in our theoretical analysis in Sec. III-A and experimental results in Sec. VI. It is worth noting that while the approach in [18] for offline optimization can prevent the accumulated noise variance in gradient estimation from growing to infinity, it cannot entirely eliminate the influence of noises on optimization accuracy. In contrast, our algorithm effectively eliminates the influence of DP noises on local gradient estimation, and thus ensures accurate convergence. Then, we prove that the proposed algorithm can ensure LDP with a finite cumulative privacy budget, even in the infinite time horizon. Furthermore, by leveraging the online eigenvector-estimation technique in [49], our proposed algorithm enables each learner to locally estimate the left normalized Perron eigenvector of the interaction graph, which allows the treatment of imbalanced graphs and hence applications in general directed networks. To the best of our knowledge, this is the first work that successfully achieves LDP in gradient-tracking-based distributed online learning and optimization over directed graphs. The main contributions are summarized as follows:

- We prove that the proposed distributed online learning algorithm converges in mean square to the optimal solution, even when the DP noises are present and the communication graph is directed. Note that existing online gradient-tracking algorithms in [24]–[26] employ the conventional gradient-tracking approach, which is susceptible to noises due to the accumulation of variance in gradient estimation [18], [22], [47].
- In addition to ensuring accurate convergence, our algorithm also achieves rigorous LDP with a finite cumulative privacy budget, even in the infinite time horizon. This stands in stark contrast to most existing DP solutions for distributed learning and optimization [35]–[37], [40]–[46], where the cumulative privacy budget grows to infinity as the number of iterations tends to infinity (implying diminishing DP protection). A key enabler for our approach to ensure a finite cumulative privacy budget in the infinite time horizon is to employ diminishing stepsizes rather than the commonly used constant stepsizes.
- Compared with existing DP solutions for distributed learning and optimization in [35]–[37], [40]–[42] which require balanced network topologies, our proposed algorithm is applicable to general directed network topologies.
- The adopted LDP framework preserves agent-level privacy for each learner's dataset without relying on any trusted third parties. This differs from the traditional DP framework employed in [35]–[37], [40]–[47], where a "centralized" data aggregator is implicitly assumed to determine the amount of injected noises.
- Different from our prior results in [47] and [48] relying on weakening factors in inter-agent coupling to ensure both learning accuracy and privacy, which unavoidably reduce the speed of convergence, our algorithm here avoids using any weakening factors, and hence can attain faster convergence speed, as confirmed in our analytical comparison in Sec. IV-B and experimental results in Sec. VI.
- We evaluate the performance of our algorithm using multiple benchmark machine-learning applications, including online logistic regression of the "mushrooms" dataset and image classification of the "MNIST" and "CIFAR-10" datasets, respectively. Moreover, the experimental results show that compared with existing state-of-the-art DP solutions in [37], [42], [44], [47], [48], our proposed algorithm provides better training and test accuracies.

The rest of the paper is organized as follows. Sec. II introduces some preliminaries and the problem formulation. Sec. III proposes our LDP approach for online gradient tracking. Sec. IV analyzes the learning accuracy of the proposed algorithm. Sec. V establishes rigorous LDP guarantees. Sec. VI provides experimental results. Sec. VII concludes the paper.

II. Preliminaries and Problem Statement

A. Notations

We use $\mathbb{R}^n$ to denote the n-dimensional real Euclidean space and $\mathbb{N}$ ($\mathbb{N}^+$) to denote the set of non-negative (positive) integers. We write $\mathbf{1}_n$ and I_n for the n-dimensional column vector of all ones and the identity matrix, respectively. For an arbitrary vector x, we denote its ith element by $[x]_i$. We write $\langle \cdot, \cdot \rangle$ for the inner product of two vectors and $\|\cdot\|$ for the standard Euclidean norm of a vector. For an arbitrary matrix A, we denote its transpose by A^T and its Frobenius norm by $\|A\|_F$. We also use other vector/matrix norms defined under a certain transformation determined by a matrix W, which will

be represented as $\|\cdot\|_W$. We write $\mathbb{P}[\mathcal{A}]$ for the probability of an event $\mathcal{A}$ and $\mathbb{E}[x]$ for the expected value of a random variable x. The notation $\lceil a \rceil$ refers to the smallest integer not less than a and $\lfloor a \rfloor$ represents the largest integer not greater than a. We use $\text{Lap}(\nu)$ to denote the Laplace distribution with a parameter $\nu > 0$, featuring a probability density function $\frac{1}{2\nu} e^{\frac{-|x|}{\nu}}$. $\text{Lap}(\nu)$ has a mean of zero and a variance of $2\nu^2$. We abbreviate independent and identically distributed by *i.i.d.*

B. Network model

We model the topology of the network over which learners communicate with each other as a directed graph $\mathcal{G} = ([m], \mathcal{E})$, where $[m] = \{1, \cdots, m\}$ denotes the agent (learner) set and $\mathcal{E} \subseteq [m] \times [m]$ represents the edge set consisting of ordered pairs of agents. Given a nonnegative matrix $W = \{w_{ij}\} \in \mathbb{R}^{m \times m}$, we define its induced directed graph as $\mathcal{G}_W = ([m], \mathcal{E}_W)$, where $(i, j) \in \mathcal{E}_W$ if and only if $w_{ij} > 0$. For a learner $i \in [m]$, it is able to receive messages from the learners in its in-neighbor set $\mathcal{N}_{W,i}^{\text{in}} = \{j \in [m] | w_{ij} > 0\}$; Similarly, learner i can also send messages to learners in its out-neighbor set $\mathcal{N}_{W,i}^{\text{out}} = \{j \in [m] | w_{ji} > 0\}$. Graph $\mathcal{G}_W$ is called strongly connected if there exists a directed path between any pair of distinct learners. In this paper, we consider a gradient-tracking-based algorithm which maintains two optimization variables [16] that can be shared on two different graphs. We represent the two directed graphs as $\mathcal{G}_R$ and $\mathcal{G}_C$, which are induced by matrices $R \in \mathbb{R}^{m \times m}$ and $C \in \mathbb{R}^{m \times m}$, respectively.

Assumption 1. *The matrices $R \in \mathbb{R}^{m \times m}$ and $C \in \mathbb{R}^{m \times m}$ have nonnegative off-diagonal entries, i.e., $R_{ij} \geq 0$ and $C_{ij} \geq 0$ for all $i \neq j$. Their diagonal entries are negative, satisfying $R_{ii} = -\sum_{j \in \mathcal{N}_{R,i}^{in}} R_{ij}$ and $C_{ii} = -\sum_{j \in \mathcal{N}_{C,i}^{out}} C_{ji}$, i.e, $R\mathbf{1} = \mathbf{0}$ and $\mathbf{1}^T C = \mathbf{0}^T$. Moreover, the induced graph $\mathcal{G}_R$ is strongly connected and $\mathcal{G}_{C^T}$ contains at least one spanning tree.*

Assumption 1 is weaker than requiring both $\mathcal{G}_R$ and $\mathcal{G}_C$ to be strongly connected in [43]–[45]. We have the following lemma on matrices R and C:

Lemma 1. *[16] Under Assumption 1, the matrix $\mathbf{R} \triangleq I + R$ has a unique positive left eigenvector u^T (corresponding to eigenvalue 1) satisfying $u^T \mathbf{1} = m$, and the matrix $\mathbf{C} \triangleq I + C$ has a unique positive right eigenvector ω (corresponding to eigenvalue 1) satisfying $\mathbf{1}^T \omega = m$.*

C. Local differential privacy

Differential privacy guarantees that when two datasets differ by only one data point (record), the output of a DP implementation does not reveal whether that specific data point was utilized. This property makes it difficult for an external adversary to identify individual data entries among all possible ones, thereby providing strong privacy protection.

In this paper, we consider an agent-level LDP framework, and thus, changes in a dataset are formalized by an adjacency relation pertaining to the local dataset of learner $i \in [m]$:

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

Remark 1. Our definition of adjacency corresponds to the so-called event-level LDP in the literature [50]. For any given t, it allows m entries in the global datasets of all learners to be different, and is more stringent than most existing results using the traditional centralized version of DP (e.g., [41]–[45]), where for any given t, only one data entry is allowed to be different. It is also worth noting that allowing one learner to have all data entries to be different (called user-level DP [50]) has been proven infeasible in distributed optimization/learning under the local model of DP [50]–[53].

According to Definition 1, two local datasets $\mathcal{D}_t^i$ and $\mathcal{D}_t'^i$ are adjacent if and only if they differ by only one entry while all other entries are the same. We denote the adjacency relationship between $\mathcal{D}_t^i$ and $\mathcal{D}_t'^i$ by $\text{Adj}(\mathcal{D}_t^i, \mathcal{D}_t'^i)$. With this understanding, we formally define LDP as follows:

Definition 2. *(Local Differential Privacy) We say that an implementation $\mathcal{A}_i(\mathcal{D}^i, \theta^{-i})$ of a randomized algorithm by learner i provides ϵ_i-local differential privacy if for any adjacent datasets $\mathcal{D}^i$ and $\mathcal{D}_t'^i$, the following inequality holds:*

$$\mathbb{P}[\mathcal{A}_i(\mathcal{D}^i, \theta^{-i}) \in \mathcal{O}_i] \leq e^{\epsilon_i} \mathbb{P}[\mathcal{A}_i(\mathcal{D}_t'^i, \theta^{-i}) \in \mathcal{O}_i], \quad (1)$$

where θ^{-i} denotes all messages received by learner i and $\mathcal{O}_i$ represents the set of all possible observations on learner i.

The privacy budget of learner i's implementation is quantified by ϵ_i. It can be seen that a smaller ϵ_i indicates closer distributions of observations under adjacent datasets, thereby ensuring a higher level of privacy protection.

Remark 2. The conventional centralized DP framework used in [35]–[37], [41]–[47] implicitly assumes mutual trust among learners to cooperatively decide each learner's DP-noise needed to satisfy a global privacy budget ϵ. In contrast, our LDP framework removes the need for such trust and allows individual learners to independently set (potentially heterogeneous) privacy budgets ϵ_i and choose the corresponding DP noises according to their individual needs. Therefore, our LDP framework provides a stronger and more user-friendly privacy framework.

D. LDP approach for distributed online learning

We consider a distributed online learning problem involving m learners. Each learner only has access to its own private dataset. At each iteration t, learner $i \in [m]$ acquires a data point $\xi_t^i = (a_t^i, b_t^i)$, which is independent and sampled from an unknown time-invariant distribution $\mathcal{P}_i$. Using the sample a_t^i and the current model parameter θ_t^i, learner i predicts a label $\hat{b}_t^i = \langle \theta_t^i, a_t^i \rangle$ with an associated loss $l(\theta_t^i, \xi_t^i)$, which quantifies the deviation between $\hat{b}_t^i$ and the true label b_t^i. This loss prompts learner i to update its model parameter from θ_t^i to θ_{t+1}^i. The objective is that, based on sequentially acquired data, all learners converge to the same optimal solution θ^* to the following stochastic optimization problem:

$$\min_{\theta\in\mathbb{R}^n} F(\theta), \quad F(\theta)=\frac{1}{m}\sum_{i=1}^{m} f_i(\theta), \tag{2}$$

where $f_i(\theta)=\mathbb{E}_{\xi^i\sim\mathcal{P}_i}\big[l(\theta,\xi^i)\big]$ represents the local objective function of learner i.

Assumption 2. *(i) Problem* (2) *has at least one optimal solution θ^*; (ii) the gradients of local objective functions are uniformly bounded, i.e., there exists some positive constant D such that we have $\|\nabla f_i(\theta)\|\le D$ for all $i\in[m]$ and $\theta\in\mathbb{R}^n$; and (iii) for any $\theta_1,\theta_2\in\mathbb{R}^n$, there exists some $\mu\ge 0$ such that $F(\theta_2)\ge F(\theta_1)+\nabla F(\theta_1)^T(\theta_2-\theta_1)+\frac{\mu}{2}\|\theta_1-\theta_2\|_2^2$ holds.*

Assumption 2(iii) represents a strongly convex (when $\mu>0$) or a convex (when $\mu=0$) condition on the global objective function $F(\theta)$, which is weaker than requiring all local objective functions $f_i(\theta)$ to be strongly convex (see, e.g., [40], [41], [46], [54]) or convex (see, e.g., [42]–[45]). Many loss functions in machine learning satisfy this assumption, with some typical examples including the linear regression loss $F(\theta)=\frac{1}{m}\sum_{i=1}^{m}(b^i-(a^i)^T\theta)^2$ and the cross-entropy loss $F(\theta)=-\frac{1}{m}\sum_{i=1}^{m}(b^i\log(\sigma((a^i)^T\theta))+(1-b^i)\log(1-\sigma((a^i)^T\theta)))$, where $\sigma(p)=\frac{1}{1+e^{-p}}$ is the sigmoid function and (a^i,b^i) represents learner i's data point. When a regularization term $\frac{\lambda}{2}\|\theta\|_2^2$ is added to these functions (to balance overfitting and underfitting a model during training), they become strongly convex.

We also need the following assumption, which is a standard assumption in distributed stochastic optimization [12], [25].

Assumption 3. *We assume that the data points $\{\xi_t^i\}$ are i.i.d. across iterations. In addition, (i) $\mathbb{E}[\nabla l(\theta,\xi_t^i)]=\nabla f_i(\theta)$; (ii) $\mathbb{E}[\|\nabla l(\theta,\xi_t^i)-\nabla f_i(\theta)\|_2^2]\le\kappa^2$; and (iii) $\|\nabla l(\theta_1,\xi_t^i)-\nabla l(\theta_2,\xi_t^i)\|_2\le L\|\theta_1-\theta_2\|_2$ for any $\theta_1,\theta_2\in\mathbb{R}^n$.*

Assumption 3(iii) implies that the function $l(\theta,\xi_t^i)$ is L-smooth. This condition is commonly used in differentially private distributed learning [38]–[40] and is satisfied by many loss functions used in machine learning [55].

Since the local objective function $f_i(\theta)$ is defined as an expectation over random data ξ^i sampled from an unknown distribution $\mathcal{P}_i$, it is inaccessible in practice and an analytical solution to problem (2) is unattainable. To tackle this issue, we focus on solving the following empirical risk minimization (ERM) problem with sequentially arriving data:

$$\min_{\theta\in\mathbb{R}^n} F_t(\theta), \quad F_t(\theta)=\frac{1}{m}\sum_{i=1}^{m} f_t^i(\theta), \tag{3}$$

where $f_t^i(\theta)=\frac{1}{t+1}\sum_{k=0}^{t} l(\theta,\xi_k^i)$ denotes the empirical local objective function of each learner $i\in[m]$.

According to the law of large numbers [56], one has $\lim_{t\to\infty}\frac{1}{t+1}\sum_{k=0}^{t} l(\theta,\xi_k^i)=\mathbb{E}_{\xi^i\sim\mathcal{P}_i}\big[l(\theta,\xi^i)\big]$. Hence, problem (3) serves as an approximation to the original problem (2). Unlike some existing online optimization results (in, e.g., [57]) where the optimal solution is time-varying, the solution to our time-varying ERM formulation in (3) converges to a constant value, i.e., the optimal solution θ^* to (2):

Lemma 2. *Denote θ_t^* as the optimal solution to problem* (3) *at time t and θ^* as the optimal solution to the original stochastic optimization problem* (2)*. Under Assumption 2 with $\mu>0$ and Assumption 3, we have*

$$\mathbb{E}[\|\theta_t^*-\theta^*\|_2^2]\le 4\kappa^2\mu^{-2}(t+1)^{-1}. \tag{4}$$

Proof. See Appendix A. □

Remark 3. Although Lemma 2 implies that our ERM problem (3)'s solution θ_t^* converges with t to a constant solution θ^*, it cannot be solved using traditional time-invariant or offline optimization methods, in, e.g., [11]–[15], due to its time-varying nature of objective functions caused by the sequential acquisition of data samples.

With this understanding, our goal is to design a distributed online learning algorithm on general directed graphs which enables individual learners to track the optimal solution θ_t^* to problem (3) under the constraints of LDP and sequentially arriving data samples. Based on the convergence result in (4), individual learners' parameters will also converge to the true optimal solution to problem (2), even under the constraints of LDP and sequentially arriving data samples.

III. Online gradient tracking with LDP

In this section, we develop an online gradient-tracking-based distributed learning algorithm over directed graphs to solve problem (2) with ensured ϵ_i-LDP. Before introducing our algorithm, we first show the limitation of conventional gradient-tracking algorithms under LDP constraints.

A. The conventional gradient tracking accumulates DP noises in gradient estimation

To preserve privacy, DP noises have to be added to messages shared in each iteration of distributed online learning. In conventional gradient-tracking-based algorithms, the injected DP noises will accumulate in the global gradient estimation, thereby significantly affecting learning accuracy.

We use the classic Push-Pull gradient-tracking algorithm in [16] as an example to illustrate the idea. In the absence of LDP constraints, i.e., when no DP noise is introduced into the information exchange among learners, the Push-Pull algorithm can be described in matrix form as follows:

$$\begin{cases}\boldsymbol{\theta}_{t+1}=\mathbf{R}\boldsymbol{\theta}_t-\lambda_t\boldsymbol{y}_t,\\ \boldsymbol{y}_{t+1}=\mathbf{C}\boldsymbol{y}_t+\nabla\boldsymbol{f}_{t+1}(\boldsymbol{\theta}_{t+1})-\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t),\end{cases}$$

where the matrices $\boldsymbol{\theta}_t$, $\boldsymbol{y}_t$, and $\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)$ are defined as $\boldsymbol{\theta}_t=[\theta_t^1,\cdots,\theta_t^m]^T\in\mathbb{R}^{m\times n}$, $\boldsymbol{y}_t=[y_t^1,\cdots,y_t^m]^T\in\mathbb{R}^{m\times n}$, and $\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)=[\nabla f_t^1(\theta_t^1),\cdots,\nabla f_t^m(\theta_t^m)]^T\in\mathbb{R}^{m\times n}$, respectively. The matrices $\mathbf{R}$ and $\mathbf{C}$ are from Lemma 1.

Using initialization $\boldsymbol{y}_0=\nabla\boldsymbol{f}_0(\boldsymbol{\theta}_0)$, we obtain $\mathbf{1}^T\boldsymbol{y}_t=\mathbf{1}^T\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)$, which means that ensuring the consensus of all y_t^i, i.e., $y_t^i=\frac{1}{m}\mathbf{1}^T\boldsymbol{y}_t$, is sufficient to guarantee each learner to track the global gradient, i.e., $y_t^i=\frac{1}{m}\mathbf{1}^T\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)$.

To achieve ϵ_i-LDP, DP noises have to be added to both shared variables $\boldsymbol{\theta}_t$ and $\boldsymbol{y}_t$. Then, the update of the conventional Push-Pull algorithm becomes

$$\begin{cases}\boldsymbol{\theta}_{t+1}=\mathbf{R}\boldsymbol{\theta}_t+\boldsymbol{\vartheta}_{R,t}-\lambda_t\boldsymbol{y}_t, & \text{(5a)}\\ \boldsymbol{y}_{t+1}=\mathbf{C}\boldsymbol{y}_t+\boldsymbol{\zeta}_{C,t}+\nabla\boldsymbol{f}_{t+1}(\boldsymbol{\theta}_{t+1})-\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t), & \text{(5b)}\end{cases}$$

where the DP noises $\boldsymbol{\zeta}_{C,t}$ and $\boldsymbol{\vartheta}_{R,t}$ are defined as $\boldsymbol{\zeta}_{C,t} = [\zeta^1_{C,t}, \cdots, \zeta^m_{C,t}]^T \in \mathbb{R}^{m\times n}$ and $\boldsymbol{\vartheta}_{R,t} = [\vartheta^1_{R,t}, \cdots, \vartheta^m_{R,t}]^T \in \mathbb{R}^{m\times n}$ with $\zeta^i_{C,t} = \sum_{j\in\mathcal{N}^{\text{in}}_{C,i}} C_{ij}\zeta^i_t$ and $\vartheta^i_{R,t} = \sum_{j\in\mathcal{N}^{\text{in}}_{R,i}} R_{ij}\vartheta^i_t$ for all $i \in [m]$, respectively.

It can be seen that even under the condition $\boldsymbol{y}_0 = \nabla \boldsymbol{f}_0(\boldsymbol{\theta}_0)$, we can only establish the following relation through induction:

$$\mathbf{1}^T \boldsymbol{y}_t = \mathbf{1}^T \left(\nabla \boldsymbol{f}_t(\boldsymbol{\theta}_t) + \textstyle\sum_{k=0}^{t-1} \boldsymbol{\zeta}_{C,k} \right), \tag{6}$$

which implies that the DP noise accumulates over time in the estimate of the global gradient. Therefore, when the gradient-estimate variable $\boldsymbol{y}_t$ is directly fed into the model parameter update (5a), learning accuracy will be compromised. This prediction is corroborated by our experimental results in Fig. 2-Fig. 4. The issue of DP-noise accumulations also exists in other gradient-tracking-based algorithms for distributed learning and optimization.

Remark 4. To circumvent the accumulation of noises in gradient estimation, recent work [18] proposes a robust gradient-tracking method for distributed offline optimization. However, this method cannot completely eliminate the influence of information-sharing noises, and thus is subject to steady-state errors. Although our recent work [22] employs a weakening factor in inter-agent interaction to attenuate noise influence and ensure optimization accuracy, such a weakening factor decreases the coupling strength among learners, which in turn reduces the speed of algorithmic convergence. Moreover, although [15], [54] consider distributed optimization in the presence of perturbation/noise, their perturbation/noise is deterministic and bounded (see Eq. (27) in [15] and Theorem 1 in [54] for details). To the contrary, commonly used differential-privacy noises (e.g., Gaussian noise and Laplace noise used in our paper) are stochastic and unbounded (under Gaussian and Laplace noises, for any given number, no matter how large it is, there is always a non-zero probability that the noise amplitude is over this given number). Hence, the perturbation/noise models considered in [15], [54] are not applicable in DP design considered here. Notably, none of these works consider privacy protection. In fact, under the constant stepsize and noise variance employed in [15], [18], [54] or the single weakening factor used in [22], it is impossible to ensure rigorous LDP in the infinite time horizon.

Recent works [37], [46] have investigated DP design for gradient-tracking algorithms. However, both of the results face the dilemma of trading optimization accuracy for privacy. To tackle this dilemma, our recent work [47] achieves accurate convergence and privacy protection simultaneously. However, this approach relies on two carefully designed weakening factors to attenuate the impact of DP noises. Such weakening factors significantly slow down algorithmic convergence, as substantiated by our experimental results in Fig. 2-Fig. 4.

Moreover, all the aforementioned works [18], [22], [37], [46], [47], [54] require static and predetermined datasets, making them unsuitable to online learning scenarios where data arrives sequentially. To the best of our knowledge, no existing work has explored LDP design for gradient-tracking-based algorithms in an online setting.

B. LDP design for online gradient tracking

We present Algorithm 1 to address problem (2) over directed graphs under the constraints of LDP and sequentially arriving data. The injected DP noises satisfy Assumption 4.

Assumption 4. *For every $i \in [m]$ and any time $t \geq 0$, the DP-noises ζ^i_t and ϑ^i_t are zero-mean and independent across iterations. The noise variance $\mathbb{E}[\|\zeta^i_t\|_2^2] = (\sigma^i_{t,\zeta})^2$ satisfies $\sigma^i_{t,\zeta} = \sigma^i_\zeta (t+1)^{-\varsigma^i_\zeta}$ with $\sigma^i_\zeta > 0$ and $\varsigma^i_\zeta \in (\frac{1}{2}, 1)$. The noise variance $\mathbb{E}[\|\vartheta^i_t\|_2^2] = (\sigma^i_{t,\vartheta})^2$ satisfies $\sigma^i_{t,\vartheta} = \sigma^i_\vartheta (t+1)^{-\varsigma^i_\vartheta}$ with $\sigma^i_\vartheta > 0$ and $\varsigma^i_\vartheta \in (\frac{1}{2}, 1)$. Moreover, the inequality $\max_{i\in[m]}\{\varsigma^i_\zeta, \varsigma^i_\vartheta\} < v < 1$ holds, where the parameter v is the decaying rate of stepsize λ_t in Algorithm 1.*

Algorithm 1 LDP design for distributed online learning (from learner i's perspective)

1: **Input:** Random initialization $\theta^i_0 \in \mathbb{R}^n$, $s^i_0 \in \mathbb{R}^n$, and $z^i_0 = \boldsymbol{e}_i \in \mathbb{R}^m$, where $\boldsymbol{e}_i$ has the ith element equal to one and all other elements equal to zero; weighting matrices $R,\ C \in \mathbb{R}^{m\times m}$; stepsize $\lambda_t = \frac{\lambda_0}{(t+1)^v}$ with $\lambda_0 > 0$ and $v \in (\frac{1}{2}, 1)$; and DP-noises ζ^i_t and ϑ^i_t satisfying Assumption 4.
2: **for** $t = 0, 1, \cdots, T-1$ **do**
3: Using all available data up to time t, i.e., ξ^i_k for $k \in [0, t]$ and the current parameter θ^i_t, learner i computes the gradient $\nabla f^i_t(\theta^i_t) = \frac{1}{t+1}\sum_{k=0}^t \nabla l(\theta^i_t, \xi^i_k)$.
4: Push $s^i_t + \zeta^i_t$ to neighbors $j \in \mathcal{N}^{\text{out}}_{C,i}$ and pull $s^j_t + \zeta^j_t$ from neighbors $j \in \mathcal{N}^{\text{in}}_{C,i}$.
5: **Update tracking variable:**
$s^i_{t+1} = (1+C_{ii})s^i_t + \sum_{j\in\mathcal{N}^{\text{in}}_{C,i}} C_{ij}(s^j_t + \zeta^j_t) + \lambda_t \nabla f^i_t(\theta^i_t)$.
6: Push $\theta^i_t + \vartheta^i_t$ to neighbors $j \in \mathcal{N}^{\text{out}}_{R,i}$ and pull $\theta^j_t + \vartheta^j_t$ from $j \in \mathcal{N}^{\text{in}}_{R,i}$.
7: **Update model parameter:**
$\theta^i_{t+1} = (1 + R_{ii})\theta^i_t + \sum_{j\in\mathcal{N}^{\text{in}}_{R,i}} R_{ij}(\theta^j_t + \vartheta^j_t) - \frac{s^i_{t+1} - s^i_t}{m[z^i_t]_i}$, where $[z^i_t]_i$ denotes the ith element of z^i_t.
8: **Locally estimate the left eigenvector of $\boldsymbol{R}$:**
$z^i_{t+1} = z^i_t + \sum_{j\in\mathcal{N}^{\text{in}}_{R,i}} R_{ij}(z^j_t - z^i_t)$.
9: **end for**

The Line 5 and Line 7 in Algorithm 1 can be written in the following matrix form:

$$\begin{cases} \boldsymbol{s}_{t+1} = \mathbf{C}\boldsymbol{s}_t + \boldsymbol{\zeta}_{C,t} + \lambda_t \nabla \boldsymbol{f}_t(\boldsymbol{\theta}_t), & \text{(7a)} \\ \boldsymbol{\theta}_{t+1} = \mathbf{R}\boldsymbol{\theta}_t + \boldsymbol{\vartheta}_{R,t} - Z_t^{-1}(\boldsymbol{s}_{t+1} - \boldsymbol{s}_t), & \text{(7b)} \end{cases}$$

where the matrices $\boldsymbol{\theta}_t$, $\boldsymbol{s}_t$, and Z_t are given by $\boldsymbol{\theta}_t = [\theta^1_t, \cdots, \theta^m_t]^T \in \mathbb{R}^{m\times n}$, $\boldsymbol{s}_t = [s^1_t, \cdots, s^m_t]^T \in \mathbb{R}^{m\times n}$, and $Z_t = \text{diag}(m[z^1_t]_1, \cdots, m[z^m_t]_m) \in \mathbb{R}^{m\times m}$, respectively.

In (7b), the difference $\boldsymbol{s}_{t+1} - \boldsymbol{s}_t$ is incorporated into the parameter update. This modification effectively addresses the issue of accumulating DP noises in global gradient estimation, as substantiated by the following relation:

$$\mathbf{1}^T(\boldsymbol{s}_{t+1} - \boldsymbol{s}_t) = \mathbf{1}^T(\boldsymbol{\zeta}_{C,t} + \lambda_t \nabla \boldsymbol{f}_t(\boldsymbol{\theta}_t)), \tag{8}$$

where in the derivation we have used (7a) and $\mathbf{1}^T C = \mathbf{0}^T$ from Assumption 1. It is clear that unlike the conventional Push-Pull gradient-tracking algorithm (5), where global gradient

estimation $\boldsymbol{y}_t$ (which is subject to accumulating DP noises as per (6)) is directly incorporated into the model parameter update, thereby affecting learning accuracy, our Algorithm 1 effectively circumvents this issue.

In addition, we introduce a local variable z_t^i in Algorithm 1 to enable each learner to locally estimate the left eigenvector u^T of $\mathbf{R}$. This eliminates the need for global information u^T, ensuring that our algorithm can be implemented in a fully distributed manner. It is worth noting that since z_t^i does not contain sensitive information, adding DP noises to it is unnecessary. Next, we present the following lemma to characterize the error of the eigenvector estimator:

Lemma 3. *[22] Under Assumption 1, the variables z_t^i in Line 8 of Algorithm 1, after scaled by m, converge to the left eigenvector $u^T = [u_1, \cdots, u_m]^T$ of $\mathbf{R}$ with a geometric rate, i.e., there exist some constants $c_z > 0$ and $\gamma_z \in (0,1)$ such that $\left|\frac{1}{m[z_t^i]_i} - \frac{1}{u_i}\right| \leq c_z \gamma_z^t$ holds for all $i \in [m]$ and any $t \geq 0$, where $[z_t^i]_i$ denotes the ith element of z_t^i.*

Remark 5. Algorithm 1 avoids using weakening factors on inter-agent interaction to attenuate the influence of DP noises, which is key in our prior results [47] and [48] to ensure both optimization accuracy and rigorous DP. Given that a weakening factor will gradually reduce the strength of inter-agent coupling, and hence, unavoidably decrease the convergence speed, our algorithm can ensure faster convergence compared with [47] and [48], which is corroborated by our analytical comparison in Sec. IV-B and experimental results in Sec. VI.

Remark 6. In Algorithm 1, each learner updates its iteration variables at the same iteration count. Although this approach may increase waiting time (as a learner needs to wait for the slowest neighbor to complete its update before moving to the next iteration), it ensures consistent learning progression among learners, which simplifies the algorithmic implementation and convergence analysis.

Remark 7. Compared with [15], [18] which consider communication/quantization noises in gradient tracking, our algorithm has fundamental differences in both algorithm structure and parameter design to ensure both rigorous differential privacy and accurate convergence. More specifically, in terms of algorithm structure, we place the stepsize in the update of tracking variables, which is necessary to ensure a decaying sensitivity and is fundamentally different from [15], [18] that place the stepsize in the update of optimization variables. In terms of parameter design, we employ decaying stepsizes, which is necessary to ensure differential privacy in the infinite time horizon and is different from the constant stepsize used in [15], [18]. In addition, the spectral-radius-based convergence analysis in [15], [18] relies on the stepsize being constant, making it inapplicable in our case where the stepsize is varying with time.

IV. Online Learning Accuracy Analysis

In this section, we quantify the learning accuracy of Algorithm 1. To this end, we present some useful lemmas.

A. Supporting lemmas

Lemma 4. *[16] Under Assumption 1, there exist vector norms $\|x\|_R \triangleq \|\tilde{R}x\|_2$ and $\|x\|_C \triangleq \|\tilde{C}x\|_2$ for all $x \in \mathbb{R}^m$, where $\tilde{R}, \tilde{C} \in \mathbb{R}^{m\times m}$ are some reversible matrices[1], such that $\|\mathbf{R} - \frac{\mathbf{1}u^T}{m}\|_R < 1$ and $\|\mathbf{C} - \frac{\omega\mathbf{1}^T}{m}\|_C < 1$ are arbitrarily close to the spectral radius of $\mathbf{R} - \frac{\mathbf{1}u^T}{m}$ and $\mathbf{C} - \frac{\omega\mathbf{1}^T}{m}$, respectively.*

According to Lemma 4 in [16] and [22], we can know that the spectral radius of the matrix $\mathbf{R} - \frac{\mathbf{1}u^T}{m}$ is equal to $1 - |\nu_R| < 1$, where ν_R is an eigenvalue of R. Lemma 4 indicates that $\|\mathbf{R} - \frac{\mathbf{1}u^T}{m}\|_R$ is arbitrarily close to the spectral radius of $\mathbf{R} - \frac{\mathbf{1}u^T}{m}$, i.e., $1 - |\nu_R|$. Without loss of generality, we denote $\|\mathbf{R} - \frac{\mathbf{1}u^T}{m}\|_R = 1 - \rho_R < 1$, where ρ_R serves as an arbitrarily close approximation of $|\nu|_R$. Similarly, we denote $\|\mathbf{C} - \frac{\omega\mathbf{1}^T}{m}\|_C = 1 - \rho_C < 1$, where ρ_C is an arbitrarily close approximation of $|\nu_C|$ with ν_C an eigenvalue of C.

Following [16] and [22], we proceed to define the matrix norms $\|\boldsymbol{x}\|_R = \|[\|\boldsymbol{x}_{(1)}\|_R, \cdots, \|\boldsymbol{x}_{(n)}\|_R]\|_2$ and $\|\boldsymbol{y}\|_C = \|[\|\boldsymbol{y}_{(1)}\|_C, \cdots, \|\boldsymbol{y}_{(n)}\|_C]\|_2$ for any matrices $\boldsymbol{x},\ \boldsymbol{y} \in \mathbb{R}^{m\times n}$, where $\boldsymbol{x}_{(i)}$ and $\boldsymbol{y}_{(i)}$ denote the ith column of $\boldsymbol{x}$ and $\boldsymbol{y}$ for $1 \leq i \leq n$, respectively. The subscript 2 denotes the 2-norm.

Lemma 5. *[16] Given an arbitrary norm $\|\cdot\|$, for any $M \in \mathbb{R}^{m\times m}$ and $\boldsymbol{x} \in \mathbb{R}^{m\times n}$, we have $\|M\boldsymbol{x}\| \leq \|M\|\|\boldsymbol{x}\|$. In particular, for any $m \in \mathbb{R}^{m\times 1}$ and $x \in \mathbb{R}^{1\times n}$, we have $\|mx\| = \|m\|\|x\|_2$.*

Lemma 6. *[16] According to the equivalence of all norms in a finite-dimensional space, there exist constants $\delta_{F,R}, \delta_{R,F}, \delta_{C,F}, \delta_{R,C}, \delta_{F,C} > 0$ such that for all $\boldsymbol{x} \in \mathbb{R}^{m\times n}$, we have $\|\boldsymbol{x}\|_F \leq \delta_{F,R}\|\boldsymbol{x}\|_R$, $\|\boldsymbol{x}\|_R \leq \delta_{R,F}\|\boldsymbol{x}\|_F$, $\|\boldsymbol{x}\|_C \leq \delta_{C,F}\|\boldsymbol{x}\|_F$, $\|\boldsymbol{x}\|_R \leq \delta_{R,C}\|\boldsymbol{x}\|_C$, and $\|\boldsymbol{x}\|_F \leq \delta_{F,C}\|\boldsymbol{x}\|_C$.*

Lemma 7. *The relation $a\gamma^t \leq \frac{1}{t^2}$ always holds for all $t > 0$ and $\gamma \in (0,1)$, where the constant a is given by $a = \frac{(\ln(\gamma)e)^2}{4}$.*

Proof. We consider a convex function $f(x) = -2\ln(x) - x\ln(\gamma) : \mathbb{R}^+ \to \mathbb{R}$, whose minimal value is $f(x^*) = -2\ln(-\frac{2}{\ln(\gamma)}) + \frac{2}{\ln(\gamma)}\ln(\gamma) = \ln(a)$. Hence, for any $t > 0$, we have $f(t) \geq \ln a$, i.e., $-2\ln(t) - t\ln(\gamma) \geq \ln(a)$, which is equivalent to $\ln(\gamma^t) \leq \ln(\frac{1}{at^2})$ and further implies Lemma 7. □

B. Online learning accuracy analysis

In this subsection, we analyze the learning accuracy of Algorithm 1 under strongly convex and general convex objective functions, respectively.

For notational simplicity, we define $\bar{s}_t = \frac{\mathbf{1}^T\boldsymbol{s}_t}{m}$, $\bar{\theta}_t = \frac{u^T\boldsymbol{\theta}_t}{m}$, $\varsigma_\zeta = \min_{i\in[m]}\{\varsigma_\zeta^i\}$, and $\varsigma_\vartheta = \min_{i\in[m]}\{\varsigma_\vartheta^i\}$. The following lemmas establish the convergence properties for $\mathbb{E}[\|\boldsymbol{s}_t - \omega\bar{s}_t\|_C^2]$ and $\mathbb{E}[\|\boldsymbol{\theta}_t - \mathbf{1}\bar{\theta}_t\|_R^2]$ under general convex objective functions.

[1]As indicated in [16] and [58], $\tilde{R}$ and $\tilde{C}$ are determined by R and C, respectively. They always exist but are hard to express in a closed form in the general case. However, in the special case where $\boldsymbol{R}$ and $\boldsymbol{C}$ are primitive and stochastic, $\tilde{R}$ and $\tilde{C}$ can be expressed as $\tilde{R} = \text{diag}(\sqrt{\pi_{\boldsymbol{R}}})$ and $\tilde{C} = \text{diag}(\sqrt{\pi_{\boldsymbol{C}}})^{-1}$, where $\text{diag}(\cdot)$ denotes the diagonal matrix with the given entries on the diagonal and $\pi_{\boldsymbol{R}}$ and $\pi_{\boldsymbol{C}}$ denote non-$\mathbf{1}_m$ Perron vectors of $\boldsymbol{R}$ and $\boldsymbol{C}$, respectively (see details in Section II-B in [58]). A detailed discussion on $\tilde{R}$ ($\tilde{C}$) is available in Lemma 5 of [18], as well as Lemma 5.6.10 of [59].

Lemma 8. *Under Assumptions 1-4 with $\mu \geq 0$, the following relation holds for Algorithm 1:*

$$\mathbb{E}[\|\boldsymbol{s}_t - \omega\bar{s}_t\|_C^2] < c_{\boldsymbol{s}1}t^{-2} + c_{\boldsymbol{s}2}t^{-2v} + c_{\boldsymbol{s}3}t^{-2\varsigma_\zeta}, \quad (9)$$

where the constants $c_{\boldsymbol{s}1}$, $c_{\boldsymbol{s}2}$, and $c_{\boldsymbol{s}3}$ are given in (41).

Proof. See Appendix B. □

Lemma 9. *Under Assumptions 1-4 with $\mu \geq 0$, the following relation holds for Algorithm 1:*

$$\mathbb{E}[\|\boldsymbol{\theta}_t - \mathbf{1}\bar{\theta}_t\|_R^2] < c_{\boldsymbol{\theta}1}t^{-2} + c_{\boldsymbol{\theta}2}t^{-2v} + c_{\boldsymbol{\theta}3}t^{-2\varsigma_\vartheta} + c_{\boldsymbol{\theta}4}t^{-2\varsigma_\zeta}, \quad (10)$$

where the constants $c_{\boldsymbol{\theta}1}$, $c_{\boldsymbol{\theta}2}$, $c_{\boldsymbol{\theta}3}$, and $c_{\boldsymbol{\theta}4}$ are given in (48).

Proof. See Appendix C. □

Based on Lemma 8 and Lemma 9, we present the learning accuracy of Algorithm 1 against the original optimal solution to problem (2) under strongly convex objective functions:

Theorem 1. *Denote θ^* as the optimal solution to the original stochastic optimization problem* (2)*. Under Assumptions 1-4 with $\mu > 0$, the parameters θ_t^i in Algorithm 1 will converge in mean square to θ^*, i.e.,*

$$\mathbb{E}[\|\theta_t^i - \theta^*\|_2^2] < 8\kappa^2\mu^{-2}t^{-1} + 2C_1t^{-\beta} = \mathcal{O}(t^{-\beta}), \quad (11)$$

for all $t > 0$, where the rate β satisfies $\beta = \min\{v + \frac{1}{2} - \alpha, 2 - v - \alpha, 2\varsigma_\vartheta - \alpha, 2\varsigma_\zeta - \alpha\}$ with $\alpha \in (v, \frac{1+v}{2})$, the positive constant κ is from Assumption 3(ii), and the constant C_1 is given by $C_1 = \max_{1\leq i\leq 4, 1\leq j\leq 17}\{c_{\boldsymbol{\theta}i}, c_{\bar{\theta}j}\}$ with $c_{\boldsymbol{\theta}1}$ to $c_{\boldsymbol{\theta}4}$ given in (48)*, and $c_{\bar{\theta}1}$ to $c_{\bar{\theta}17}$ given in Eqs.* (69)-(71)*.*

Proof. See Appendix D. □

Theorem 1 establishes the convergence of Algorithm 1 to the optimal solution to problem (2) under DP noises. This differs from most existing DP solutions for distributed learning and optimization [37], [40]–[46], which are always subject to optimization errors under rigorous DP constraints. In fact, besides ensuring convergence accuracy, our algorithm guarantees rigorous LDP even in the infinite time horizon, which will be substantiated in Sec. V.

Unlike most existing results on distributed online optimization [40]–[45] which focus on dynamic or static regrets with respect to the optimal solution to problem (3) (which only approximates the optimal solution to (2)), Theorem 1 provides a direct quantitative measure of the learning error with respect to the optimal solution θ^* to the problem (2) at each iteration. Moreover, Theorem 1 shows that the convergence speed of Algorithm 1 is $\mathcal{O}(t^{-\beta})$ with $\beta = \min\{v + \frac{1}{2} - \alpha, 2 - v - \alpha, 2\varsigma_\vartheta - \alpha, 2\varsigma_\zeta - \alpha\}$. This speed outpaces that of the distributed online learning algorithm in our prior work [48] by a factor of $\mathcal{O}(t^{\frac{v+1-2\alpha}{2}})$ (the convergence speed in [48] is $\mathcal{O}(t^{-\beta'})$ with $\beta' = \min\{1-v, 2\varsigma_\vartheta - 1\}$). In addition, the algorithm in [48] only characterizes the deviation between the learned parameter θ_t^i and the optimal solution θ_t^* to an approximated formulation of (2). Hence, Theorem 1 provides stronger and more precise convergence than the result in [48].

Remark 8. By characterizing the constant C_1 in Theorem 1, we can obtain $\mathbb{E}[\|\theta_t^i - \theta^*\|_2^2] \leq \mathcal{O}\big(\big(\frac{L^3}{\rho_R^4\rho_C^3} + \frac{1}{\mu^3}\big)t^{-\beta}\big)$. It is clear that a larger strongly convex coefficient μ, a smaller Lipschitz constant L, and larger ρ_R and ρ_C (i.e., better-connected networks) lead to faster convergence.

Next, we establish the convergence result for general convex objective functions.

Theorem 2. *Under Assumptions 1-4 with $\mu \geq 0$, the objective function values $F(\theta_t^i)$ will converge in mean to the minimal objective function value $F(\theta^*)$, i.e.,*

$$\mathbb{E}[F(\theta_t^i) - F(\theta^*)] \leq \frac{(1-v)\sum_{i=1}^4 \bar{c}_{\boldsymbol{\theta}i}t^{v-1}}{2\lambda_0(1-\frac{1}{2^{1-v}})} = \mathcal{O}(t^{-\bar{\beta}}), \quad (12)$$

for all $t > 0$, where the rate $\bar{\beta}$ satisfies $\bar{\beta} = 1 - v$ and the constants $\bar{c}_{\boldsymbol{\theta}1}$ to $\bar{c}_{\boldsymbol{\theta}4}$ are given in Eqs. (84)-(86)*, respectively.*

Proof. See Appendix E. □

Theorem 2 characterizes the convergence of $F(\theta_t^i)$ to the minimal objective function value $F(\theta^*)$. Moreover, the convergence speed specified in Theorem 2 (i.e., $\mathcal{O}(t^{-(1-v)})$) is twice as fast as that in our prior work [48], which converges at a speed of $\mathcal{O}(t^{-\frac{1-v}{2}})$ for general convex objective functions.

Remark 9. According to the definitions of $\bar{c}_{\boldsymbol{\theta}1}$ to $\bar{c}_{\boldsymbol{\theta}4}$ given in Eqs. (84)-(86), we have $\mathbb{E}[F(\theta_t^i) - F(\theta^*)] \leq \mathcal{O}\big(\frac{L^2}{\rho_R^4\rho_C^3}t^{-(1-v)}\big)$, which implies that a smaller Lipschitz constant L and larger ρ_R and ρ_C (i.e., better-connected networks) lead to faster convergence.

V. Local differential-privacy analysis

In this section, we prove that besides accurate convergence, Algorithm 1 can also simultaneously ensure rigorous ϵ_i-LDP for each learner, even in the infinite time horizon. To this end, we first introduce the concept of sensitivity for learner i's implementation $\mathcal{A}_i$:

Definition 3. *(Sensitivity) Let $\mathcal{D}_t^i$ and $\mathcal{D}_t'^i$ be any two adjacent datasets for learner i at each time instant t. The sensitivity of learner i's implementation $\mathcal{A}_i$ at time t is*

$$\Delta_{t+1}^i = \max_{Adj(\mathcal{D}_t^i, \mathcal{D}_t'^i)} \|\mathcal{A}_i(\mathcal{D}_t^i, \theta_t^{-i}) - \mathcal{A}_i(\mathcal{D}_t'^i, \theta_t^{-i})\|_1, \quad (13)$$

where θ_t^{-i} represents all messages received by learner i at time instant t.

According to Definition 3, under Algorithm 1, learner i's implementation involves two sensitivities: $\Delta_{t,s}^i$ and $\Delta_{t,\theta}^i$, which correspond to the two shared variables s_t^i and θ_t^i, respectively.

With this understanding, we have the following lemma:

Lemma 10. *For any given $T \in \mathbb{N}^+$ or $T = \infty$, if learner i injects to each of its shared variables s_t^i and θ_t^i at each time $t \in \{1, \cdots, T\}$ noise vectors ζ_t^i and ϑ_t^i consisting of n independent Laplace noises with parameters $\nu_{t,\zeta}^i$ and $\nu_{t,\vartheta}^i$, respectively, then learner i's implementation $\mathcal{A}_i$ is ϵ_i-locally differentially private with the cumulative privacy budget from time $t = 1$ to $t = T$ upper bounded by $\sum_{t=1}^T(\frac{\Delta_{t,s}^i}{\nu_{t,\zeta}^i} + \frac{\Delta_{t,\theta}^i}{\nu_{t,\vartheta}^i})$.*

Proof. The lemma can be obtained following the same line of reasoning of Lemma 2 in [35]. □

For privacy analysis, we also need the following lemma:

Lemma 11. *Denote $\{\psi_t\}$ as a nonnegative sequence. If there exists a sequence $\beta_t = \frac{\beta_0}{(t+1)^q}$ with some $\beta_0 > 0$ and $q > 0$ such that $\psi_{t+1} \leq (1-c)\psi_t + \beta_t$ holds for all $c \in (0, 1)$,*

then we always have $\psi_t \leq C_2\beta_t$ for all $t \in \mathbb{N}$, where the constant C_2 is given by $C_2 = (\frac{4q}{e\ln(\frac{2}{2-c})})^q(\frac{v_0(1-c)}{\beta_0} + \frac{2}{c})$.

Proof. See Appendix F. □

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

Assumption 5 is commonly used in DP design for distributed optimization and learning [42]–[44].

Without loss of generality, we consider adjacent datasets $\mathcal{D}_t^i$ and $\mathcal{D}_t'^i$ that differ in the k-th element, i.e., ξ_k^i in $\mathcal{D}_t^i$ and $\xi_k'^i$ in $\mathcal{D}_t'^i$ are different. For the sake of clarity, the parameters learned from $\mathcal{D}_t^i$ and $\mathcal{D}_t'^i$ are denoted as θ_{t+1}^i and $\theta_{t+1}'^i$, respectively.

Theorem 3. *Under Assumptions 1-5, if each element of ϑ_t^i and ζ_t^i follows the Laplace distributions Lap$(\nu_{t,\vartheta}^i)$ and Lap$(\nu_{t,\zeta}^i)$, respectively, with $(\sigma_{t,\vartheta}^i)^2 = 2(\nu_{t,\vartheta}^i)^2$ and $(\sigma_{t,\zeta}^i)^2 = 2(\nu_{t,\zeta}^i)^2$ satisfying Assumption 4, then θ_t^i (resp. $F(\theta_t^i)$ in the general convex case) in Algorithm 1 converges in mean square to the optimal solution θ^* to the optimization problem* (2) *(resp. in mean to $F(\theta^*)$). Furthermore,*

1) For any finite number of iterations T, each learner i's implementation of Algorithm 1 is ϵ_i-locally differentially private with a cumulative privacy budget bounded by $\sum_{t=1}^T(\frac{\sqrt{2}\varrho_{t,s}(t+1)^{\varsigma_\zeta^i}}{\sigma_\zeta^i} + \frac{\sqrt{2}\varrho_{t,\theta}(t+1)^{\varsigma_\vartheta^i}}{\sigma_\vartheta^i})$ with $\varrho_{t,s} = 2c_l\sum_{p=1}^t(1-\min_{i\in[m]}\{|C_{ii}|\})^{t-p}\lambda_{p-1}$, $\varrho_{0,s} = 0$, and $\varrho_{t,\theta} = \sum_{p=1}^t(1-\min_{i\in[m]}\{|R_{ii}|\})^{t-p}(c_z\gamma_z^{p-1} + \frac{1}{|u_i|})(\varrho_{p,s} + \varrho_{p-1,s})$.

2) The cumulative privacy budget is finite even when the number of iterations T tends to infinity, i.e., when $T \to \infty$, the cumulative privacy budget is bounded by $\sum_{t=1}^\infty\left(\frac{2\sqrt{2}C_4c_l\lambda_0(c_z\gamma_z^t+C_0)}{\sigma_\vartheta^i(t+1)^{1+v-\varsigma_\vartheta^i}} + \frac{2\sqrt{2}C_4c_l\lambda_0(c_z\gamma_z^t+C_0)}{(C_0-\frac{1}{|u_i|})\sigma_\zeta^i(t+1)^{1+v-\varsigma_\zeta^i}}\right) < \infty$, where C_0 and C_4 are given in (21) *and* (24)*, respectively.*

Proof. The convergence result follows naturally from Theorem 1 (resp. Theorem 2).

1) To prove the statements on privacy, we first analyze the sensitivity of learner i's implementation under Algorithm 1.

According to the definition of sensitivity in (13), we have $s_t^j + \zeta_t^j = s_t'^j + \zeta_t'^j$ and $\theta_t^j + \vartheta_t^j = \theta_t'^j + \vartheta_t'^j$ for each time $t \geq 0$ and $j \in \mathcal{N}_i$. Since we assume that the k-th data point is different between $\mathcal{D}_t^i = \{\xi_1^i, \cdots, \xi_k^i, \cdots, \xi_t^i\}$ and $\mathcal{D}_t'^i = \{\xi_1^i, \cdots, \xi_k'^i, \cdots, \xi_t^i\}$, we have $\xi_p^i = \xi_p'^i$ for all $p \neq k$. However, since the difference in loss functions kicks in at time k, i.e., $l(\theta, \xi_k^i) \neq l(\theta, \xi_k'^i)$, we have $s_t^i \neq s_t'^i$ and $\theta_t^i \neq \theta_t'^i$. Hence, for learner i's implementation of Algorithm 1, we have

$$\begin{aligned}\|s_{t+1}^i - s_{t+1}'^i\|_1 = \big\|&(1+C_{ii})(s_t^i - s_t'^i)\\ &+ \tfrac{\lambda_t}{t+1}\textstyle\sum_{p=0,p\neq k}^t(\nabla l(\theta_t^i, \xi_p^i) - \nabla l(\theta_t'^i, \xi_p^i))\\ &+ \tfrac{\lambda_t}{t+1}(\nabla l(\theta_t^i, \xi_k^i) - \nabla l(\theta_t'^i, \xi_k'^i))\big\|_1,\end{aligned} \tag{14}$$

Letting $c_C = \min_{i\in[m]}\{|C_{ii}|\}$, the sensitivity $\Delta_{t+1,s}^i$ satisfies

$$\Delta_{t+1,s}^i \leq (1-c_C)\Delta_{t,s}^i + \tfrac{\lambda_t}{t+1}\textstyle\sum_{p=0}^t\|\nabla l(\theta_t^i, \xi_p^i) - \nabla l(\theta_t'^i, \xi_p'^i)\|_1, \tag{15}$$

where we have used $\xi_p^i = \xi_p'^i$ for all $p \in [0, t]$ and $p \neq k$.

By using Assumption 5 and the relation $\Delta_{0,s}^i = 0$, we iterate (15) from 0 to $t-1$ to obtain

$$\Delta_{t,s}^i \leq 2c_l\textstyle\sum_{p=1}^t(1-c_C)^{t-p}\lambda_{p-1}. \tag{16}$$

Similarly, we use Line 7 in Algorithm 1 to obtain

$$\begin{aligned}\|\theta_{t+1}^i - \theta_{t+1}'^i\|_1 = \big\|&(1+R_{ii})(\theta_t^i - \theta_t'^i)\\ &- \tfrac{1}{m[z_t^i]_i}(s_{t+1}^i - s_{t+1}'^i) + \tfrac{1}{m[z_t^i]_i}(s_t^i - s_t'^i)\big\|_1.\end{aligned}$$

Letting $c_R = \min_{i\in[m]}\{|R_{ii}|\}$ and using Lemma 3, the sensitivity $\Delta_{t+1,\theta}^i$ satisfies

$$\begin{aligned}\Delta_{t+1,\theta}^i \leq{}& (1-c_R)\Delta_{t,\theta}^i + c_z\gamma_z^t\Delta_{t+1,s}^i + c_z\gamma_z^t\Delta_{t,s}^i\\ &+ \tfrac{1}{|u_i|}\Delta_{t+1,s}^i + \tfrac{1}{|u_i|}\Delta_{t,s}^i.\end{aligned} \tag{17}$$

By using the relation $\Delta_{0,\theta}^i = 0$ and iterating (17) from 0 to $t-1$, we obtain

$$\Delta_{t,\theta}^i \leq \textstyle\sum_{p=1}^t(1-c_R)^{t-p}(c_z\gamma_z^{p-1} + \frac{1}{|u_i|})(\Delta_{p,s}^i + \Delta_{p-1,s}^i). \tag{18}$$

The inequalities (16) and (18) imply that for learner i, the T-iteration cumulative privacy budget are bounded by $\sum_{t=1}^T(\frac{\sqrt{2}\varrho_{t,s}(t+1)^{\varsigma_\zeta^i}}{\sigma_\zeta^i} + \frac{\sqrt{2}\varrho_{t,\theta}(t+1)^{\varsigma_\vartheta^i}}{\sigma_\vartheta^i})$, with $\varrho_{t,s}$ and $\varrho_{t,\theta}$ given in the theorem statement.

2) The Lipschitz property in Assumption 3(iii) implies that for the same data ξ_p^i, we can rewrite (14) as

$$\Delta_{t+1,s}^i \leq (1-c_C)\Delta_{t,s}^i + \tfrac{\sqrt{n}Lt\lambda_t}{t+1}\Delta_{t,\theta}^i + \tfrac{2c_l\lambda_t}{t+1}, \tag{19}$$

where in the derivation we have used Assumption 5.

By substituting (19) into (17), we have

$$\begin{aligned}\Delta_{t+1,\theta}^i \leq{}& \left(1 - c_R + \tfrac{\sqrt{n}Lc_z(t\gamma_z^t\lambda_t)}{t+1}\right)\Delta_{t,\theta}^i + \tfrac{1}{|u_i|}\Delta_{t+1,s}^i\\ &+ (2-c_C)c_z\gamma_z^t\Delta_{t,s}^i + \tfrac{2c_lc_z\gamma_z^t\lambda_t}{t+1} + \tfrac{1}{|u_i|}\Delta_{t,s}^i.\end{aligned} \tag{20}$$

By selecting positive constants:

$$C_3 < \min\left\{\tfrac{c_R}{2}, \tfrac{c_C}{2}\right\} \text{ and } C_0 > \max\left\{\tfrac{4}{|u_i|(c_C-2C_3)}, \tfrac{1}{|u_i|}\right\}, \tag{21}$$

we multiply both sides of (19) by C_0 and combine (19) and (20) to obtain

$$\begin{aligned}&\Delta_{t+1,\theta}^i + \left(C_0 - \tfrac{1}{|u_i|}\right)\Delta_{t+1,s}^i\\ &\leq \left(1 - c_R + \tfrac{\sqrt{n}Lc_z(t\gamma_z^t\lambda_t)}{t+1} + \tfrac{C_0\sqrt{n}Lt\lambda_t}{t+1}\right)\Delta_{t,\theta}^i\\ &\quad + \left((2-c_C)c_z\gamma_z^t + C_0(1-c_C) + \tfrac{1}{|u_i|}\right)\Delta_{t,s}^i\\ &\quad + \tfrac{2c_lc_z\gamma_z^t\lambda_t + 2C_0c_l\lambda_t}{t+1}.\end{aligned} \tag{22}$$

Since λ_t and γ_z^t are decaying sequences, there must exist some $T_0 \geq 0$ such that $\frac{c_R}{2} \geq \frac{\sqrt{n}Lc_z(t\gamma_z^t\lambda_t) + C_0\sqrt{n}Lt\lambda_t}{t+1}$ and $(2-c_C)c_z\gamma_z^t \leq \frac{C_0c_C}{2}$ hold for all $t \geq T_0$. Hence, we arrive at

$$\begin{aligned}\Delta_{t+1,\theta}^i + \big(C_0 - \tfrac{1}{|u_i|}\big)\Delta_{t+1,s}^i \leq{}& (1-C_3)\big(\Delta_{t,\theta}^i\\ &+ \big(C_0 - \tfrac{1}{|u_i|}\big)\Delta_{t,s}^i\big) + \tfrac{2c_lc_z\gamma_z^t\lambda_t + 2C_0c_l\lambda_t}{t+1},\end{aligned} \tag{23}$$

for all $t \geq T_0$, where we used $(1-C_3)(C_0 - \frac{1}{|u_i|}) > C_0(1 - \frac{c_C}{2}) + \frac{1}{|u_i|}$ according to the definitions of C_0 and C_3.

We further define a constant $C_4 > 0$ as follows:

$$\begin{aligned}C_4 = \max\Bigg\{&\left(\tfrac{4(1+v)}{e\ln(\frac{2}{2-C_3})}\right)^{1+v}\tfrac{2}{1-C_3},\\ &\max_{0\leq t\leq T_0, i\in[m]}\left\{\tfrac{\left(\Delta_{t,\theta}^i + \left(C_0 - \frac{1}{|u_i|}\right)\Delta_{t,s}^i\right)(t+1)}{2c_lc_z\gamma_z^t\lambda_t + 2C_0c_l\lambda_t}\right\}\Bigg\}.\end{aligned} \tag{24}$$

Combining Lemma 11 and (23), we obtain

$$\Delta^i_{t+1,\theta}+(C_0-\tfrac{1}{|u_i|})\Delta^i_{t+1,s}\leq C_4\tfrac{2c_lc_z\gamma_z^t\lambda_0+2C_0c_l\lambda_0}{(t+1)^{1+v}},\quad(25)$$

for all $t>0$. By using Lemma 10, we arrive at

$$\sum_{t=1}^{\infty}\left(\tfrac{\Delta^i_{t,\theta}}{\nu^i_{t,\vartheta}}+\tfrac{\Delta^i_{t,s}}{\nu^i_{t,\zeta}}\right)\leq\sum_{t=1}^{\infty}\Big(\tfrac{2\sqrt{2}C_4c_l\lambda_0(c_z\gamma_z^t+C_0)}{\sigma^i_\vartheta(t+1)^{1+v-\varsigma^i_\vartheta}}+\tfrac{2\sqrt{2}C_4c_l\lambda_0(c_z\gamma_z^t+C_0)}{(C_0-\frac{1}{|u_i|})\sigma^i_\zeta(t+1)^{1+v-\varsigma^i_\zeta}}\Big),\quad(26)$$

implying that the cumulative privacy budget is finite since $1+v-\max\{\varsigma^i_\vartheta,\varsigma^i_\zeta\}>1$ always holds. □

Theorem 3 proves that the privacy budget is finite even when the number of iterations T tends to infinity, thereby establishing rigorous privacy protection in the infinite time horizon. We have thus shown that Algorithm 1 can simultaneously ensure accurate learning and rigorous ϵ_i-LDP for each learner. This is fundamentally different from existing DP solutions for distributed learning and optimization [40]–[45], which allow the cumulative privacy budget to grow to infinity, implying diminishing privacy protection as the number of iterations tends to infinity.

Remark 10. A key reason for Algorithm 1 to ensure a finite cumulative privacy budget in the infinite time horizon under diminishing noise variances is that our algorithm design leads to diminishing sensitivity. Specifically, Lemma 10 implies that when the cumulative privacy budget $\sum_{t=1}^{\infty}\left(\frac{\Delta^i_{t,\theta}}{\nu^i_{t,\vartheta}}+\frac{\Delta^i_{t,s}}{\nu^i_{t,\zeta}}\right)$ is bounded (where $\Delta^i_{t,\theta}$ and $\Delta^i_{t,s}$ are the sensitivities and $\nu^i_{t,\vartheta}$ and $\nu^i_{t,\zeta}$ are the parameters of DP-noise variances for θ^i_t and s^i_t, respectively), learner i's implementation of an iterative algorithm is ϵ_i-locally differentially private in the infinite time horizon. According to Eq. (25), our algorithm design ensures that the sensitivities $\Delta^i_{t,\theta}$ and $\Delta^i_{t,s}$ (both of which are on the order of $\mathcal{O}(t^{-(1+v)})$) decay faster than the DP-noise variances $\nu^i_{t,\vartheta}$ and $\nu^i_{t,\zeta}$ (on the order of $\mathcal{O}(t^{-\varsigma^i_\vartheta})$ and $\mathcal{O}(t^{-\varsigma^i_\zeta})$, respectively). More specifically, our design ensures $\sum_{t=1}^{\infty}\left(\frac{\Delta^i_{t,\theta}}{\nu^i_{t,\vartheta}}+\frac{\Delta^i_{t,s}}{\nu^i_{t,\zeta}}\right)\leq\sum_{t=1}^{\infty}\mathcal{O}(t^{-(1+v-\varsigma^i_\vartheta)})+\mathcal{O}(t^{-(1+v-\varsigma^i_\zeta)})<\infty$ by requiring the parameters to satisfy $1+v-\max\{\varsigma^i_\vartheta,\varsigma^i_\zeta\}>1$. Therefore, we can ensure that the cumulative privacy budget is always finite.

Remark 11. Theorem 3 proves that our algorithm can circumvent the tradeoff between privacy and learning accuracy. A key enabler for our algorithm to resolve this tradeoff is to use diminishing stepsizes and DP-noise variances. Specifically, if we use a constant stepsize (i.e., making $v=0$), the cumulative privacy budget in Eq. (26) will grow to infinity since $1-\max\{\varsigma^i_\vartheta,\varsigma^i_\zeta\}<1$ holds. Furthermore, if we use constant noise variances (i.e., $\mathbb{E}[\|\zeta^i_t\|_2^2]=\mathbb{E}[\|\vartheta^i_t\|_2^2]=\sigma^2$), although a finite cumulative privacy budget can be achieved in the infinite time horizon, a steady-state optimization error (on the order of $m\sigma^2$) will appear in both $\mathbb{E}[\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\|_2^2]$ and $\mathbb{E}[\|\boldsymbol{s}_{t+1}-\boldsymbol{s}_t\|_2^2]$ (see Eq. (7b) and Eq. (8)), making it impossible for our algorithm to converge in mean square to an exact optimal solution.

Remark 12. Compared with the privacy analysis in our prior work [48] for undirected graphs, which only involves a single optimization variable, the privacy analysis here is much more complicated due to the involvement of two optimization variables s^i_t and θ^i_t, whose dynamics are strongly coupled.

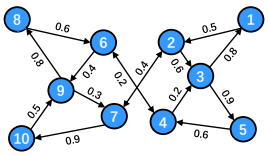


Fig. 1. The interaction graph $\mathcal{G}_R$ of ten learners

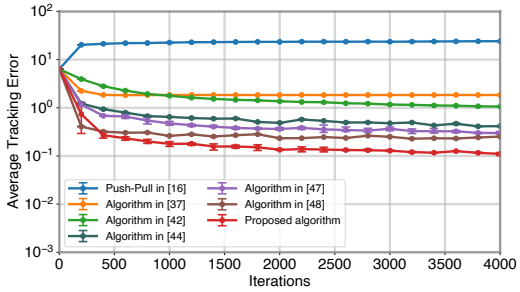


Fig. 2. Comparison of online logistic regression results by using the "mushrooms" dataset. The error bar represents standard derivation.

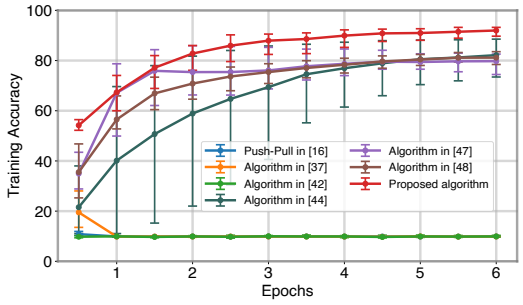
Remark 13. From Theorem 1 and Theorem 3, one can see that under a given sequence λ_t, if noise parameter sequences $\nu^i_{t,\vartheta}$ and $\nu^i_{t,\zeta}$ ensure a differential-privacy level of ϵ_i, then scaling the sequences $\nu^i_{t,\vartheta}$ and $\nu^i_{t,\zeta}$ by a constant $\frac{1}{c}>0$ can achieve any desired level of $c\epsilon_i$-LDP without losing provable convergence.

VI. Numerical Experiments

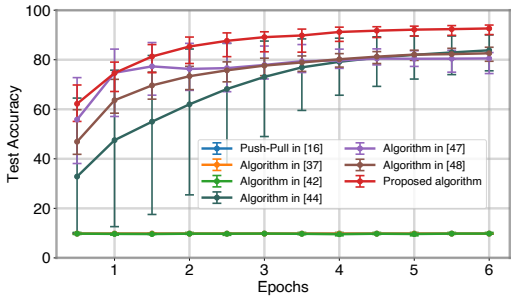
We evaluated the performance of Algorithm 1 through three machine-learning applications: linear regression using the "mushrooms" dataset and image classification using the "MNIST" and "CIFAR-10" datasets, respectively. In each experiment, we compared Algorithm 1 with existing DP solutions for distributed learning and optimization, including the DiaDSP algorithm [37], the DP-oriented gradient-tracking-based algorithm [47], the distributed online stochastic subgradient algorithm [42], the distributed online optimization algorithm [44], and the distributed online learning algorithm [48]. For a fair comparison, we set the privacy budget for these algorithms as the maximum ϵ_i across all learners used in our Algorithm 1, which corresponds to the weakest level of privacy protection among all learners. Additionally, we evaluated the conventional Push-Pull gradient-tracking algorithm [16] (i.e., algorithm (5)) under the same DP noises as those used in Algorithm 1. The interaction pattern associated with the weight matrix R was consistent across all experiments and is depicted in Fig. 1. The weight matrix C was set as the transpose of R.

A. Logistic regression using the "mushrooms" dataset

We first evaluated the performance of Algorithm 1 using l_2-logistic regression based classification of the "mushrooms" dataset [60]. The loss function for learner i is given by $l(\theta,\xi^i_t)=\frac{1}{N^i_t}\sum_{s=1}^{N^i_t}(1-b^i_s(a^i_s)^T\theta-\log(s((a^i_s)^T\theta))+\frac{r^i_t}{2}\|\theta\|^2$, where N^i_t is the number of samples at time t, $s(q)=(1+e^{-q})^{-1}$ is the sigmoid function defined, (a^i_s,b^i_s) is learner i's data point, and $r^i_t>0$ is a regularization parameter proportional

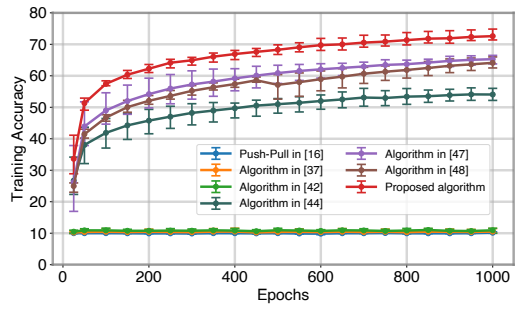


(a) Training accuracy

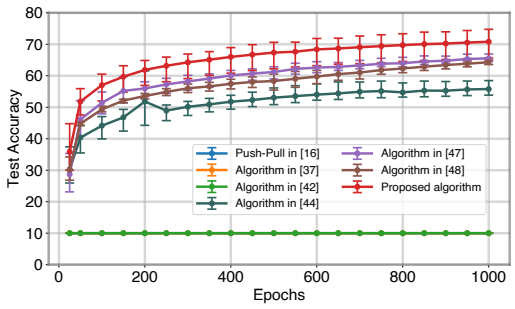


(b) Test accuracy

Fig. 3. Comparison of neural network training results by using the "MNIST" dataset.



(a) Training accuracy



(b) Test accuracy

Fig. 4. Comparison of neural network training results by using the "CIFAR-10" dataset.

to N_t^i. In each iteration, we randomly selected 10 samples and distributed them among the 10 learners.

In this experiment, we set the stepsize as $\lambda_t = (t+1)^{-0.61}$ and the DP-noise variances as $\nu_{t,\vartheta}^i = (t+1)^{-\varsigma_\vartheta^i}$ and $\nu_{t,\zeta}^i = (t+1)^{-\varsigma_\zeta^i}$ with $\varsigma_\vartheta^i = \varsigma_\zeta^i = 0.5+0.01i$ for $i = 1, 2, \cdots, 10$. The optimal solution θ^* was obtained using a noise-free centralized gradient descent algorithm. In our comparison, we employed the same stepsize and DP noises for the conventional Push-Pull gradient-tracking algorithm [16]. For other algorithms, we selected near-optimal stepsizes, ensuring that doubling the stepsizes would lead to non-convergent behaviors for these algorithms. In particular, the weakening factor for the algorithm in [48] was set to $\gamma_t = (t+1)^{-0.7}$, in accordance with the guidelines provided in [48]. The weakening factors for the algorithm in [47] were set to $\gamma_{1,t} = (t+1)^{-0.95}$ and $\gamma_{2,t} = (t+1)^{-0.75}$, in line with the guidelines provided in [47].

Fig. 2 shows the evolution of the average tracking errors $\frac{1}{10}\sum_{i=1}^{10}\|\theta_t^i - \theta^*\|$ [see Theorem 1]. Clearly, our Algorithm 1 outperforms existing results in terms of optimization accuracy. Moreover, it can be seen that the DP noise indeed accumulates in the conventional Push-Pull algorithm in [16], leading to non-convergent learning results [see Sec. III-A].

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

In the second experiment, we evaluated Algorithm 1 by training a convolutional neural-network (CNN) ResNet-18 on the "MNIST" dataset [61]. During each iteration, each learner was trained on 40 randomly selected images. In this experiment, we chose the stepsize as $\lambda_t = 0.6(t+1)^{-0.61}$ and the DP-noise variances as $\nu_{t,\vartheta}^i = 0.01(t+1)^{-\varsigma_\vartheta^i}$ and $\nu_{t,\zeta}^i = 0.01(t+1)^{-\varsigma_\zeta^i}$ with $\varsigma_\vartheta^i = \varsigma_\zeta^i = 0.5 + 0.01i$ for $i = 1, 2, \cdots, 10$. We used the best stepsizes that we could find for the existing algorithms used in the comparison. The weakening factors for the algorithms in [48] and [47] remained consistent with those employed in the previous logistic regression experiment.

Fig. 3 shows that the conventional Push-Pull algorithm [16], the algorithm in [37], and the algorithm in [42] are incapable of effectively training the CNN model under DP-noise injections. Moreover, our Algorithm 1 has better training and test accuracies than the differentially private distributed online optimization algorithm [44], the DP-oriented gradient-tracking-based algorithm [47], and the algorithm in [48].

To compare the strength of enabled privacy protection, we ran the DLG attack model proposed in [28], which is a powerful inference algorithm capable of reconstructing raw data from shared gradient/model updates. The training/test accuracies and the DLG attacker's inference errors for all compared algorithms are summarized in Table I. It can be seen that our algorithm can provide stronger privacy protection (i.e., a higher final DLG inference error) and better training/test accuracies than existing counterparts.

TABLE I
COMPARISON OF TRAINING AND TEST ACCURACIES AND DLG ATTACKER'S INFERENCE ERRORS

Algorithms	Training accuracy (%)	Test accuracy(%)	Final DLG error
Push-Pull in [16]	9.74	9.80	8.14
Algorithm in [37]	9.18	9.80	2.42
Algorithm in [42]	9.44	9.80	5.11
Algorithm in [44]	82.15	83.80	7.43
Algorithm in [47]	90.48	92.22	14.42
Algorithm in [48]	85.11	86.07	15.15
Proposed algorithm	91.98	92.62	14.85

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

The third experiment evaluated Algorithm 1 using a CNN model and the "CIFAR-10" dataset [62], which provides a greater diversity and complexity than the "MNIST" dataset. The CNN architecture and parameters were the same as those used in the previous experiment on the "MNIST" dataset.

The results are summarized in Fig. 4, which once again confirms the advantage of our proposed algorithm over existing counterparts in terms of both training and test accuracies.

VII. Conclusions

In this study, we proposed a distributed learning algorithm under the constraints of differential privacy and sequentially arriving data. We proved that the proposed algorithm converges in mean square to the exact optimal solution, even in the presence of DP noises and general directed graphs. Simultaneously, we also proved that the proposed algorithm can ensure rigorous ϵ_i-LDP with a finite cumulative privacy budget, even when the number of iterations grows to infinity. To the best of our knowledge, this is the first algorithm that is able to simultaneously achieve provable convergence and rigorous ϵ_i-LDP (with a finite cumulative privacy budget) in distributed online learning over directed graphs. Experimental comparisons using multiple benchmark machine-learning applications confirm the advantage of our proposed algorithm over existing counterparts.

Appendix

For the convenience of derivation, we define $\bar{\zeta}_{C,t}=\frac{\mathbf{1}^T\boldsymbol{\zeta}_{C,t}}{m}$, $\bar{\vartheta}_{R,t}=\frac{u^T\boldsymbol{\vartheta}_{R,t}}{m}$, $\nabla\bar{f}_t(\boldsymbol{\theta}_t)=\frac{\mathbf{1}^T\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)}{m}$, $\nabla\bar{f}(\boldsymbol{\theta}_t)=\frac{\mathbf{1}^T\nabla\boldsymbol{f}(\boldsymbol{\theta}_t)}{m}$, $U=\mathrm{diag}(u_1,\cdots,u_m)$, $\bar{\mathbf{C}}=\mathbf{C}-\frac{\omega\mathbf{1}\mathbf{1}^T}{m}$, $\bar{\mathbf{R}}=\mathbf{R}-\frac{\mathbf{1}u^T}{m}$, $\Pi_\omega=I-\frac{\omega\mathbf{1}\mathbf{1}^T}{m}$, $\Pi_u=I-\frac{\mathbf{1}u^T}{m}$, $\Pi_U=U^{-1}-\frac{\mathbf{1}\mathbf{1}^T}{m}$, $\Pi_U^e=(I-\frac{\mathbf{1}u^T}{m})(Z_t^{-1}-U^{-1})$, and $\sigma_\zeta^+=\max_{i\in[m]}\{\sigma_\zeta^i\}$, $\sigma_\vartheta^+=\max_{i\in[m]}\{\sigma_\vartheta^i\}$. We denote $\langle\cdot,\cdot\rangle_C$ and $\langle\cdot,\cdot\rangle_R$ by the inner products induced by the norm $\|\cdot\|_C$ and $\|\cdot\|_R$, respectively.

A. Proof of Lemma 2

Using the relationship $F_t(\theta_t^*)\le F_t(\theta^*)$ and the mean value theorem, we obtain

$$F(\theta_t^*)-F(\theta^*)\le\|\nabla F(\chi)-\nabla F_t(\chi)\|_2\|\theta_t^*-\theta^*\|_2, \quad (27)$$

with $\chi=\alpha\theta_t^*+(1-\alpha)\theta^*$ for some constant $\alpha\in(0,1)$.

The definitions of $\nabla F_t(\cdot)$ and $\nabla F(\cdot)$ imply

$$\begin{aligned}&\mathbb{E}[\|\nabla F_t(\chi)-\nabla F(\chi)\|_2]\\&\le\tfrac{1}{m}\textstyle\sum_{i=1}^m\tfrac{1}{t+1}\sum_{k=0}^t\mathbb{E}[\|\nabla l(\chi,\xi_k^i)-\mathbb{E}[\nabla l(\chi,\xi_k^i)]\|_2].\end{aligned} \quad (28)$$

Given that the data points ξ_k^i are *i.i.d.* across iterations, we use Assumption 3(ii) to obtain

$$\textstyle\sum_{k=0}^t\mathbb{E}\left[\|\nabla l(\chi,\xi_k^i)-\mathbb{E}[\nabla l(\chi,\xi_k^i)]\|_2\right]\le\kappa\sqrt{t+1}. \quad (29)$$

Substituting (29) into (28) yields $\mathbb{E}[\|\nabla F_t(\chi)-\nabla F(\chi)\|_2]\le\frac{\kappa}{\sqrt{t+1}}$. By using (27), we have

$$\mathbb{E}\left[F(\theta_t^*)-F(\theta^*)\right]\le\tfrac{\kappa}{\sqrt{t+1}}\mathbb{E}\left[\|\theta_t^*-\theta^*\|_2\right]. \quad (30)$$

Assumption 2(iii) with $\mu>0$ implies $\frac{\mu}{2}\|\theta_t^*-\theta^*\|_2^2\le F(\theta_t^*)-F(\theta^*)$. Combing this relation and (30), we arrive at $\frac{\mu}{2}\mathbb{E}[\|\theta_t^*-\theta^*\|_2^2]\le\frac{\kappa}{\sqrt{t+1}}\mathbb{E}[\|\theta_t^*-\theta^*\|_2]$, which implies (4).

B. Proof of Lemma 8

Left multiplying both sides of (7a) by $\frac{1}{m}\mathbf{1}^T$ and using the relation $\mathbf{1}^TC=\mathbf{0}$, we obtain $\bar{s}_{t+1}=\frac{\mathbf{1}^T}{m}(\boldsymbol{s}_t+\boldsymbol{\zeta}_{C,t}+\lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t))$. Combing this relation with (7a) and $\bar{\mathbf{C}}\omega=\mathbf{0}$ leads to

$$\boldsymbol{s}_{t+1}-\omega\bar{s}_{t+1}=\bar{\mathbf{C}}(\boldsymbol{s}_t-\omega\bar{s}_t)+\Pi_\omega\boldsymbol{\zeta}_{C,t}+\lambda_t\Pi_\omega\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t), \quad (31)$$

where we have used the definition of Π_ω. Using the definition $\|\bar{\mathbf{C}}\|_C=1-\rho_C<1$ and the inequality $(a+b)^2\le(1+\epsilon)a^2+(1+\epsilon^{-1})b^2$ for any scalars a, b, and $\epsilon>0$ (setting $\epsilon=\frac{1}{1-\rho_C}-1$, implying $1+\epsilon^{-1}=\frac{1}{\rho_C}$), we obtain

$$\mathbb{E}[\|\boldsymbol{s}_{t+1}-\omega\bar{s}_{t+1}\|_C^2]\le(1-\rho_C)\mathbb{E}[\|\boldsymbol{s}_t-\omega\bar{s}_t\|_C^2]+\Phi_{t,s}, \quad (32)$$

where the term $\Phi_{t,s}$ is given by

$$\Phi_{t,s}=\tfrac{\lambda_t^2}{\rho_C}\|\Pi_\omega\|_C^2\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|_C^2]+\|\Pi_\omega\|_C^2\mathbb{E}[\|\boldsymbol{\zeta}_{C,t}\|_C^2]. \quad (33)$$

We proceed to characterize the $\Phi_{t,s}$ in (33). By using the definition of $\nabla f_i(\theta_t^i)$, Assumptions 2(ii), and 3(ii), we have

$$\mathbb{E}[\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|_C^2]\le 2m\delta_{C,F}^2(\kappa^2+D^2). \quad (34)$$

Substituting the relation $\mathbb{E}[\|\boldsymbol{\zeta}_{C,t}\|_C^2]\le m\delta_{C,F}^2\sum_{i,j}(C_{ij}\sigma_{t,\zeta}^j)^2$ and (34) into (33) yields

$$\Phi_{t,s}=\tfrac{\tau_{s1}}{(t+1)^{2v}}+\tfrac{\tau_{s2}}{(t+1)^{2\varsigma_\zeta}}, \quad (35)$$

where τ_{s1} and τ_{s2} are given by $\tau_{s1}=\frac{2m\delta_{C,F}^2\|\Pi_\omega\|_C^2\|(\kappa^2+D^2)\lambda_0^2}{\rho_C}$ and $\tau_{s2}=m\delta_{C,F}^2\|\Pi_\omega\|_C^2\sum_{i,j}(C_{ij})^2(\sigma_\zeta^+)^2$, respectively.

Now, we iterate (32) from 0 to t to obtain

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{s}_{t+1}-\omega\bar{s}_{t+1}\|_C^2]&\le(1-\rho_C)^{t+1}\mathbb{E}[\|\boldsymbol{s}_0-\omega\bar{s}_0\|_C^2]\\&\quad+\textstyle\sum_{p=0}^{t-1}(1-\rho_C)^{t-p}\Phi_{p,s}+\Phi_{t,s}.\end{aligned} \quad (36)$$

When $t=0$, we have $\mathbb{E}[\|\boldsymbol{s}_1-\omega\bar{s}_1\|_C^2]\le(1-\rho_C)\mathbb{E}[\|\boldsymbol{s}_0-\omega\bar{s}_0\|_C^2]+\tau_{s1}+\tau_{s2}$. When $t>0$, we estimate each item on the right hand side of (36).

1) Lemma 7 and $(t+1)^{-2}<t^{-2}$ for all $t>0$ imply

$$(1-\rho_C)^{t+1}\mathbb{E}[\|\boldsymbol{s}_0-\omega\bar{s}_0\|_C^2]<c_{s0}t^{-2}\mathbb{E}[\|\boldsymbol{s}_0-\omega\bar{s}_0\|_C^2], \quad (37)$$

with the constant $c_{s0}=\frac{4}{(e\ln(1-\rho_C))^2}$.

2) For scalars $a,b,c,d>0$ satisfying $\frac{c}{a}>\frac{d}{b}$, the relationship $\frac{d}{b}<\frac{c+d}{a+b}<\frac{c}{a}$ always holds. This inequality implies $\frac{1}{(t-p)^2}<(\frac{p+1}{t})^2$ for all $p\in[0,t)$. Using this inequality, Lemma 7, and the relation $(\frac{p+1}{t})^2<(\frac{p+1}{t})^{2v}$ (where $\frac{p+1}{t}\in(0,1]$), we obtain

$$\tfrac{(1-\frac{\rho_C}{2})^{t-p}}{(p+1)^{2v}}\le\tfrac{4}{(e\ln(1-\frac{\rho_C}{2}))^2(t-p)^2}\tfrac{1}{(p+1)^{2v}}=\tfrac{4t^{-2v}}{(e\ln(1-\frac{\rho_C}{2}))^2}. \quad (38)$$

By using inequality (38), $1-\rho_C\le(1-\frac{\rho_C}{2})^2$, and $\sum_{p=0}^{t-1}(1-\frac{\rho_C}{2})^{t-p}<\frac{1-(1-\frac{\rho_C}{2})^t}{1-(1-\frac{\rho_C}{2})}\le\frac{2}{\rho_C}$, we obtain

$$\textstyle\sum_{p=0}^{t-1}\frac{(1-\rho_C)^{t-p}}{(p+1)^{2v}}<\sum_{p=0}^{t-1}\left(1-\frac{\rho_C}{2}\right)^{t-p}\frac{4t^{-2v}}{(e\ln(1-\frac{\rho_C}{2}))^2}. \quad (39)$$

Using an argument similar to the derivation of (39) yields

$$\textstyle\sum_{p=0}^{t-1}(1-\rho_C)^{t-p}\Phi_{p,s}<\bar{c}_{s0}(t^{-2v}+t^{-2\varsigma_\zeta}), \quad (40)$$

with the constant $\bar{c}_{s0}=\frac{8}{\rho_C(e\ln(1-\frac{\rho_C}{2}))^2}$.

By substituting (37) and (40), and the relationship $\Phi_{t,s}\le$

$\tau_{s1}t^{-2v}+\tau_{s2}t^{-2\varsigma_\zeta}$ into (36), we arrive at

$$\mathbb{E}[\|\boldsymbol{s}_t-\omega\bar{s}_t\|_C^2]<c_{\boldsymbol{s}1}t^{-2}+c_{\boldsymbol{s}2}t^{-2v}+c_{\boldsymbol{s}3}t^{-2\varsigma_\zeta}, \tag{41}$$

where the constants $c_{\boldsymbol{s}1}$, $c_{\boldsymbol{s}2}$, and $c_{\boldsymbol{s}3}$ are given by $c_{\boldsymbol{s}1}=\max\{1-\rho_C,c_{s0}\}\mathbb{E}[\|\boldsymbol{s}_0-\omega\bar{s}_0\|_C^2]$, $c_{\boldsymbol{s}2}=\bar{c}_{s0}+\tau_{s1}$, and $c_{\boldsymbol{s}3}=\bar{c}_{s0}+\tau_{s2}$ with τ_{s1} and τ_{s2} given in (35) and c_{s0} and $\bar{c}_{s0}$ given in (37) and (39), respectively.

C. Proof of Lemma 9

Left multiplying both sides of (7b) by $\frac{u^T}{m}$ and using the relations $u^TU^{-1}=\mathbf{1}^T$ and $u^T\mathbf{R}=u^T$, we obtain

$$\bar{\theta}_{t+1}=\tfrac{u^T\boldsymbol{\theta}_t}{m}+\tfrac{u^T\boldsymbol{\vartheta}_{R,t}}{m}-\left(\tfrac{\mathbf{1}^T}{m}+\tfrac{u^T(Z_t^{-1}-U^{-1})}{m}\right)(\boldsymbol{s}_{t+1}-\boldsymbol{s}_t). \tag{42}$$

Based on dynamics (7a) and the relation $Cv=\mathbf{0}$, we have

$$\boldsymbol{s}_{t+1}-\boldsymbol{s}_t=C(\boldsymbol{s}_t-\omega\bar{s}_t)+\boldsymbol{\zeta}_{C,t}+\lambda_t\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t). \tag{43}$$

Substituting (43) into (42) and using the relationships $\mathbf{1}^TC=\mathbf{0}^T$ and $\bar{\mathbf{R}}\mathbf{1}=(I+R-\frac{\mathbf{1}u^T}{m})\mathbf{1}=\mathbf{0}$ lead to

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1}-\mathbf{1}\bar{\theta}_{t+1}\|_R^2\le&\left\|\Pi_u\boldsymbol{\vartheta}_{R,t}-(\Pi_U+\Pi_U^e)\boldsymbol{\zeta}_{C,t}\right\|_R^2\\&+\big[\|\bar{\mathbf{R}}\|_R\|\boldsymbol{\theta}_t-\mathbf{1}\bar{\theta}_t\|_R+(\|\Pi_UC\|_R+\|\Pi_U^eC\|_R)\|\boldsymbol{s}_t-\omega\bar{s}_t\|_R\\&+\lambda_t(\|\Pi_U\|_R+\|\Pi_U^e\|_R)\|\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t)\|_R\big]^2\\&+2\left\langle\bar{\mathbf{R}}(\boldsymbol{\theta}_t-\mathbf{1}\bar{\theta}_t)-(\Pi_U+\Pi_U^e)C(\boldsymbol{s}_t-\omega\bar{s}_t)\right.\\&\left.-\lambda_t(\Pi_U+\Pi_U^e)\nabla\boldsymbol{f}_t(\boldsymbol{\theta}_t),\Pi_u\boldsymbol{\vartheta}_{R,t}-(\Pi_U+\Pi_U^e)\boldsymbol{\zeta}_{C,t}\right\rangle_R.\end{aligned}$$

Using an argument similar to the derivation of (36), we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_{t+1}-\mathbf{1}\bar{\theta}_{t+1}\|_R^2]\le&(1-\rho_R)^{t+1}\mathbb{E}[\|\boldsymbol{\theta}_0-\mathbf{1}\bar{\theta}_0\|_R^2]\\&+\textstyle\sum_{p=0}^{t-1}(1-\rho_R)^{t-p}\Phi_{p,\theta}+\Phi_{t,\theta},\end{aligned} \tag{44}$$

where $\Phi_{t,\theta}$ is given by

$$\Phi_{t,\theta}=\tau_{\theta1}\mathbb{E}[\|\boldsymbol{s}_t-\omega\bar{s}_t\|_C^2]+\tfrac{\tau_{\theta2}}{(t+1)^{2v}}+\tfrac{\tau_{\theta3}}{(t+1)^{2\varsigma_\vartheta}}+\tfrac{\tau_{\theta4}}{(t+1)^{2\varsigma_\zeta}}, \tag{45}$$

with the positive constants $\tau_{\theta1}=\frac{4\delta_{R,C}^2\|C\|_R^2(\|\Pi_U\|_R^2+\|\Pi_U^e\|_R^2)}{\rho_R}$, $\tau_{\theta2}=\frac{2m\lambda_0^2\delta_{R,F}^2(\kappa^2+D^2)}{\delta_{R,C}^2\|C\|_R^2}$, $\tau_{\theta3}=2\|\Pi_u\|_R^2\delta_{R,F}^2\sum_{i,j}(R_{ij})^2(\sigma_\vartheta^+)^2$, and $\tau_{\theta4}=2\|\Pi_U+\Pi_U^e\|_R^2\delta_{R,F}^2\sum_{i,j}(C_{ij})^2(\sigma_\zeta^+)^2$.

When $t=0$, we have $\mathbb{E}[\|\boldsymbol{\theta}_1-\mathbf{1}\bar{\theta}_1\|_R^2]\le(1-\rho_R)\mathbb{E}[\|\boldsymbol{\theta}_0-\mathbf{1}\bar{\theta}_0\|_R^2]+\tau_{\theta1}\mathbb{E}[\|\boldsymbol{s}_0-\omega\bar{s}_0\|_C^2]+\tau_{\theta2}+\tau_{\theta3}+\tau_{\theta4}$. When $t>0$, we analyze each item on the right hand side of (44).

Using an argument similar to the derivation of (37), the first term on the right hand side of (44) satisfies

$$(1-\rho_R)^{t+1}\mathbb{E}[\|\boldsymbol{\theta}_0-\mathbf{1}\bar{\theta}_0\|_R^2]<c_{\theta0}t^{-2}\mathbb{E}[\|\boldsymbol{\theta}_0-\mathbf{1}\bar{\theta}_0\|_R^2], \tag{46}$$

with the constant $c_{\theta0}=\frac{4}{(e\ln(1-\rho_R))^2}$.

By using (9) and (45) and following an argument similar to the derivation of (40), we obtain

$$\begin{aligned}\textstyle\sum_{p=0}^{t-1}(1-\rho_R)^{t-p}\Phi_{p,\theta}<&\,c_{\theta0}\mathbb{E}[\|\boldsymbol{\theta}_0-\omega\bar{\theta}_0\|_C^2]t^{-2}\\&+\bar{c}_{\theta0}\tau_{\theta1}(c_{\boldsymbol{s}1}t^{-2}+c_{\boldsymbol{s}2}t^{-2v}+c_{\boldsymbol{s}3}t^{-2\varsigma_\zeta})\\&+\bar{c}_{\theta0}(\tau_{\theta2}t^{-2v}+\tau_{\theta3}t^{-2\varsigma_\vartheta}+\tau_{\theta4}t^{-2\varsigma_\zeta}),\end{aligned} \tag{47}$$

where the constants $\bar{c}_{\theta0}$ is given by $\bar{c}_{\theta0}=\frac{8}{\rho_R(e\ln(1-\frac{\rho_R}{2}))^2}$, $\tau_{\theta1}$ to $\tau_{\theta4}$ are given in (45), and $\bar{c}_{\theta0}$ is given in (46).

Substituting (46), (47), and (45) into (44) yields

$$\mathbb{E}\big[\|\boldsymbol{\theta}_t-\mathbf{1}\bar{\theta}_t\|_R^2\big]<c_{\boldsymbol{\theta}1}t^{-2}+c_{\boldsymbol{\theta}2}t^{-2v}+c_{\boldsymbol{\theta}3}t^{-2\varsigma_\vartheta}+c_{\boldsymbol{\theta}4}t^{-2\varsigma_\zeta}, \tag{48}$$

where the constant $c_{\boldsymbol{\theta}1}$, $c_{\boldsymbol{\theta}2}$, $c_{\boldsymbol{\theta}3}$, and $c_{\boldsymbol{\theta}4}$ are given by $c_{\boldsymbol{\theta}1}=\max\{1-\rho_R,c_{\theta0}\}\mathbb{E}[\|\boldsymbol{\theta}_0-\mathbf{1}\bar{\theta}_0\|_R^2]+\max\{1,c_{\theta0}\}\tau_{\theta1}\mathbb{E}[\|\boldsymbol{s}_0-\omega\bar{s}_0\|_C^2]+\tau_{\theta1}c_{\boldsymbol{s}1}(\bar{c}_{\theta0}+1)$, $c_{\boldsymbol{\theta}2}=(\tau_{\theta1}c_{\boldsymbol{s}2}+\tau_{\theta2})(\bar{c}_{\theta0}+1)$, $c_{\boldsymbol{\theta}3}=\tau_{\theta3}(\bar{c}_{\theta0}+1)$, and $c_{\boldsymbol{\theta}4}=(\tau_{\theta1}c_{\boldsymbol{s}3}+\tau_{\theta4})(\bar{c}_{\theta0}+1)$, respectively.

D. Proof of Theorem 1

Substituting (8) into (42) and using Assumption 4 yield

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_{t+1}-\theta_t^*\|_2^2]\le&\left(1+\tfrac{\lambda_t\mu}{2}\right)\mathbb{E}[\|\bar{\theta}_t-\lambda_t\nabla\bar{f}(\mathbf{1}\bar{\theta}_t)-\theta_t^*\|_2^2]\\&+\left(1+\tfrac{2}{\lambda_t\mu}\right)\mathbb{E}\big[\big\|\lambda_t\nabla\bar{f}(\mathbf{1}\bar{\theta}_t)-\lambda_t\nabla\bar{f}_t(\boldsymbol{\theta}_t)\\&-\tfrac{u^T(Z_t^{-1}-U^{-1})}{m}(\boldsymbol{s}_{t+1}-\boldsymbol{s}_t)\big\|_2^2\big]+\mathbb{E}[\|\bar{\vartheta}_{R,t}-\bar{\zeta}_{C,t}\|_2^2].\end{aligned} \tag{49}$$

The definition $\nabla\bar{f}(\mathbf{1}\bar{\theta}_t)=\frac{\mathbf{1}^T\nabla\boldsymbol{f}(\mathbf{1}\bar{\theta}_t)}{m}$ implies $\nabla\bar{f}(\mathbf{1}\bar{\theta}_t)=\nabla F(\bar{\theta}_t)$. Then, we have

$$\begin{aligned}\|\bar{\theta}_t-\lambda_t\nabla\bar{f}(\mathbf{1}\bar{\theta}_t)-\theta_t^*\|_2^2=&\,\|\bar{\theta}_t-\theta_t^*\|_2^2\\&-2\lambda_t\langle\nabla F(\bar{\theta}_t),\bar{\theta}_t-\theta_t^*\rangle+\lambda_t^2\|\nabla F(\bar{\theta}_t)\|_2^2.\end{aligned} \tag{50}$$

By using Assumption 2(iii) with $\mu\ge0$, we have $F(\theta_t^*)-F(\bar{\theta}_t)\ge-\nabla F(\theta_t^*)^T(\bar{\theta}_t-\theta_t^*)+\frac{\mu}{2}\|\bar{\theta}_t-\theta_t^*\|_2^2$. Combining this relationship and Assumption 2(ii) with (50), we arrive at

$$\begin{aligned}\|\bar{\theta}_t-\lambda_t\nabla\bar{f}(\mathbf{1}\bar{\theta}_t)-\theta_t^*\|_2^2\le&(1-\lambda_t\mu)\|\bar{\theta}_t-\theta_t^*\|_2^2\\&-2\lambda_t(F(\bar{\theta}_t)-F(\theta_t^*))+\lambda_t^2D^2.\end{aligned} \tag{51}$$

Using the mean value theorem and (4) in Lemma 2, one has

$$\mathbb{E}[F(\bar{\theta}_t)-F(\theta_t^*)]\ge\mathbb{E}[F(\theta^*)-F(\theta_t^*)]\ge-\tfrac{2D\kappa}{\mu\sqrt{t+1}}. \tag{52}$$

Using Assumption 3(ii)-(iii) and the definitions of $\nabla\bar{f}_t(\boldsymbol{\theta}_t)$, $\nabla\bar{f}(\boldsymbol{\theta}_t)$, and $\nabla f_i(\theta_t^i)$, we have

$$\mathbb{E}[\|\nabla\bar{f}(\mathbf{1}\bar{\theta}_t)-\nabla\bar{f}_t(\boldsymbol{\theta}_t)\|_2^2]\le\tfrac{2\kappa^2}{t+1}+2\delta_{F,R}^2L^2\mathbb{E}\left[\|\boldsymbol{\theta}_t-\mathbf{1}\bar{\theta}_t\|_R^2\right]. \tag{53}$$

By taking the norm $\|\cdot\|_F$ on both sides of (43) and then using an argument similar to the derivation of (34), we have

$$\begin{aligned}\|\boldsymbol{s}_{t+1}-\boldsymbol{s}_t\|_F^2\le&\,3\delta_{F,C}^2\|C\|_C^2\|\boldsymbol{s}_t-\omega\bar{s}_t\|_C^2\\&+3\|\boldsymbol{\zeta}_{C,t}\|_F^2+6m(\kappa^2+D^2)\lambda_t^2.\end{aligned} \tag{54}$$

Incorporating (51)-(54) into (49) and then combining (49) and the inequality $\|\bar{\theta}_{t+1}-\theta_{t+1}^*\|_2^2\le(1+\frac{\lambda_t\mu}{4})\|\bar{\theta}_{t+1}-\theta_t^*\|_2^2+(1+\frac{4}{\lambda_t\mu})\|\theta_{t+1}^*-\theta_t^*\|_2^2$, one obtains

$$\mathbb{E}[\|\bar{\theta}_{t+1}-\theta_{t+1}^*\|_2^2]\le\left(1-\tfrac{\lambda_t\mu}{4}\right)\mathbb{E}[\|\bar{\theta}_t-\theta_t^*\|_2^2]+\Phi_{t,\bar{\theta}}, \tag{55}$$

where $\Phi_{t,\bar{\theta}}$ is given by

$$\begin{aligned}\Phi_{t,\bar{\theta}}=&\,\hat{c}_{\bar{\theta}0}\big(\tfrac{2\lambda_t\kappa^2}{t+1}+2\lambda_t\delta_{F,R}^2L^2\mathbb{E}[\|\boldsymbol{\theta}_t-\mathbf{1}\bar{\theta}_t\|_R^2]\\&+\tfrac{3\delta_{F,C}^2\|C\|_C^2\|u\|_2^2c_z^2}{m^2}\tfrac{\gamma_z^{2t}}{\lambda_t}\mathbb{E}[\|\boldsymbol{s}_t-\omega\bar{s}_t\|_C^2]\\&+\tfrac{3\|u\|_2^2c_z^2}{m^2}\tfrac{\gamma_z^{2t}}{\lambda_t}\mathbb{E}[\|\boldsymbol{\zeta}_{C,t}\|_F^2]+\tfrac{6(\kappa^2+D^2)\|u\|_2^2c_z^2}{m}\gamma_z^{2t}\lambda_t\big)\\&+2\big(1+\tfrac{\lambda_0\mu}{4}\big)\mathbb{E}[\|\bar{\vartheta}_{R,t}\|_2^2+\|\bar{\zeta}_{C,t}\|_2^2]+\tfrac{\hat{c}_{\bar{\theta}0}\mu}{4}\big(\tfrac{4D\kappa}{\mu}\tfrac{\lambda_t}{\sqrt{t+1}}\\&+D^2\lambda_t^2\big)+\big(1+\tfrac{4}{\lambda_t\mu}\big)\mathbb{E}[\|\theta_{t+1}^*-\theta_t^*\|_2^2],\end{aligned} \tag{56}$$

with $\hat{c}_{\bar{\theta}0}\triangleq\frac{\lambda_0^2\mu^2+6\lambda_0\mu+8}{2\mu}$.

By iterating (55) from 0 to t and using the relation $\prod_{p=0}^t(1-\frac{\lambda_p\mu}{4})\le e^{-\frac{\mu}{4}\sum_{p=0}^t\lambda_p}$, we arrive at

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_{t+1}-\theta_{t+1}^*\|_2^2]\le&\,e^{-\frac{\mu}{4}\sum_{p=0}^t\lambda_p}\mathbb{E}\left[\|\bar{\theta}_0-\theta_0^*\|_2^2\right]\\&+\textstyle\sum_{p=1}^t\Phi_{p-1,\bar{\theta}}e^{-\frac{\mu}{4}\sum_{q=p}^t\lambda_q}+\Phi_{t,\bar{\theta}}.\end{aligned} \tag{57}$$

We estimate the first term on the right hand side of (57).

Since $\frac{\lambda_0}{(p+1)^v} \geq \frac{\lambda_0}{(t+1)^v}$ holds for all $t \geq p$ and $(t+1)^v \leq 2^v t^v$ holds for all $t > 0$, we have $\sum_{p=0}^{t} \lambda_p \geq \frac{\lambda_0}{(t+1)^v}(t+1) \geq \frac{\lambda_0}{2^v t^{v-1}}$, which implies $e^{\frac{\mu}{4}\sum_{p=0}^{t}\lambda_p} \geq e^{\frac{\mu}{4}\frac{\lambda_0}{2^v t^{v-1}}}$. Using Taylor expansion $e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!}$, we have that there must exist some $n_0 \in \mathbb{N}^+$ such that $e^x \geq \frac{x^{n_0}}{n_0!}$ holds when x is nonnegative. Setting $n_0 \triangleq \lceil \frac{1}{1-v} \rceil$, we have $(1-v)n_0 \geq 1$, which implies

$$e^{\frac{\mu}{4}\sum_{p=0}^{t}\lambda_p} \geq \frac{1}{n_0!}\Big(\frac{\mu\lambda_0}{4\times 2^v}\Big)^{n_0} t^{(1-v)n_0} \geq \frac{\Big(\frac{\mu\lambda_0}{4\times 2^v}\Big)^{\frac{1}{1-v}} t}{(\frac{1}{1-v}+1)!}. \quad (58)$$

Substituting (58) into the first term on the right hand side of (57), we arrive at

$$e^{-\frac{\mu}{4}\sum_{p=0}^{t}\lambda_p}\mathbb{E}[\|\bar{\theta}_0 - \theta_0^*\|_2^2] \leq c_{\bar{\theta}1} t^{-1}, \quad (59)$$

where $c_{\bar{\theta}1}$ is given by $c_{\bar{\theta}1} = (\frac{2-v}{1-v})!\big(\frac{\mu\lambda_0}{4\times 2^v}\big)^{\frac{1}{v-1}}\mathbb{E}[\|\bar{\theta}_0 - \theta_0^*\|_2^2]$.

We proceed to analyze the second and third terms on the right hand side of (57). We select a constant $\alpha \in (v, \frac{1+v}{2})$. Since $e^{-\frac{\mu}{4}\sum_{q=\lceil t-t^\alpha\rceil+1}^{t}\lambda_q} < 1$ is valid and $e^{-\frac{\mu}{4}\sum_{q=p}^{t}\lambda_q} \leq e^{-\frac{\mu}{4}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q}$ holds for all $p \in [1, \lceil t-t^\alpha\rceil]$, we obtain

$$\begin{aligned}&\textstyle\sum_{p=1}^{t}\Phi_{p-1,\bar{\theta}} e^{-\frac{\mu}{4}\sum_{q=p}^{t}\lambda_q} + \Phi_{t,\bar{\theta}}\\ &\leq \textstyle\sum_{p=0}^{\lfloor t-t^\alpha\rfloor}\Phi_{p,\bar{\theta}} e^{-\frac{\mu}{4}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} + \sum_{p=\lceil t-t^\alpha\rceil}^{t}\Phi_{p,\bar{\theta}}.\end{aligned} \quad (60)$$

We now analyze the first term on the right hand side of (60). To this end, we first characterize the term $e^{-\frac{\mu}{4}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q}$. Given $\frac{1}{(q+1)^v} \geq \frac{1}{(t+1)^v}$ for all $q \in [\lceil t-t^\alpha\rceil, t]$, we have

$$\textstyle\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q \geq \frac{\lambda_0}{(t+1)^v}(t - \lceil t-t^\alpha\rceil + 1) \geq \frac{\lambda_0 t^{\alpha-v}}{2^v}, \quad (61)$$

where we have used $\lceil t-t^\alpha\rceil \leq t-t^\alpha+1$ and $(t+1)^v \leq 2^v t^v$.

Using an argument similar to the derivation of (58), we set $n_0 \triangleq \lceil \frac{1}{\alpha-v}\rceil$ (i.e., $(\alpha-v)n_0 \geq 1$) for the Taylor expansion to obtain $e^{\frac{\mu}{4}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} \geq \frac{1}{(\frac{1}{\alpha-v}+1)!}(\frac{\mu\lambda_0}{4\times 2^v})^{\frac{1}{\alpha-v}} t$. Then, the first term on the right hand side of (60) satisfies

$$\textstyle\sum_{p=0}^{\lfloor t-t^\alpha\rfloor}\Phi_{p,\bar{\theta}} e^{-\frac{\mu}{4}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} < \Big(\Phi_{0,\bar{\theta}} + \sum_{p=1}^{\infty}\Phi_{p,\bar{\theta}}\Big) c t^{-1}, \quad (62)$$

where the constant c is given by $c = (\frac{\alpha-v+1}{\alpha-v})!\left(\frac{\mu\lambda_0}{4\times 2^v}\right)^{\frac{1}{v-\alpha}}$.

To proceed, we need to estimate an upper bound on $\Phi_{t,\bar{\theta}}$ in (56). To this end, we first prove the following relations:

1) By using (10) and $t^{-2v} \leq 4^v(t+1)^{-2v}$, we have

$$\begin{aligned}\lambda_t\mathbb{E}\left[\|\boldsymbol{\theta}_t - \mathbf{1}\bar{\theta}_t\|_R^2\right] < \lambda_0\Big(&\tfrac{4c_{\boldsymbol{\theta}1}}{(t+1)^{v+2}} + \tfrac{4^v c_{\boldsymbol{\theta}2}}{(t+1)^{3v}}\\ &+ \tfrac{4^{\varsigma_\vartheta} c_{\boldsymbol{\theta}3}}{(t+1)^{v+2\varsigma_\vartheta}} + \tfrac{4^{\varsigma_\zeta} c_{\boldsymbol{\theta}4}}{(t+1)^{v+2\varsigma_\zeta}}\Big).\end{aligned} \quad (63)$$

2) Lemma 7 implies $\gamma_z^{2t} \leq \frac{16}{(e\ln(\gamma_z))^4 t^4} \leq \frac{16\times 2^4}{(e\ln(\gamma_z))^4(t+1)^4}$. Combing this relationship with (9) in Lemma 8, we obtain

$$\tfrac{\gamma_z^{2t}}{\lambda_t}\mathbb{E}\left[\|\boldsymbol{s}_t - \omega\bar{s}_t\|_C^2\right] < \tfrac{16\times 2^4(c_{\boldsymbol{s}1}t^{-2} + c_{\boldsymbol{s}2}t^{-2v} + c_{\boldsymbol{s}3}t^{-2\varsigma_\zeta})}{\lambda_0(e\ln(\gamma_z))^4(t+1)^{4-v}}.$$

3) Lemma 7 implies $\gamma_z^{2t}\lambda_t \leq \frac{16\times 2^4\lambda_0}{(e\ln(\gamma_z))^4(t+1)^{4+v}}$.

4) Using the relation $\mathbb{E}[\|\theta_{t+1}^* - \theta_t^*\|_2^2] \leq \frac{16(\kappa^2+D^2)}{(t+1)^2}(2\mu^{-2} + L^{-2})$ from Lemma 1 in [48] yields

$$\tfrac{\mathbb{E}\left[\|\theta_{t+1}^* - \theta_t^*\|_2^2\right]}{\lambda_t} \leq \tfrac{16(\kappa^2+D^2)}{\lambda_0(t+1)^{2-v}}\left(\tfrac{2}{\mu^2} + \tfrac{1}{L^2}\right). \quad (64)$$

Substituting (63)-(64) into (56), we obtain

$$\begin{aligned}\Phi_{t,\bar{\theta}} < &\tfrac{\tau_{\bar{\theta}1}}{(t+1)^{v+2}} + \tfrac{\tau_{\bar{\theta}2}}{(t+1)^{v+2}} + \tfrac{\tau_{\bar{\theta}3}}{(t+1)^{3v}} + \tfrac{\tau_{\bar{\theta}4}}{(t+1)^{v+2\varsigma_\vartheta}}\\ &+ \tfrac{\tau_{\bar{\theta}5}}{(t+1)^{v+2\varsigma_\zeta}} + \tfrac{\tau_{\bar{\theta}6}}{(t+1)^{6-v}} + \tfrac{\tau_{\bar{\theta}7}}{(t+1)^{4+v}} + \tfrac{\tau_{\bar{\theta}8}}{(t+1)^{4-v+2\varsigma_\zeta}}\\ &+ \tfrac{\tau_{\bar{\theta}9}}{(t+1)^{4-v+2\varsigma_\zeta}} + \tfrac{\tau_{\bar{\theta}10}}{(t+1)^{4+v}} + \tfrac{\tau_{\bar{\theta}11}}{(t+1)^{2\varsigma_\vartheta}} + \tfrac{\tau_{\bar{\theta}12}}{(t+1)^{2\varsigma_\zeta}}\\ &+ \tfrac{\tau_{\bar{\theta}13}}{(t+1)^{v+\frac{1}{2}}} + \tfrac{\tau_{\bar{\theta}14}}{(t+1)^{2v}} + \tfrac{\tau_{\bar{\theta}15}}{(t+1)^2} + \tfrac{\tau_{\bar{\theta}16}}{(t+1)^{2-v}},\end{aligned} \quad (65)$$

where $\tau_{\bar{\theta}1} = 2\hat{c}_{\bar{\theta}0}\kappa^2\lambda_0$, $\tau_{\bar{\theta}2} = 4\kappa^{-2}c_{\boldsymbol{\theta}1}\delta_{F,R}^2 L^2\tau_{\bar{\theta}1}$, $\tau_{\bar{\theta}3} = 4^{v-1}c_{\boldsymbol{\theta}1}^{-1}c_{\boldsymbol{\theta}2}\tau_{\bar{\theta}2}$, $\tau_{\bar{\theta}4} = 4^{\varsigma_\vartheta-1}c_{\boldsymbol{\theta}1}^{-1}c_{\boldsymbol{\theta}3}\tau_{\bar{\theta}2}$, $\tau_{\bar{\theta}5} = 4^{\varsigma_\zeta-1}c_{\boldsymbol{\theta}1}^{-1}c_{\boldsymbol{\theta}4}\tau_{\bar{\theta}2}$, $\tau_{\bar{\theta}6} = \frac{3\times 2^{10}c_{\boldsymbol{s}1}\hat{c}_{\bar{\theta}0}\delta_{F,C}^2\|C\|_C^2\|u\|_2^2 c_z^2}{m^2\lambda_0(e\ln(\gamma_z))^4}$, $\tau_{\bar{\theta}7} = \frac{4^v c_{\boldsymbol{s}2}\tau_{\bar{\theta}6}}{4c_{\boldsymbol{s}1}}$, $\tau_{\bar{\theta}8} = \frac{4^{\varsigma_\zeta}c_{\boldsymbol{s}3}\tau_{\bar{\theta}6}}{4c_{\boldsymbol{s}1}}$, $\tau_{\bar{\theta}9} = \frac{3\times 2^8\hat{c}_{\bar{\theta}0}\|u\|_2^2 c_z^2\sum_{i,j}(C_{ij})^2(\sigma_\zeta^+)^2}{m^2\lambda_0(e\ln(\gamma_z))^4}$, $\tau_{\bar{\theta}10} = \frac{3\times 2^9\hat{c}_{\bar{\theta}0}(\kappa^2+D^2)\|u\|_2^2 c_z^2\lambda_0}{m(e\ln(\gamma_z))^4}$, $\tau_{\bar{\theta}11} = \frac{(4+\lambda_0\mu)\sum_{i,j}(R_{ij})^2(\sigma_\vartheta^+)^2}{2}$, $\tau_{\bar{\theta}12} = \frac{(4+\lambda_0\mu)\sum_{i,j}(C_{ij})^2(\sigma_\zeta^+)^2}{2}$, $\tau_{\bar{\theta}13} = \hat{c}_{\bar{\theta}0}D\kappa\lambda_0$, $\tau_{\bar{\theta}14} = \frac{\hat{c}_{\bar{\theta}0}\mu\lambda_0^2 D^2}{4}$, $\tau_{\bar{\theta}15} = \frac{16(\kappa^2+D^2)(2L^2+\mu^2)}{\mu^2 L^2}$, and $\tau_{\bar{\theta}16} = \frac{4\tau_{\bar{\theta}15}}{\lambda_0\mu}$.

By plugging (65) into $\sum_{p=1}^{\infty}\Phi_{p,\bar{\theta}}$, we can estimate the second term on the right hand side of (62). To illustrate this idea, we use $\sum_{p=1}^{\infty}\tau_{\bar{\theta}1}(p+1)^{-1-v}$ as an example:

$$\textstyle\sum_{p=1}^{\infty}\frac{\tau_{\bar{\theta}1}}{(t+1)^{1+v}} \leq \int_1^{\infty}\frac{\tau_{\bar{\theta}1}}{x^{1+v}}dx \leq \frac{\tau_{\bar{\theta}1}}{(1+v-1)2^{1-(1+v)}}. \quad (66)$$

Applying an argument similar to the derivation of (66) to the other items on the right hand of $\sum_{p=1}^{\infty}\Phi_{p,\bar{\theta}}$ yields $\sum_{p=1}^{\infty}\Phi_{p,\bar{\theta}} < \frac{\tau_{\bar{\theta}1}2^v}{v} + \frac{\tau_{\bar{\theta}2}2^{v+1}}{v+1} + \frac{\tau_{\bar{\theta}3}2^{3v-1}}{3v-1} + \frac{\tau_{\bar{\theta}4}2^{2\varsigma_\vartheta+v-1}}{v+2\varsigma_\vartheta-1} + \frac{\tau_{\bar{\theta}5}2^{2\varsigma_\zeta+v-1}}{v+2\varsigma_\zeta-1} + \frac{\tau_{\bar{\theta}6}2^{5-v}}{5-v} + \frac{\tau_{\bar{\theta}7}2^{v+3}}{v+3} + \frac{\tau_{\bar{\theta}8}2^{2\varsigma_\zeta-v+3}}{2\varsigma_\zeta-v+3} + \frac{\tau_{\bar{\theta}9}2^{2\varsigma_\zeta-v+3}}{2\varsigma_\zeta-v+3} + \frac{\tau_{\bar{\theta}10}2^{v+3}}{v+3} + \frac{\tau_{\bar{\theta}11}2^{2\varsigma_\vartheta-1}}{2\varsigma_\vartheta-1} + \frac{\tau_{\bar{\theta}12}2^{2\varsigma_\zeta-1}}{2\varsigma_\zeta-1} + \frac{\tau_{\bar{\theta}13}2^{v-\frac{1}{2}}}{v-\frac{1}{2}} + \frac{\tau_{\bar{\theta}14}2^{2v-1}}{2v-1} + 2\tau_{\bar{\theta}15} + \frac{\tau_{\bar{\theta}16}2^{1-v}}{1-v} \triangleq c'$. Combining $\Phi_{0,\bar{\theta}} = \sum_{i=1}^{16}\tau_{\bar{\theta}i}$ with (62) yields that the first term on the right hand side of (60) satisfies

$$\textstyle\sum_{p=0}^{\lfloor t-t^\alpha\rfloor}\Phi_{p,\bar{\theta}} e^{-\frac{\mu}{4}\sum_{q=\lceil t-t^\alpha\rceil}^{t}\lambda_q} < c_{\bar{\theta}2} t^{-1}, \quad (67)$$

with $c_{\bar{\theta}2} = (\frac{\alpha-v+1}{\alpha-v})!(\frac{\mu\lambda_0}{4\times 2^v})^{\frac{1}{v-\alpha}}(\sum_{i=1}^{16}\tau_{\bar{\theta}i} + c')$.

By plugging (65) into $\sum_{p=\lceil t-t^\alpha\rceil}^{t}\Phi_{p,\bar{\theta}}$, we can estimate the second term on the right hand side of (60). To illustrate this idea, we use $\sum_{p=\lceil t-t^\alpha\rceil}^{t}\tau_{\bar{\theta}2}(p+1)^{-v-2}$ as an example: Since the relation $\frac{1}{(p+1)^{v+2}} \leq \frac{1}{(\lceil t-t^\alpha\rceil+1)^{v+2}}$ holds for all $p \in [\lceil t - t^\alpha\rceil, t]$ and any $\alpha \in (v, \frac{1+v}{2})$, we have $\sum_{p=\lceil t-t^\alpha\rceil}^{t}\frac{1}{(p+1)^{v+2}} \leq \frac{1}{(\lceil t-t^\alpha\rceil+1)^{v+2}}(t - \lceil t-t^\alpha\rceil + 1)$. Since $\lceil t-t^\alpha\rceil + 1 \geq t(1-\alpha)$ is valid for all $\alpha \in (0,1)$, we obtain

$$\textstyle\sum_{p=\lceil t-t^\alpha\rceil}^{t}\frac{\tau_{\bar{\theta}2}}{(p+1)^{v+2}} \leq \frac{\tau_{\bar{\theta}2}(t^\alpha+1)}{t^{v+2}(1-\alpha)^{v+2}} \leq \frac{2\tau_{\bar{\theta}2}t^{\alpha-(v+2)}}{(1-\alpha)^{v+2}}. \quad (68)$$

Using an argument similar to the derivation of (68) to the other items on the right hand side of $\sum_{p=\lceil t-t^\alpha\rceil}^{t}\Phi_{p,\bar{\theta}}$ yields

$$\begin{aligned}&\textstyle\sum_{p=\lceil t-t^\alpha\rceil}^{t}\Phi_{p,\bar{\theta}} < c_{\bar{\theta}3}t^{\alpha-(v+1)} + c_{\bar{\theta}4}t^{\alpha-(v+2)} + c_{\bar{\theta}5}t^{\alpha-3v}\\ &+ c_{\bar{\theta}6}t^{\alpha-(v+2\varsigma_\vartheta)} + c_{\bar{\theta}7}t^{\alpha-(v+2\varsigma_\zeta)} + c_{\bar{\theta}8}t^{\alpha-(6-v)}\\ &+ c_{\bar{\theta}9}t^{\alpha-(v+4)} + c_{\bar{\theta}10}t^{\alpha-(4-v+2\varsigma_\zeta)} + c_{\bar{\theta}11}t^{\alpha-2\varsigma_\vartheta} + c_{\bar{\theta}12}t^{\alpha-2\varsigma_\zeta}\\ &+ c_{\bar{\theta}13}t^{\alpha-(v+\frac{1}{2})} + c_{\bar{\theta}14}t^{\alpha-2v} + c_{\bar{\theta}15}t^{\alpha-2} + c_{\bar{\theta}16}t^{\alpha-(2-v)},\end{aligned} \quad (69)$$

where $c_{\bar{\theta}3} = \frac{2\tau_{\bar{\theta}1}}{(1-\alpha)^{v+1}}$, $c_{\bar{\theta}4} = \frac{2\tau_{\bar{\theta}2}}{(1-\alpha)^{v+2}}$, $c_{\bar{\theta}5} = \frac{2\tau_{\bar{\theta}3}}{(1-\alpha)^{3v}}$, $c_{\bar{\theta}6} = \frac{2\tau_{\bar{\theta}4}}{(1-\alpha)^{v+2\varsigma_\vartheta}}$, $c_{\bar{\theta}7} = \frac{2\tau_{\bar{\theta}5}}{(1-\alpha)^{v+2\varsigma_\zeta}}$, $c_{\bar{\theta}8} = \frac{2\tau_{\bar{\theta}6}}{(1-\alpha)^{6-v}}$, $c_{\bar{\theta}9} = \frac{2(\tau_{\bar{\theta}7}+\tau_{\bar{\theta}10})}{(1-\alpha)^{v+4}}$, $c_{\bar{\theta}10} = \frac{2(\tau_{\bar{\theta}8}+\tau_{\bar{\theta}9})}{(1-\alpha)^{4-v+2\varsigma_\zeta}}$, $c_{\bar{\theta}11} = \frac{2\tau_{\bar{\theta}11}}{(1-\alpha)^{2\varsigma_\vartheta}}$, $c_{\bar{\theta}12} = \frac{2\tau_{\bar{\theta}12}}{(1-\alpha)^{2\varsigma_\zeta}}$, $c_{\bar{\theta}13} = \frac{2\tau_{\bar{\theta}13}}{(1-\alpha)^{v+\frac{1}{2}}}$, $c_{\bar{\theta}14} = \frac{2\tau_{\bar{\theta}14}}{(1-\alpha)^{2v}}$, $c_{\bar{\theta}15} = \frac{2\tau_{\bar{\theta}15}}{(1-\alpha)^2}$, and $c_{\bar{\theta}16} = \frac{2\tau_{\bar{\theta}16}}{(1-\alpha)^{2-v}}$.

Substituting (67) and (69) into (60) and then plugging (59) and (60) into (57), we arrive at

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_{t+1}-\theta^*_{t+1}\|_2^2] &< (c_{\bar{\theta}1}+c_{\bar{\theta}2})t^{-1}+c_{\bar{\theta}3}t^{\alpha-(v+1)}\\ &+c_{\bar{\theta}4}t^{\alpha-(v+2)}+c_{\bar{\theta}5}t^{\alpha-3v}+c_{\bar{\theta}6}t^{\alpha-(v+2\varsigma_\vartheta)}+c_{\bar{\theta}7}t^{\alpha-(v+2\varsigma_\zeta)}\\ &+c_{\bar{\theta}8}t^{\alpha-(6-v)}+c_{\bar{\theta}9}t^{\alpha-(v+4)}+c_{\bar{\theta}10}t^{\alpha-(4-v+2\varsigma_\zeta)}\\ &+c_{\bar{\theta}11}t^{\alpha-2\varsigma_\vartheta}+c_{\bar{\theta}12}t^{\alpha-2\varsigma_\zeta}+c_{\bar{\theta}13}t^{\alpha-(v+\frac{1}{2})}\\ &+c_{\bar{\theta}14}t^{\alpha-2v}+c_{\bar{\theta}15}t^{\alpha-2}+c_{\bar{\theta}16}t^{\alpha-(2-v)},\end{aligned}\tag{70}$$

for all $t>0$, where the constants $c_{\bar{\theta}1}$ is given in (59), $c_{\bar{\theta}2}$ is given in (67), $c_{\bar{\theta}3}$ to $c_{\bar{\theta}16}$ are given in (69).

By plugging $\Phi_{0,\bar{\theta}}=\sum_{i=1}^{16}\tau_{\bar{\theta}i}$ into (55), we obtain

$$\mathbb{E}[\|\bar{\theta}_1-\theta^*_1\|_2^2]\le c_{\bar{\theta}17},\tag{71}$$

with $c_{\bar{\theta}17}=(1-\frac{\lambda_0\mu}{4})\mathbb{E}\big[\|\bar{\theta}_0-\theta^*_0\|_2^2\big]+\sum_{i=1}^{16}\tau_{\bar{\theta}i}$.

By using Lemma 5 and Lemma 6, we obtain

$$\|\boldsymbol{\theta}_t-\mathbf{1}\theta^*_t\|_F^2\le 2\delta_{F,R}^2\|\boldsymbol{\theta}_t-\mathbf{1}\bar{\theta}_t\|_R^2+2m\|\bar{\theta}_t-\theta^*_t\|_2^2.\tag{72}$$

Taking the expectation on both sides of (72) and combining the result with (10), (70), and (71), we arrive at

$$\begin{aligned}\mathbb{E}[\|\theta^i_t-\theta^*_t\|_2^2] &< 2\delta_{F,R}^2\left(c_{\boldsymbol{\theta}1}t^{-2}+c_{\boldsymbol{\theta}2}t^{-2v}+c_{\boldsymbol{\theta}3}t^{-2\varsigma_\vartheta}+c_{\boldsymbol{\theta}4}t^{-2\varsigma_\zeta}\right)\\ &+2m\big(\max\{c_{\bar{\theta}1},c_{\bar{\theta}2},c_{\bar{\theta}17}\}t^{-1}+c_{\bar{\theta}3}t^{\alpha-v-1}+c_{\bar{\theta}4}t^{\alpha-v-2}\\ &+c_{\bar{\theta}5}t^{\alpha-3v}+c_{\bar{\theta}6}t^{\alpha-v-2\varsigma_\vartheta}+c_{\bar{\theta}7}t^{\alpha-v-2\varsigma_\zeta}+c_{\bar{\theta}8}t^{\alpha-6+v}\\ &+c_{\bar{\theta}9}t^{\alpha-v-4}+c_{\bar{\theta}10}t^{\alpha-4+v-2\varsigma_\zeta}+c_{\bar{\theta}11}t^{\alpha-2\varsigma_\vartheta}+c_{\bar{\theta}12}t^{\alpha-2\varsigma_\zeta}\\ &+c_{\bar{\theta}13}t^{\alpha-v-\frac{1}{2}}+c_{\bar{\theta}14}t^{\alpha-2v}+c_{\bar{\theta}15}t^{\alpha-2}+c_{\bar{\theta}16}t^{\alpha-2+v}\big),\end{aligned}\tag{73}$$

for all $t>0$. Here, the constant α satisfies $\alpha\in(v,\frac{1+v}{2})$.

Substituting (4) and (73) into the triangle inequality $\|\theta^i_t-\theta^*\|_2^2\le 2\|\theta^i_t-\theta^*_t\|_2^2+2\|\theta^*_t-\theta^*\|_2^2$, we arrive at (11).

E. Proof of Theorem 2

For the convenience of derivation, we introduce an auxiliary variable $s\in[0,t]$. By plugging (8) into (42), we obtain

$$\mathbb{E}[\|\bar{\theta}_{t+1-s}-\theta^*\|_2^2]\le\mathbb{E}[\|\bar{\theta}_{t-s}-\theta^*\|_2^2]+\textstyle\sum_{i=1}^{6}\Gamma_i,\tag{74}$$

where Γ_1 to Γ_6 are given by

$$\begin{cases}\Gamma_1=\mathbb{E}\big[\big\|\lambda_{t-s}\nabla\bar{f}_{t-s}(\mathbf{1}\bar{\theta}_{t-s})-\lambda_{t-s}\nabla\bar{f}_{t-s}(\boldsymbol{\theta}_{t-s})\\ \qquad-\frac{u^T(Z_{t-s}^{-1}-U^{-1})}{m}(\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s})\big\|_2^2\big],\\ \Gamma_2=\mathbb{E}[\|\bar{\vartheta}_{R,t-s}-\bar{\zeta}_{C,t-s}\|_2^2],\quad \Gamma_3=\mathbb{E}[\|\lambda_{t-s}\nabla\bar{f}_{t-s}(\mathbf{1}\bar{\theta}_{t-s})\|_2^2],\\ \Gamma_4=2\lambda_{t-s}\mathbb{E}\big[\big\langle\nabla\bar{f}_{t-s}(\mathbf{1}\bar{\theta}_{t-s})-\nabla\bar{f}_{t-s}(\boldsymbol{\theta}_{t-s})+\bar{\vartheta}_{R,t-s}-\bar{\zeta}_{C,t-s}\\ \qquad-\frac{u^T(Z_{t-s}^{-1}-U^{-1})}{m}(\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s}),\lambda_{t-s}\nabla\bar{f}_{t-s}(\mathbf{1}\bar{\theta}_{t-s})\big\rangle\big],\\ \Gamma_5=2\mathbb{E}\big[\big\langle\bar{\theta}_{t-s}-\theta^*,\lambda_{t-s}\nabla\bar{f}_{t-s}(\mathbf{1}\bar{\theta}_{t-s})-\lambda_{t-s}\nabla\bar{f}_{t-s}(\boldsymbol{\theta}_{t-s})\\ \qquad+\bar{\vartheta}_{R,t-s}-\bar{\zeta}_{C,t-s}-\frac{u^T(Z_{t-s}^{-1}-U^{-1})}{m}(\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s})\big\rangle\big],\\ \Gamma_6=2\mathbb{E}[\langle\bar{\theta}_{t-s}-\theta^*,\lambda_{t-s}\nabla\bar{f}_{t-s}(\mathbf{1}\bar{\theta}_{t-s})\rangle_2].\end{cases}\tag{75}$$

We further characterize each item in (75):

1) By using Assumption 3(iii) and Lemma 3, we have

$$\begin{aligned}\Gamma_1\le 6L^2\delta_{F,R}^2\lambda_{t-s}^2\mathbb{E}[\|\boldsymbol{\theta}_{t-s}-\mathbf{1}\bar{\theta}_{t-s}\|_R^2]+\frac{12\kappa^2\lambda_{t-s}^2}{t-s+1}\\ +\frac{2\|u\|_2^2c_z^2}{m^2}\gamma_z^{2(t-s)}\mathbb{E}[\|\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s}\|_F^2].\end{aligned}\tag{76}$$

2) Based on the definitions of $\bar{\vartheta}_{R,t-s}$ and $\bar{\vartheta}_{C,t-s}$, we have

$$\Gamma_2\le\frac{2\|u\|_2^2(\sigma_\vartheta^+)^2\sum_{i,j}(R_{ij})^2}{m^2(t-s+1)^{2\varsigma_\vartheta}}+\frac{2(\sigma_\zeta^+)^2\sum_{i,j}(C_{ij})^2}{m^2(t-s+1)^{2\varsigma_\zeta}}.$$

3) Using an argument similar to the derivation of (34) yields

$$\Gamma_3\le 2(\kappa^2+D^2)\lambda_{t-s}^2.\tag{77}$$

4) By utilizing Assumption 4, (76), (77), and the relation $2\langle a,b\rangle\le\|a\|^2+\|b\|^2$ for any vectors a and b, we have

$$\begin{aligned}\Gamma_4\le 2(\kappa^2+D^2)\lambda_{t-s}^2+6L^2\delta_{F,R}^2\lambda_{t-s}^2\mathbb{E}[\|\boldsymbol{\theta}_{t-s}-\mathbf{1}\bar{\theta}_{t-s}\|_R^2]\\ +\frac{2\|u\|_2^2c_z^2}{m^2}\gamma_z^{2(t-s)}\mathbb{E}[\|\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s}\|_F^2]+\frac{12\kappa^2\lambda_{t-s}^2}{t-s+1}.\end{aligned}$$

5) By using Assumption 4 and (76) and defining $a_{t-s}\triangleq\frac{1}{(t-s+1)^r}$ with $r\in(\frac{1}{2},v)$, we obtain

$$\begin{aligned}\Gamma_5\le a_{t-s}\lambda_{t-s}\mathbb{E}[\|\bar{\theta}_{t-s}-\theta^*\|_2^2]\\ +\frac{6L^2\delta_{F,R}^2\lambda_{t-s}^2}{a_{t-s}}\mathbb{E}[\|\boldsymbol{\theta}_{t-s}-\mathbf{1}\bar{\theta}_{t-s}\|_R^2]+\frac{12\kappa^2\lambda_{t-s}^2}{(t-s+1)a_{t-s}}\\ +\frac{2\|u\|_2^2c_z^2}{m^2}\frac{\gamma_z^{2(t-s)}}{a_{t-s}\lambda_{t-s}}\mathbb{E}[\|\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s}\|_F^2].\end{aligned}$$

6) By defining $a_{t-s}\triangleq\frac{1}{(t-s+1)^r}$ with $r\in(\frac{1}{2},v)$, we have

$$\begin{aligned}\Gamma_6\le-2\lambda_{t-s}\mathbb{E}[F(\theta^i_{t+1})-F(\theta^*)]\\ +\lambda_{t-s}a_{t-s}\mathbb{E}[\|\bar{\theta}_{t-s}-\theta^*\|_2^2]+\frac{\lambda_{t-s}\kappa^2}{a_{t-s}(t-s+1)}.\end{aligned}\tag{78}$$

Substituting (76)-(78) into (74), we arrive at

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_{t+1-s}-\theta^*\|_2^2]\le-2\lambda_{t-s}\mathbb{E}[F(\theta^i_{t+1})-F(\theta^*)]\\ +(1+2\lambda_{t-s}a_{t-s})\mathbb{E}[\|\bar{\theta}_{t-s}-\theta^*\|_2^2]+\Phi_{t-s},\end{aligned}\tag{79}$$

where the term Φ_{t-s} is given by

$$\begin{aligned}\Phi_{t-s}=6L^2\delta_{F,R}^2\Big(2\lambda_{t-s}^2+\frac{\lambda_{t-s}}{a_{t-s}}\Big)\mathbb{E}[\|\boldsymbol{\theta}_{t-s}-\mathbf{1}\bar{\theta}_{t-s}\|_R^2]\\ +\frac{2\|u\|_2^2c_z^2}{m^2}\Big(2\gamma_z^{2(t-s)}+\frac{\gamma_z^{2(t-s)}}{a_{t-s}\lambda_{t-s}}\Big)\mathbb{E}[\|\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s}\|_F^2]\\ +\frac{2\|u\|_2^2(\sigma_\vartheta^+)^2\sum_{i,j}(R_{ij})^2}{m^2(t-s+1)^{2\varsigma_\vartheta}}+\frac{2(\sigma_\zeta^+)^2\sum_{i,j}(C_{ij})^2}{m^2(t-s+1)^{2\varsigma_\zeta}}\\ +4(\kappa^2+D^2)\lambda_{t-s}^2+\frac{\kappa^2(24\lambda_{t-s}^2+\frac{13\lambda_{t-s}}{a_{t-s}})}{(t-s+1)}.\end{aligned}\tag{80}$$

We define $\bar{t}=t-s+1$ for all $\bar{t}>0$ and drop the negative term $-2\lambda_{t-s}\mathbb{E}[F(\theta^i_{t+1})-F(\theta^*)]$ to rewrite (79) as follows:

$$\mathbb{E}\left[\|\bar{\theta}_{\bar{t}}-\theta^*\|_2^2\right]\le(1+2\lambda_{\bar{t}-1}a_{\bar{t}-1})\mathbb{E}\left[\|\bar{\theta}_{\bar{t}-1}-\theta^*\|_2^2\right]+\Phi_{\bar{t}-1}.\tag{81}$$

By iterating (81) from 0 to $\bar{t}-1$, one yields

$$\mathbb{E}[\|\bar{\theta}_{\bar{t}}-\theta^*\|_2^2]\le\big(\textstyle\prod_{p=0}^{\bar{t}-1}(1+2\lambda_pa_p)\big)\big(\mathbb{E}[\|\bar{\theta}_0-\theta^*\|_2^2]+\sum_{p=0}^{\bar{t}-1}\Phi_p\big).\tag{82}$$

Since $\ln(\prod_{p=0}^{\bar{t}-1}(1+2\lambda_pa_p))\le\frac{2\lambda_0(r+v)}{r+v-1}$ is valid, we have $\prod_{p=0}^{\bar{t}-1}(1+2\lambda_pa_p)\le e^{\frac{2\lambda_0(r+v)}{r+v-1}}$. We use this inequality and replace $\bar{t}$ with $t-s+1$ to rewrite (82) as follows

$$\mathbb{E}[\|\bar{\theta}_{t-s+1}-\theta^*\|_2^2]\le e^{\frac{2\lambda_0(r+v)}{r+v-1}}\Big(\mathbb{E}[\|\bar{\theta}_0-\theta^*\|_2^2]+\textstyle\sum_{s=0}^{t}\Phi_{t-s}\Big),\tag{83}$$

where in the derivation we used $\sum_{p=0}^{t-s}\Phi_p=\sum_{s=0}^{t}\Phi_{t-s}$.

We proceed to estimate an upper bound on $\sum_{s=0}^{t}\Phi_{t-s}$:

1) Considering that $a_{t-s}\lambda_{t-s}\le\lambda_0$ implies $\lambda_{t-s}^2\le\frac{\lambda_{t-s}}{a_{t-s}}\lambda_0$, we combine (10) and $(t-s+1)^p\le 2^p(t-s)^p$ to obtain

$$6L^2\delta_{F,R}^2\textstyle\sum_{s=0}^{t}(2\lambda_{t-s}^2+\frac{\lambda_{t-s}}{a_{t-s}})\mathbb{E}[\|\boldsymbol{\theta}_{t-s}-\mathbf{1}\bar{\theta}_{t-s}\|_R^2]<\bar{c}_{\boldsymbol{\theta}1},\tag{84}$$

with $\bar{c}_{\boldsymbol{\theta}1}=6L^2\delta_{F,R}^2\lambda_0(2\lambda_0+1)(\frac{4c_{\boldsymbol{\theta}1}(v-r+2)}{v-r+1}+\frac{4^vc_{\boldsymbol{\theta}2}(3v-r)}{3v-r-1}+\frac{4^{\varsigma_\vartheta}c_{\boldsymbol{\theta}3}(v-r+2\varsigma_\vartheta)}{v-r+2\varsigma_\vartheta-1}+\frac{4^{\varsigma_\zeta}c_{\boldsymbol{\theta}4}(v-r+2\varsigma_\zeta)}{v-r+2\varsigma_\zeta-1})$.

2) The relation $a_{t-s}\lambda_{t-s}\leq\lambda_0$ implies $\gamma_z^{2(t-s)}\leq\frac{\gamma_z^{2(t-s)}}{a_{t-s}\lambda_{t-s}}\lambda_0$. Using this relation, (54), Lemma 7, and Lemma 8 yields

$\frac{2\|u\|_2^2c_z^2}{m^2}\sum_{s=0}^t(2\gamma_z^{2(t-s)}+\frac{\gamma_z^{2(t-s)}}{a_{t-s}\lambda_{t-s}})\mathbb{E}[\|\boldsymbol{s}_{t+1-s}-\boldsymbol{s}_{t-s}\|_F^2]<\bar{c}_{\boldsymbol{\theta}2}$,

with $\bar{c}_{\boldsymbol{\theta}2}=\frac{2^9\|u\|_2^2c_z^2(2\lambda_0+1)}{(\ln(\gamma_z)e)^4m^2\lambda_0}\big(\frac{12\delta_{F,C}^2\|C\|_C^2c_{\boldsymbol{s}1}(6-v-r)}{5-v-r}+\frac{(3\times4^v\delta_{F,C}^2\|C\|_C^2c_{\boldsymbol{s}2})(4+v-r)}{3+v-r}+\frac{(6m(\kappa^2+D^2)\lambda_0^2)(4+v-r)}{3+v-r}+\frac{(3\times4^{\varsigma_\zeta}\delta_{F,C}^2\|C\|_C^2c_{\boldsymbol{s}3}+3\sum_{i,j}(C_{ij})^2(\sigma_\zeta^+)^2)(4+2\varsigma_\zeta-v-r)}{3+2\varsigma_\zeta-v-r}\big)$.

3) Applying the relation $\sum_{s=0}^t\frac{1}{(t-s+1)^p}\leq\frac{p}{p-1}$ to the rest of terms on the right hand side of $\sum_{s=0}^t\Phi_{t-s}$ yields

$$\sum_{s=0}^t\frac{2\|u\|_2^2(\sigma_\vartheta^+)^2\sum_{i,j}(R_{ij})^2}{m^2(t-s+1)^{2\varsigma_\vartheta}}+\sum_{s=0}^t\frac{2(\sigma_\zeta^+)^2\sum_{i,j}(C_{ij})^2}{m^2(t-s+1)^{2\varsigma_\zeta}}+\sum_{s=0}^t4(\kappa^2+D^2)\lambda_{t-s}^2+\frac{\kappa^2(24\lambda_{t-s}^2+\sum_{s=0}^t\frac{13\lambda_{t-s}}{a_{t-s}})}{(t-s+1)}\leq\bar{c}_{\boldsymbol{\theta}3},\tag{85}$$

with $\bar{c}_{\boldsymbol{\theta}3}=\frac{4\|u\|_2^2(\sigma_\vartheta^+)^2\sum_{i,j}(R_{ij})^2\varsigma_\vartheta}{m^2(2\varsigma_\vartheta-1)}+\frac{4(\sigma_\zeta^+)^2\sum_{i,j}(C_{ij})^2\varsigma_\zeta}{m^2(2\varsigma_\zeta-1)}+\frac{8v\lambda_0^2(\kappa^2+D^2)}{2v-1}+\frac{12\kappa^2(2v+1)\lambda_0^2}{v}+\frac{13\kappa^2\lambda_0(v+1-r)}{v-r}$.

Summing up both sides of (84)-(85) yields $\sum_{s=0}^t\Phi_{t-s}=\sum_{i=1}^3\bar{c}_{\boldsymbol{\theta}i}$. Further substituting this relationship into (83), we have $\mathbb{E}[\|\bar{\theta}_{t-s+1}-\theta^*\|_2^2]\leq\bar{c}'$ for all $s\in[0,t]$, where $\bar{c}'\triangleq e^{\frac{2\lambda_0(r+v)}{r+v-1}}(\mathbb{E}\left[\|\bar{\theta}_0-\theta^*\|_2^2\right]+\sum_{i=1}^3\bar{c}_{\boldsymbol{\theta}i})$.

We sum up both sides of (79) from 0 to t to obtain

$$2\sum_{s=0}^t\lambda_{t-s}\mathbb{E}\big[F(\theta_{t+1}^i)-F(\theta^*)\big]\leq(1+2\lambda_0)\mathbb{E}\big[\|\bar{\theta}_0-\theta^*\|_2^2\big]+\sum_{s=0}^{t-1}2\lambda_{t-s}a_{t-s}\mathbb{E}\big[\|\bar{\theta}_{t-s}-\theta^*\|_2^2\big]\leq\bar{c}_{\boldsymbol{\theta}4},\tag{86}$$

where $\bar{c}_{\boldsymbol{\theta}4}$ is given by $\bar{c}_{\boldsymbol{\theta}4}=\frac{2(v+r)e^{\frac{2\lambda_0(r+v)}{r+v-1}}}{v+r-1}(\mathbb{E}\left[\|\bar{\theta}_0-\theta^*\|_2^2\right]+\sum_{i=1}^3\bar{c}_{\boldsymbol{\theta}i})+(1+2\lambda_0)\mathbb{E}[\|\bar{\theta}_0-\theta^*\|_2^2]$. Using the relation $2\lambda_0((t+2)^{1-v}-1)\geq2\lambda_0(1-\frac{1}{2^{1-v}})(t+1)^{1-v}$, we arrive at (12).

F. Proof of Lemma 11

We first prove the following inequality for all $p\in[0,t]$:

$$\frac{\beta_t}{\beta_{t-p}}\geq c_0\big(1-\tfrac{c}{2}\big)^p,\tag{87}$$

where the constant c_0 is given by $c_0=\big(\frac{e\ln(\frac{2}{2-c})}{2q}\big)^q$.

When constants $a,b,c,d>0$ satisfy $\frac{c}{a}\geq\frac{d}{b}$, the relationship $\frac{d}{b}\leq\frac{c+d}{a+b}\leq\frac{c}{a}$ always holds. This inequality further implies $\frac{t+2-p}{t+1}=\frac{t+1-p+1}{t+1-p+p}\geq\frac{1}{p}$ and $\frac{t-p+1}{t-p+2}\geq\frac{1}{2}$. Therefore, we have

$$\frac{\beta_t}{\beta_{t-p}}=\big(\tfrac{t-p+1}{t+1}\big)^q\geq\tfrac{1}{(2p)^q}.\tag{88}$$

Consider a convex function $f(x)=-q\ln(x)-x\ln(1-\frac{c}{2}):\mathbb{R}^+\to\mathbb{R}$ (where $\mathbb{R}^+$ represents the set of positive real numbers), whose derivative satisfies $f'(x)=-\frac{q}{x}-\ln(1-\frac{c}{2})$, implying the minimum point at $x^*=-\frac{q}{\ln(1-\frac{c}{2})}$ with the minimal value $f(x^*)=-q\ln\big(-\frac{q}{\ln(1-\frac{c}{2})}\big)+\frac{q}{\ln(1-\frac{c}{2})}\ln\big(1-\frac{c}{2}\big)=\ln(c_02^q)$. Hence, for any $p\in\mathbb{N}^+$, we have $f(p)\geq\ln(c_02^q)$, which is equivalent to $\ln\big((1-\frac{c}{2})^p\big)\leq\ln(\frac{1}{c_02^qp^q})$. Combining this relation with (88) yields (87).

By iterating $\psi_{t+1}\leq(1-c)\psi_t+\beta_t$ from 0 to t, we obtain

$$\psi_{t+1}\leq(1-c)(1-c)^t\psi_0+\sum_{p=0}^t\beta_{t-p}(1-c)^p.\tag{89}$$

By using $(1-c)^p\leq(1-\frac{c}{2})^{2p}$, we have $\sum_{p=0}^t\beta_{t-p}(1-c)^p\leq\beta_t\sum_{p=0}^t\frac{\beta_{t-p}}{\beta_t}(1-\frac{c}{2})^{2p}$. Based on (87), one yields

$$\sum_{p=0}^t\beta_{t-p}(1-c)^p\leq\tfrac{\beta_t}{c_0}\sum_{p=0}^t\big(1-\tfrac{c}{2}\big)^p\leq\tfrac{2\beta_t}{cc_0}.\tag{90}$$

Using again inequality (87), we obtain

$$(1-c)^t\psi_0\leq\big(1-\tfrac{c}{2}\big)^t\psi_0\leq\tfrac{\psi_0}{c_0\beta_0}\beta_t.\tag{91}$$

Substituting (90) and (91) into (89), we arrive at $\psi_t\leq\frac{1}{c_0}\big(\frac{\psi_0(1-c)}{\beta_0}+\frac{2}{c}\big)\beta_{t-1}$. Further using the relationship $\beta_{t-1}\leq2^q\beta_t$ and the definition of c_0 yields Lemma 11.

REFERENCES

[1] J. B. Predd, S. B. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 56–69, 2006.

[2] C. Yu, X. Wang, X. Xu, M. Zhang, H. Ge, J. Ren, L. Sun, B. Chen, and G. Tan, "Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 735–748, 2019.

[3] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognit. Lett.*, vol. 141, pp. 61–67, 2021.

[4] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5330–5340, 2017.

[5] Y. Bo and Y. Wang, "Quantization avoids saddle points in distributed optimization," *Proc. Natl. Acad. Sci.*, vol. 121, no. 17, p. e2319625121, 2024.

[6] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-bayesian learning," *IEEE Trans. Autom. Contr.*, vol. 62, no. 11, pp. 5538–5553, 2017.

[7] Z. Chen and Y. Wang, "Locally differentially private decentralized stochastic bilevel optimization with guaranteed convergence accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2024.

[8] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative deep learning in fixed topology networks," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5904–5914, 2017.

[9] A. Koloskova, T. Lin, and S. U. Stich, "An improved analysis of gradient tracking for decentralized machine learning," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 11422–11435, 2021.

[10] S. A. Alghunaim and K. Yuan, "A unified and refined convergence analysis for non-convex decentralized learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 3264–3279, 2022.

[11] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *Proc. 54th IEEE Conf. Decis. Contr.*, pp. 2055–2060, IEEE, 2015.

[12] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Math. Program.*, vol. 187, pp. 409–457, 2021.

[13] R. Xin, U. A. Khan, and S. Kar, "A fast randomized incremental gradient method for decentralized nonconvex optimization," *IEEE Trans. Autom. Contr.*, vol. 67, no. 10, pp. 5150–5165, 2021.

[14] Y. Sun, G. Scutari, and A. Daneshmand, "Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation," *SIAM J. Optim.*, vol. 32, no. 2, pp. 354–385, 2022.

[15] Y. Xiong, L. Wu, K. You, and L. Xie, "Quantized distributed gradient tracking algorithm with linear convergence in directed networks," *IEEE Trans. Autom. Contr.*, vol. 68, no. 9, pp. 5638–5645, 2022.

[16] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push–pull gradient methods for distributed optimization in networks," *IEEE Trans. Autom. Contr.*, vol. 66, no. 1, pp. 1–16, 2020.

[17] C. Xi, V. S. Mai, R. Xin, E. H. Abed, and U. A. Khan, "Linear convergence in optimization over directed graphs with row-stochastic matrices," *IEEE Trans. Autom. Contr.*, vol. 63, no. 10, pp. 3558–3565, 2018.

[18] S. Pu, "A robust gradient tracking method for distributed optimization over directed networks," in *Proc. 59th IEEE Conf. Decis. Contr.*, pp. 2335–2341, IEEE, 2020.

[19] R. Xin and U. A. Khan, "Distributed heavy-ball: A generalization and acceleration of first-order methods with gradient tracking," *IEEE Trans. Autom. Contr.*, vol. 65, no. 6, pp. 2627–2633, 2019.

[20] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2597–2633, 2017.

[21] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Math. Program.*, vol. 176, pp. 497–544, 2019.

[22] Y. Wang and T. Başar, "Gradient-tracking based distributed optimization with guaranteed optimality under noisy information sharing," *IEEE Trans. Autom. Contr.*, vol. 68, no. 8, pp. 4796–4811, 2022.

[23] H. M. Gomes, M. Grzenda, R. Mello, J. Read, M. H. Le Nguyen, and A. Bifet, "A survey on semi-supervised learning for delayed partially labelled data streams," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–42, 2022.

[24] Y. Zhang, R. J. Ravier, M. M. Zavlanos, and V. Tarokh, "A distributed online convex optimization algorithm with improved dynamic regret," in *Proc. 58th IEEE Conf. Decis. Contr.*, pp. 2449–2454, IEEE, 2019.

[25] R. Xin, U. A. Khan, and S. Kar, "An improved convergence analysis for decentralized online stochastic non-convex optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 1842–1858, 2021.

[26] G. Carnevale, F. Farina, I. Notarnicola, and G. Notarstefano, "GTAdam: Gradient tracking with adaptive momentum for distributed online optimization," *IEEE Trans. Contr. Netw. Syst.*, vol. 10, no. 3, pp. 1436–1448, 2022.

[27] D. A. Burbano-L, J. George, R. A. Freeman, and K. M. Lynch, "Inferring private information in wireless sensor networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 4310–4314, IEEE, 2019.

[28] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 14774–14784.

[29] C. Zhang and Y. Wang, "Enabling privacy-preservation in decentralized optimization," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 2, pp. 679–689, 2018.

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

[31] Y. Mo and R. M. Murray, "Privacy preserving average consensus," *IEEE Trans. Autom. Contr.*, vol. 62, no. 2, pp. 753–765, 2016.

[32] 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, 2018.

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

[34] H. Gao, Y. Wang, and A. Nedić, "Dynamics based privacy preservation in decentralized optimization," *Automatica*, vol. 151, p. 110878, 2023.

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

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

[37] 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. Contr.*, vol. 67, no. 2, pp. 722–737, 2021.

[38] B. Jayaraman, L. Wang, D. Evans, and Q. Gu, "Distributed learning without distress: Privacy-preserving empirical risk minimization," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[39] R. Chourasia, J. Ye, and R. Shokri, "Differential privacy dynamics of langevin diffusion and noisy gradient descent," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 14771–14781, 2021.

[40] 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.

[41] M. Yuan, J. Lei, and Y. Hong, "Differentially private distributed online mirror descent algorithm," *Neurocomputing*, p. 126531, 2023.

[42] 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.

[43] Q. Lü, X. Liao, T. Xiang, H. Li, and T. Huang, "Privacy masking stochastic subgradient-push algorithm for distributed online optimization," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3224–3237, 2020.

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

[45] 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.

[46] 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. Contr. Netw. Syst.*, vol. 10, no. 4, pp. 2165–2177, 2023.

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

[48] Z. Chen and Y. Wang, "Locally differentially private distributed online learning with guaranteed optimality," *IEEE Trans. Autom. Contr. (Early Access)*, 2024.

[49] V. S. Mai and E. H. Abed, "Distributed optimization over weighted directed graphs using row stochastic matrix," in *Poc. Am. Contr. Conf.*, pp. 7165–7170, 2016.

[50] 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.

[51] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[52] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *Int. Conf. Learn. Represent.*, 2018.

[53] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symp. Secur. Priv.*, pp. 691–706, 2019.

[54] M. Bin, I. Notarnicola, and T. Parisini, "Stability, linear convergence, and robustness of the wang-elia algorithm for distributed consensus optimization," in *IEEE 61st Conf. Decis. Contr.*, pp. 1610–1615, 2022.

[55] L. Berrada, A. Zisserman, and P. Mudigonda, "Smooth loss functions for deep top-k classification," in *Int. Conf. Learn. Represent.*, 2018.

[56] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*, vol. 9. Philadelphia, PA, USA: SIAM, 2009.

[57] E. Dall'Anese, A. Simonetto, S. Becker, and L. Madden, "Optimization and learning with information streams: Time-varying algorithms and applications," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 71–83, 2020.

[58] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *Proc. 58th IEEE Conf. Decis. Contr.*, pp. 8353–8358, IEEE, 2019.

[59] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[60] "Classification and regression datasets." https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

[61] Y. LeCun, "The MNIST database of handwritten digits," 1998.

[62] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," 2010.

Ziqin Chen received the Ph.D. degree in automation from the University of Science and Technology of China, Hefei, China, in 2020. She is currently a postdoctoral associate at the Department of Electrical Computer and Engineering, Clemson University, USA. Her was a postdoctoral fellow at Tongji University, China, from 2020 to 2022. Her current research interests include differential privacy, distributed 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.