Differentially-Private Distributed Optimization with Guaranteed Optimality

Yongqiang Wang, Angelia Nedić

***Abstract*—Privacy protection is gaining increased attention in distributed optimization and learning. As differential privacy is becoming a de facto standard for privacy preservation, recently results have emerged integrating differential privacy with distributed optimization. However, to ensure rigorous differential privacy (with a finite cumulative privacy budget), all existing approaches have to sacrifice provable convergence to the optimal solution. In this paper, we propose a differentially-private distributed optimization algorithm that can ensure, for the first time, both rigorous ϵ-differential privacy and optimality, even on the infinite time horizon. Numerical simulation results confirm the effectiveness of the proposed approach.**

I. Introduction

The problem of optimizing a global objective function through the cooperation of multiple agents has gained increased attention in recent years. In the problem, each agent only has access to a local objective function, and can only communicate with its local neighbors. The agents cooperate to minimize the summation of all individual agents' local objective functions. Such a distributed optimization problem can be formulated in the following general form:

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

where m is the number of agents, $\theta \in \mathbb{R}^d$ is a decision variable common to all agents, while $f_i : \mathbb{R}^d \to \mathbb{R}$ is a local objective function private to agent i.

Plenty of approaches have been reported to solve the above distributed optimization problem since the seminal work of [2], with some of the commonly used approaches including gradient methods (e.g., [3], [4], [5], [6], [7], [8]), distributed alternating direction method of multipliers (e.g., [9], [10]), and distributed Newton methods (e.g., [11]). Among these approaches, gradient-based approaches are gaining increased traction due to their efficiency in both computation complexity and storage requirement, which is particularly appealing for agents with limited computational or storage capabilities.

Despite the enormous success of gradient based distributed optimization algorithms, they all explicitly share optimization variables and/or gradient estimates in every iteration, which becomes a problem in applications involving sensitive data. For example, in the rendezvous problem where a group of agents uses distributed optimization to cooperatively find an optimal assembly point, participating agents may want to keep their initial positions private, which is particularly important in unfriendly environments [10]. In sensor network based localization, the positions of sensor agents should be kept private in sensitive (hostile) environments as well [10], [12]. In fact, without an effective privacy mechanism in place, the results in [10], [12], [13] show that a participating agent's sensitive information, such as position, can be easily inferred by an adversary or other participating agents in distributed-optimization based rendezvous and localization. Another example underscoring the importance of privacy protection in distributed optimization is machine learning where exchanged data may contain sensitive information such as medical records or salary information [14]. In fact, recent results in [15] show that without a privacy mechanism in place, an adversary can use shared information to precisely recover the raw data used for training (pixel-wise accurate for images and token-wise matching for texts).

To address the pressing need for privacy protection in distributed optimization, recently plenty of efforts have been reported. One approach resorts to partially homomorphic encryption, which has been employed in both our own prior results [10], [16], and others [17], [18]. However, such approaches incur heavy communication and computation overhead. Another approach employs the structural properties of distributed optimization to inject temporally or spatially correlated uncertainties, which can also provide privacy protection in distributed optimization (see [14], [19], [20] as well as our own results [21]). However, since the uncertainties injected by these approaches are correlated, their enabled privacy is usually restricted. Time-varying random stepsizes [22] and quantization errors [23] can also be exploited to achieve a certain level of privacy protection in distributed optimization. As Differential Privacy (DP) is immune to arbitrary post-processing (including, e.g., statistical inferences), and can provide strong privacy protection for a participating agent even when all its neighbors are compromised [24], it is gradually becoming a de facto standard for privacy protection. In fact, as DP has achieved remarkable successes in various applications [25], [26], [27], [28], [29], some efforts have also been reported incorporating DP-noise into distributed optimization. For example, approaches have been proposed to obscure shared information in distributed optimization by injecting DP-noise to exchanged messages [12], [30], [31], [32], or objective functions [33]. However, while obscuring information, directly incorporating persistent DP-noise into existing algorithms also unavoidably compromises the accuracy of optimization, leading to a fundamental trade-off between

An extended version of the manuscript [1] has been submitted to IEEE Transactions on Automatic Control. The work was supported in part by the National Science Foundation under Grants ECCS-1912702, CCF-2106293, and CCF-2106336.

Yongqiang Wang is with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA yongqiw@clemson.edu

Angelia Nedić is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA angelia.nedich@asu.edu

privacy and accuracy. In fact, recently the investigation in [15] indicates that directly incorporating DP-noise can achieve reasonable privacy protection *"only when the noise variance is large enough to degrade accuracy [15]."*

In this paper, we propose to tailor gradient methods for differentially-private distributed optimization. More specifically, motivated by the observation that inter-agent coupling becomes unnecessary after convergence, we propose to gradually weaken coupling strength in distributed optimization to attenuate DP-noise that is added to every shared message. We judiciously design the weakening factor sequence such that the consensus and convergence to an optimal solution are ensured even in the presence of persistent DP-noise. To our knowledge, this is the first time that both differential privacy and provable optimality are ensured simultaneously in distributed optimization.

Notations: We use $\mathbb{R}^d$ to denote the Euclidean space of dimension d. We write I_d for the identity matrix of dimension d, and $\mathbf{1}_d$ for the d-dimensional column vector will all entries equal to 1; in both cases we suppress the dimension when clear from the context. For a vector x, x_i denotes its ith element. We use $\langle\cdot,\cdot\rangle$ to denote the inner product. We write $\|A\|$ for the matrix norm induced by the vector norm $\|\cdot\|$, unless stated otherwise. We let A^T denote the transpose of a matrix A. A matrix is column-stochastic when its entries are nonnegative and elements in every column add up to one. A square matrix A is said to be doubly-stochastic when both A and A^T are column-stochastic. For two vectors u and v with the same dimension, we use $u \leq v$ to represent the relationship that every element of the vector $u - v$ is nonpositive. Often, we abbreviate *almost surely* by *a.s.*

II. Problem Formulation and Preliminaries

A. On distributed optimization

We describe the local interaction among agents using a weight matrix $W = \{W_{ij}\}$, where $W_{ij} > 0$ if agent j and agent i can directly communicate with each other, and $W_{ij} = 0$ otherwise. For an agent $i \in [m]$, its neighbor set $\mathbb{N}_i$ is defined as the collection of agents j such that $W_{ij} > 0$. We define $W_{ii} \triangleq -\sum_{j\in\mathbb{N}_i} W_{ij}$ for all $i \in [m]$, where $\mathbb{N}_i$ is the neighbor set of agent i. Furthermore, We make the following assumption on W:

Assumption 1. *The matrix $W = \{w_{ij}\} \in \mathbb{R}^{m\times m}$ is symmetric and satisfies $\mathbf{1}^T W = \mathbf{0}^T$, $W\mathbf{1} = \mathbf{0}$, $\|I + W - \frac{\mathbf{1}\mathbf{1}^T}{m}\| < 1$.*

The optimization problem (1) can be reformulated as the following equivalent multi-agent optimization problem:

$$\min_{x\in\mathbb{R}^{md}} f(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x_i) \text{ s.t. } x_1 = x_2 = \cdots = x_m, \quad (2)$$

where $x_i \in \mathbb{R}^d$ is agent i's decision variable and the collection of the agents' variables is $x = [x_1^T, x_2^T, \ldots, x_m^T]^T \in \mathbb{R}^{md}$.

We make the following assumption on objective functions:

Assumption 2. *Problem* (1) *has an optimal solution θ^*. The objective function $F(\cdot)$ is convex and each $f_i(\cdot)$ has Lipschitz continuous gradients over $\mathbb{R}^d$, i.e., for some $L > 0$,*

$$\|\nabla f_i(u) - \nabla f_i(v)\| \leq L\|u - v\|, \quad \forall i \in [m] \text{ and } \forall u, v \in \mathbb{R}^d.$$

Under Assumption 2, the optimization problem (2) has an optimal solution $x^* = [(\theta^*)^T, (\theta^*)^T, \ldots, (\theta^*)^T]^T \in \mathbb{R}^{md}$.

In the analysis of our methods, we use the following results:

Lemma 1 ([34], Lemma 11, page 50)**.** *Let $\{v^k\}$, $\{u^k\}$, $\{\alpha^k\}$, and $\{\beta^k\}$ be random nonnegative scalar sequences such that $\sum_{k=0}^{\infty}\alpha^k < \infty$ and $\sum_{k=0}^{\infty}\beta^k < \infty$ a.s. and*

$$\mathbb{E}\left[v^{k+1}|\mathcal{F}^k\right] \leq (1+\alpha^k)v^k - u^k + \beta^k, \quad \forall k \geq 0 \quad a.s.$$

where $\mathcal{F}^k = \{v^\ell, u^\ell, \alpha^\ell, \beta^\ell;\, 0 \leq \ell \leq k\}$. Then $\sum_{k=0}^{\infty} u^k < \infty$ and $\lim_{k\to\infty} v^k = v$ for a random variable $v \geq 0$ a.s.

Lemma 2. *Let $\{v^k\}$,$\{\alpha^k\}$, and $\{p^k\}$ be random nonnegative scalar sequences, and $\{q^k\}$ be a deterministic nonnegative scalar sequence satisfying $\sum_{k=0}^{\infty}\alpha^k < \infty$ a.s., $\sum_{k=0}^{\infty} q^k = \infty$, $\sum_{k=0}^{\infty} p^k < \infty$ a.s., and the following inequality*

$$\mathbb{E}\left[v^{k+1}|\mathcal{F}^k\right] \leq (1+\alpha^k-q^k)v^k + p^k, \quad \forall k \geq 0 \quad a.s.$$

where $\mathcal{F}^k = \{v^\ell, \alpha^\ell, p^\ell; 0 \leq \ell \leq k\}$. Then, $\sum_{k=0}^{\infty} q^k v^k < \infty$ and $\lim_{k\to\infty} v^k = 0$ hold a.s.

Proof. See proof in our extended version [1]. ■

Lemma 3. *Consider the problem $\min_{z\in\mathbb{R}^d}\phi(z)$, where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. Assume that the optimal solution set Z^* of the problem is nonempty. Let $\{z^k\}$ be a random sequence such that for any optimal solution $z^* \in Z^*$,*

$$\mathbb{E}\left[\|z^{k+1} - z^*\|^2|\mathcal{F}^k\right]$$
$$\leq (1+\alpha^k)\|z^k - z^*\|^2 - \eta^k\left(\phi(z^k) - \phi(z^*)\right) + \beta^k,\ \forall k \geq 0$$

holds a.s., where $\mathcal{F}^k = \{z^\ell, \alpha^\ell, \beta^\ell,\ \ell = 0, 1, \ldots, k\}$, $\{\alpha^k\}$ and $\{\beta^k\}$ are random nonnegative scalar sequences satisfying $\sum_{k=0}^{\infty}\alpha^k < \infty$, $\sum_{k=0}^{\infty}\beta^k < \infty$ a.s. , while $\{\eta^k\}$ is a deterministic nonnegative scalar sequence with $\sum_{k=0}^{\infty}\eta^k = \infty$. Then, $\{z^k\}$ converges a.s. to some solution $z^* \in Z^*$.

Proof. See proof in our extended version [1]. ■

Lemma 4. *Let $\{v^k\}$ be a nonnegative sequence, and $\{\alpha^k\}$ and $\{\beta^k\}$ be nonnegative scalar sequences satisfying $\sum_{k=0}^{\infty}\alpha^k = \infty$, $\lim_{k\to\infty}\alpha^k = 0$, and $\lim_{k\to\infty}\frac{\beta^k}{\alpha^k} = 0$. If there exists a $K \geq 0$ such that the following relation holds for all $k \geq K$:*

$$v^{k+1} \leq (1-\alpha^k)v^k + \beta^k,$$

then there always exits a constant C such that $v^k \leq C\frac{\beta^k}{\alpha^k}$ for all $k \geq K$.

Proof. See proof in our extended version [1]. ■

B. On differential privacy

We consider Laplace noise for DP. For a constant $\nu > 0$, $\mathrm{Lap}(\nu)$ denotes the Laplace distribution with probability density function $\frac{1}{2\nu}e^{-\frac{|x|}{\nu}}$. This distribution has mean zero and variance $2\nu^2$. Following [35], for the convenience of DP analysis, we represent the distributed optimization problem $\mathcal{P}$ in (1) by four parameters $(\mathcal{X}, \mathcal{S}, F, \mathcal{G}_W)$, where $\mathcal{X} = \mathbb{R}^n$ is the domain of optimization, $\mathcal{S} \subseteq \{\mathbb{R}^n \mapsto \mathbb{R}\}$ is a set of real-valued objective functions, with $f_i \in \mathcal{S}$, and $F(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x)$, and $\mathcal{G}_W$ is the induced graph by matrix W. Then we define adjacency as follows:

Definition 1. *Two distributed optimization problems $\mathcal{P}$ and $\mathcal{P}'$ are adjacent if the following conditions hold:*

- $\mathcal{X} = \mathcal{X}'$, $\mathcal{S} = \mathcal{S}'$, *and* $\mathcal{G}_W = \mathcal{G}'_W$, *i.e., the domain of optimization, the set of individual objective functions, and the communication graphs are identical;*
- *there exists an* $i \in [m]$ *such that* $f_i \neq f'_i$ *but* $f_j = f'_j$ *for all* $j \in [m]$, $j \neq i$.

It can be seen that two distributed optimization problems are adjacent if and only if one agent changes its individual objective function while all others parameters are identical.

Given a distributed optimization problem, we represent an execution of such an algorithm as $\mathcal{A}$, which is an infinite sequence of the optimization variables, i.e., $\mathcal{A} = \{x^0, x^1, \cdots\}$. We consider adversaries that can observe all communicated messages in the network. Therefore, the observation part of an execution is the infinite sequence of shared messages, which is represented by $\mathcal{O}$. Given a distributed optimization problem $\mathcal{P}$ and an initial state x^0, we define the observation mapping as $\mathcal{R}_{\mathcal{P},x^0}(\mathcal{A}) \triangleq \mathcal{O}$. Given a distributed optimization problem $\mathcal{P}$, observation sequence $\mathcal{O}$, and an initial state x^0, $\mathcal{R}^{-1}_{\mathcal{P},x^0}(\mathcal{O})$ is the set of executions $\mathcal{A}$ that can generate observation $\mathcal{O}$.

Definition 2. *(ϵ-DP [35]). For a given $\epsilon > 0$, an iterative algorithm for problem (1) is ϵ-differentially private if for any two adjacent $\mathcal{P}$ and $\mathcal{P}'$, any set of observation sequences $\mathcal{O}_s \subseteq \mathbb{O}$ (with $\mathbb{O}$ denoting the set of all possible observation sequences), and any initial state x^0, we always have*

$$\mathbb{P}[\mathcal{R}^{-1}_{\mathcal{P},x^0}(\mathcal{O}_s)] \leq e^{\epsilon}\mathbb{P}[\mathcal{R}^{-1}_{\mathcal{P}',x^0}(\mathcal{O}_s)], \tag{3}$$

where the probability $\mathbb{P}$ is taken over the randomness over iteration processes.

The definition of ϵ-DP ensures that an adversary having access to all shared messages in the network cannot gain information with a significant probability of any participating agent's objective function. It can also be seen that a smaller ϵ means a higher level of privacy protection.

III. The proposed algorithm

To achieve a strong DP, independent DP-noise should be injected in every round of message sharing and, hence, constantly affects the algorithm through inter-agent interactions, leading to significant reduction in optimization accuracy. Motivated by this observation, we propose to gradually weaken inter-agent interactions to reduce the influence of DP-noise on optimization accuracy. Interestingly, we prove that by judiciously designing the interaction weakening mechanism, we can still ensure convergence of all agents to a common optimal solution even in the presence of persistent DP-noise.

Algorithm 1: Differentially private distributed optimization

Parameters: Stepsize λ^k and weakening factor γ^k.

Every agent i maintains one state x_i^k, which is initialized with a random vector in $\mathbb{R}^d$.

for $k = 1, 2, \ldots$ **do**

a) Every agent j adds persistent DP-noise ζ_j^k to its state x_j^k, and then sends the obscured state $x_j^k + \zeta_j^k$ to agent $i \in \mathbb{N}_j$.

b) After receiving $x_j^k + \zeta_j^k$ from all $j \in \mathbb{N}_i$, agent i updates its state as follows:

$$x_i^{k+1} = x_i^k + \sum_{j\in\mathbb{N}_i} \gamma^k w_{ij}(x_j^k + \zeta_j^k - x_i^k) - \lambda^k \nabla f_i(x_i^k). \tag{4}$$

c) **end**

The sequence $\{\gamma^k\}$ diminishes with time and is used to suppress the influence of persistent DP-noise ζ_j^k on the convergence point of the iterates. The stepsize sequence $\{\lambda^k\}$ and attenuation sequence $\{\gamma^k\}$ have to be designed appropriately to guarantee the almost sure convergence of all $\{x_i^k\}$ to a common optimal solution θ^*. The persistent DP-noise processes $\{\zeta_i^k\}, i \in [m]$ have zero-mean and γ^k-bounded (conditional) variances, as specified below:

Assumption 3. *For every $i \in [m]$ and every k, conditional on the state x_i^k, the random noise ζ_i^k satisfies $\mathbb{E}\left[\zeta_i^k \mid x_i^k\right] = 0$ and $\mathbb{E}\left[\|\zeta_i^k\|^2 \mid x_i^k\right] = (\sigma_i^k)^2$ for all $k \geq 0$, and*

$$\sum_{k=0}^{\infty}(\gamma^k)^2 \max_{i\in[m]}(\sigma_i^k)^2 < \infty, \tag{5}$$

where $\{\gamma^k\}$ is the attenuation sequence from Algorithm 1. The initial random vectors satisfy $\mathbb{E}\left[\|x_i^0\|^2\right] < \infty$, $\forall i \in [m]$.

Remark 1. *Given that γ^k decreases with time, (5) can be satisfied even when $\{\sigma_i^k\}$ increases with time. For example, under $\gamma^k = \mathcal{O}(\frac{1}{k^{0.9}})$, an increasing $\{\sigma_i^k\}$ with increasing rate no faster than $\mathcal{O}(k^{0.3})$ still satisfies the summable condition in (5). Allowing $\{\sigma_i^k\}$ to increase with time is key to enabling the strong ϵ-DP, as elaborated later in Theorem 2.*

IV. Convergence analysis

We first extend Lemma 1 to deal with random vectors:

Lemma 5. *Let $\{\mathbf{v}^k\} \subset \mathbb{R}^d$ and $\{\mathbf{u}^k\} \subset \mathbb{R}^p$ be random nonnegative vector sequences, and $\{a^k\}$ and $\{b^k\}$ be random nonnegative scalar sequences such that*

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq (V^k + a^k\mathbf{1}\mathbf{1}^T)\mathbf{v}^k + b^k\mathbf{1} - H^k\mathbf{u}^k, \quad \forall k \geq 0$$

holds a.s., where $\{V^k\}$ and $\{H^k\}$ are random sequences of nonnegative matrices and $\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right]$ denotes the conditional expectation given $\mathbf{v}^\ell, \mathbf{u}^\ell, a^\ell, b^\ell, V^\ell, H^\ell$ for $\ell = 0, 1, \ldots, k$. Assume that $\{a^k\}$ and $\{b^k\}$ satisfy $\sum_{k=0}^{\infty} a^k < \infty$ and $\sum_{k=0}^{\infty} b^k < \infty$ a.s., and that there exists a (deterministic) vector $\pi > 0$ such that $\pi^T V^k \leq \pi^T$ and $\pi^T H^k \geq 0$ hold a.s. for all $k \geq 0$. Then, 1) $\{\pi^T\mathbf{v}^k\}$ converges to some random variable $\pi^T\mathbf{v} \geq 0$ a.s.; 2) $\{\mathbf{v}^k\}$ is bounded a.s., and 3) $\sum_{k=0}^{\infty}\pi^T H^k\mathbf{u}^k < \infty$ holds a.s.

Proof. See proof in our extended version [1]. ∎

Based on Lemma 3 and Lemma 5, we can prove the following general convergence results:

Lemma 6. *Assume that problem (1) has a solution. Suppose that a distributed algorithm generates sequences $\{x_i^k\} \subseteq \mathbb{R}^d$ such that a.s. we have for any optimal solution θ^*,*

$$\begin{aligned}
&\begin{bmatrix} \mathbb{E}\left[\|\bar{x}^{k+1}-\theta^*\|^2|\mathcal{F}^k\right] \\ \mathbb{E}\left[\sum_{i=1}^m \|x_i^{k+1}-\bar{x}^{k+1}\|^2|\mathcal{F}^k\right] \end{bmatrix} \\
&\le \left(\begin{bmatrix} 1 & \frac{\gamma^k}{m} \\ 0 & 1-\kappa\gamma^k \end{bmatrix} + a^k \mathbf{1}\mathbf{1}^T \right) \begin{bmatrix} \|\bar{x}^k-\theta^*\|^2 \\ \sum_{i=1}^m \|x_i^k-\bar{x}^k\|^2 \end{bmatrix} \\
&\quad + b^k\mathbf{1} - c^k \begin{bmatrix} F(\bar{x}^k)-F(\theta^*) \\ 0 \end{bmatrix}, \quad \forall k \ge 0
\end{aligned} \tag{6}$$

where $\bar{x}^k = \frac{1}{m}\sum_{i=1}^m x_i^k$, $\mathcal{F}^k = \{x_i^\ell,\, i \in [m],\, 0 \le \ell \le k\}$, the random nonnegative scalar sequences $\{a^k\}$, $\{b^k\}$ satisfy $\sum_{k=0}^\infty a^k < \infty$ and $\sum_{k=0}^\infty b^k < \infty$ a.s., the deterministic nonnegative sequences $\{c^k\}$ and $\{\gamma^k\}$ satisfy $\sum_{k=0}^\infty c^k = \infty$ and $\sum_{k=0}^\infty \gamma^k = \infty$, and the scalar $\kappa > 0$ satisfies $\kappa\gamma^k < 1$ for all $k \ge 0$. Then, we have $\lim_{k\to\infty}\|x_i^k - \bar{x}^k\| = 0$ a.s. for all i, and there is a solution $\tilde{\theta}^*$ such that $\lim_{k\to\infty}\|\bar{x}^k - \tilde{\theta}^*\| = 0$ a.s.

Proof. See Appendix A. ∎

Using Lemma 6, we are in position to establish convergence of Algorithm 1:

Theorem 1. *Under Assumption 1, Assumption 2, and Assumption 3, Algorithm 1 converges to a solution of problem (1) a.s. when nonnegative sequences $\{\gamma^k\}$ and $\{\lambda^k\}$ satisfy $\sum_{k=0}^\infty \gamma^k = \infty$, $\sum_{k=0}^\infty \lambda^k = \infty$, and $\sum_{k=0}^\infty \frac{(\lambda^k)^2}{\gamma^k} < \infty$.*

Proof. See Appendix B. ∎

Remark 2. *Communication imperfections can be modeled as channel noises [36], [37], which can be regarded as the DP-noise here. Therefore, Algorithm 1 can also counteract such communication imperfections in distributed optimization.*

Remark 3. *Because the evolution of x_i^k to the optimal solution satisfies the conditions in Lemma 6, we can leverage Lemma 6 to examine the convergence speed. The first relationship in (10) (i.e., $\sum_{k=0}^\infty \kappa\gamma^k \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 < \infty$) implies that $\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$ decreases to zero with a rate no slower than $\mathcal{O}(\frac{1}{k\gamma^k})$, and hence we have x_i^k converging to $\bar{x}^k$ no slower than $\mathcal{O}(\frac{1}{(k\gamma^k)^{0.5}})$. Moreover, given that a^k and b^k in (11) in Lemma 6's proof in the appendix are summable (and hence decrease to zero no slower than $\mathcal{O}(\frac{1}{k})$) and c^k in (11) corresponds to λ^k (which is square summable and hence decreases to zero no slower than $\mathcal{O}(\frac{1}{k^{0.5}})$), we have that $\bar{x}^k$ converges to an optimal solution with a speed no worse than $\mathcal{O}(\frac{1}{k^{0.5}})$ [34]. Therefore, the convergence of every x_i^k to an optimal solution, which is equivalent to the combination of the convergence of x_i^k to $\bar{x}^k$ and the convergence of $\bar{x}^k$ to an optimal solution, should be no slower than $\mathcal{O}(\frac{1}{(k\gamma^k)^{0.5}})$. (For example, under $\gamma^k = \mathcal{O}(\frac{1}{k^{0.6}})$, $\mathcal{O}(\frac{1}{(k\gamma^k)^{0.5}})$ is $\mathcal{O}(\frac{1}{k^{0.2}})$.)*

V. Privacy analysis

Similar to [35], we define the sensitivity of an algorithm to problem (1) as follows:

Definition 3. *At each iteration k, any initial state x^0 and any adjacent distributed optimization problems $\mathcal{P}$ and $\mathcal{P}'$, the sensitivity of an algorithm is*

$$\Delta^k \triangleq \sup_{\mathcal{O}\in\mathbb{O}} \left\{ \sup_{x\in\mathcal{R}_{\mathcal{P},x^0}^{-1}(\mathcal{O}),\, x'\in\mathcal{R}_{\mathcal{P}',x^0}^{-1}(\mathcal{O})} \|x^{k+1} - x'^{k+1}\|_1 \right\}. \tag{7}$$

Lemma 7. *At each iteration k, if each agent adds a noise vector $\zeta_i^k \in \mathbb{R}^p$ consisting of p independent Laplace noises with parameter ν^k such that $\sum_{k=1}^T \frac{\Delta^k}{\nu^k} \le \epsilon$, then Algorithm 1 is ϵ-differentially private for iterations from $k = 1$ to $k = T+1$.*

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

As indicated in [35], since the change of an objective function f_i can be arbitrary in Definition 1, we have to make the following assumption to ensure bounded sensitivity:

Assumption 4. *The gradients of all individual objective functions are bounded, i.e., there exists a constant C such that $\|\nabla f_i(x)\|_1 \le C$ holds for all $x \in \mathbb{R}^p$ and $1 \le i \le m$.*

Theorem 2. *Under Assumptions 1, 2, 4, if nonnegative sequences $\{\lambda^k\}$ and $\{\gamma^k\}$ satisfy the conditions in Theorem 1, and all elements of ζ_i^k are drawn independently from Laplace distribution $\mathrm{Lap}(\nu^k)$ with $(\sigma_i^k)^2 = 2(\nu^k)^2$ satisfying Assumption 3, then all agents in Algorithm 1 will converge a.s. to an optimal solution. Moreover,*

1) *Algorithm 1 is ϵ-differentially private with the cumulative privacy budget bounded by $\epsilon \le \sum_{k=1}^T \frac{2C\lambda^k}{\nu^k}$ for iterations from $k = 1$ to $k = T+1$ where C is from Assumption 4. And the cumulative privacy budget is always finite for $T \to \infty$ when the sequence $\{\frac{\lambda^k}{\nu^k}\}$ is summable;*
2) *Suppose that two sequences $\{\nu'^k\}$ and $\{\lambda^k\}$ have a finite sequence-ratio sum $\Phi_{\lambda,\nu'} \triangleq \sum_{k=1}^\infty \frac{\lambda^k}{\nu'^k}$. Then setting the Laplace noise parameter ν^k as $\nu^k = \frac{2C\Phi_{\lambda,\nu'}}{\epsilon}\nu'^k$ ensures that Algorithm 1 is ϵ-differentially private for any $\epsilon > 0$ even when the number of iterations goes to infinity;*
3) *In the special case where $\lambda^k = \frac{1}{k}$ and $\gamma^k = \frac{1}{k^{0.9}}$, setting $\nu^k = \frac{2C\Phi}{\epsilon}k^{0.3}$ with $\Phi \triangleq \sum_{k=1}^\infty \frac{1}{k^{1.3}} \approx 3.93$ (which can be verified to satisfy Assumption 3) ensures that Algorithm 1 is always ϵ-differentially private for any $\epsilon > 0$ even when the number of iterations goes to infinity.*

Proof. Since the Laplace noise satisfies Assumption 3, the convergence results follow naturally from Theorem 1.

To prove the three statements on privacy, we first prove that the sensitivity of the algorithm satisfies $\Delta^k \le 2C\lambda^k$. Given two adjacent distributed optimization problems $\mathcal{P}$ and $\mathcal{P}'$, for any given fixed observation $\mathcal{O}$ and initial state x^0, the sensitivity is determined by $\|\mathcal{R}_{\mathcal{P},x^0}^{-1}(\mathcal{O}) - \mathcal{R}_{\mathcal{P}',x^0}^{-1}(\mathcal{O})\|_1$ according to Definition 3. Since in $\mathcal{P}$ and $\mathcal{P}'$, there is only one objective function that is different, we represent this different objective function as the ith one, i.e., f_i in $\mathcal{P}$ and f_i' in $\mathcal{P}'$, without loss of generality. We define $o^k \triangleq x_i^k + \sum_{j\in\mathbb{N}_i^{\text{in}}} \gamma^k w_{ij}(x_j^k - x_i^k)$ and $o'^k \triangleq x_i'^k + \sum_{j\in\mathbb{N}_i^{\text{in}}} \gamma^k w_{ij}(x_j'^k - x_i'^k)$, which are accessible

to adversaries under f_i and f_i', respectively. Because the observations under $\mathcal{P}$ and $\mathcal{P}'$ are identical, we have that o^k and o'^k should be the same according to the definition of sensitivity. Therefore, we have the following relationship:

$$\begin{aligned}&\|\mathcal{R}^{-1}_{\mathcal{P},x^0}(\mathcal{O}) - \mathcal{R}^{-1}_{\mathcal{P}',x^0}(\mathcal{O})\|_1\\ =&\left\|o^k - \lambda^k\nabla f_i(x_i^k) - \left(o'^k - \lambda^k\nabla f_i'(x_i'^k)\right)\right\|_1\\ =&\left\|\lambda^k\nabla f_i(x_i^k) - \lambda^k\nabla f_i'(x_i'^k)\right\|_1 \le 2C\lambda^k,\end{aligned} \tag{8}$$

where the last inequality is obtained using Assumption 4.

Using Lemma 7, we can easily obtain $\epsilon \le \sum_{k=1}^{T}\frac{2C\lambda^k}{\nu^k}$. Hence, ϵ will always be finite even when T tends to infinity if the sequence $\{\frac{\lambda^k}{\nu^k}\}$ is summable, i.e., $\sum_{k=0}^{\infty}\frac{\lambda^k}{\nu^k} < \infty$.

By scaling ν^k proportionally and using the linear relationship between ϵ and $\frac{1}{\nu^k}$, the second statement can be easily obtained based on the first statement. The third statement can also be easily proven by specializing the selection of λ^k, γ^k, and ν^k sequences. ■

Different from [35] which has to use a summable stepsize (geometrically-decreasing stepsize, more specifically) to ensure a finite privacy budget ϵ when $k \to \infty$, here we ensure a finite ϵ even when the stepsize sequence is non-summable. Allowing stepsize sequences to be non-summable is key to avoiding optimization errors in [35] and achieve almost sure convergence. In fact, to our knowledge, this is the first time that almost-sure convergence is achieved under rigorous ϵ-DP for an infinite number of iterations.

Remark 4. *In Theorem 2, to ensure that the privacy budget $\epsilon = \sum_{k=1}^{\infty}\frac{2C\lambda^k}{\nu^k}$ is finite even when $k \to \infty$, the Laplace noise parameter ν^k has to increase with time since $\{\lambda^k\}$ is non-summable. An increasing ν^k will make the relative level between noise ζ_i^k and signal x_i^k increase with time. However, since the increase in ν^k is outweighed by the decrease of γ^k (see Assumption 3), the actual noise fed into the algorithm, i.e., $\gamma^k\mathrm{Lap}(\nu^k)$, still decays with time, which makes it possible for Algorithm 1 to ensure a.s. convergence to an optimal solution. Moreover, according to Theorem 1, such a.s. convergence is not affected by scaling ν^k by any constant coefficient $\frac{1}{\epsilon} > 0$ so as to achieve any desired level of ϵ-DP, as long as the Laplace noise parameter ν^k (with associated variance $(\sigma_i^k)^2 = 2(\nu^k)^2$) satisfies Assumption 3.*

VI. Numerical Experiments

We evaluate the performance of the proposed algorithm using a canonical distributed estimation problem where a network of m sensors collectively estimate an unknown parameter $\theta \in \mathbb{R}^d$. More specifically, we assume that each sensor i has a noisy measurement of the parameter, $z_i = M_i\theta + w_i$, where $M_i \in \mathbb{R}^{s\times d}$ is the measurement matrix of agent i and w_i is Gaussian measurement noise of unit variance. Then the maximum likelihood estimation of parameter θ can be solved using the optimization problem formulated as (1), with each $f_i(\theta)$ given as $f_i(\theta) = \|z_i - M_i\theta\|^2 + \varsigma\|\theta\|^2$ where ς is a regularization parameter [5].

We consider $m = 5$ sensors interacting on a randomly generated connected graph. In the evaluation, we set $s = 3$ and $d = 2$. To evaluate the performance of the proposed Algorithm, we injected Laplace based DP-noise with parameter $\nu^k = 1 + 0.1k^{0.3}$ in every message shared in all iterations. We set the stepsize λ^k and diminishing sequence γ^k as $\lambda^k = \frac{0.02}{1+0.1k}$ and $\gamma^k = \frac{1}{1+0.1k^{0.9}}$, respectively, which satisfy the conditions in Theorem 1 and Theorem 2. In the evaluation, we ran our algorithm for 100 times and calculated the average as well as the variance of the optimization error as a function of the iteration index. The result is given by the blue curve and error bars in Fig. 1. For comparison, we also ran the existing distributed gradient descent (DGD) approach in [3] under the same noise, and the differential-privacy approach for distributed optimization (PDOP) in [12] under the same privacy budget. Note that PDOP uses geometrically decreasing stepsizes (which are summable) to ensure a finite privacy budget, but the fast decreasing stepsize also leads to optimization errors. The evolution of the average optimization error and variance of the DGD and PDOP approaches are given by the red and black curves/error bars in Fig. 1, respectively. It is clear that the proposed algorithm has a comparable convergence speed but much better optimization accuracy.

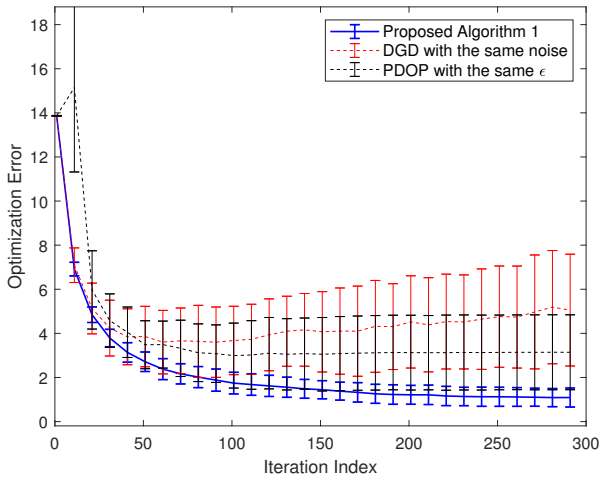

Fig. 1. Comparison of Algorithm 1 with existing distributed gradient descent algorithm (DGD) in [3] (under the same noise) and the differential-privacy approach for decentralized optimization PDOP in [12] (under the same privacy budget) using the distributed estimation problem

VII. Conclusions and Discussions

Although DP is becoming the de facto standard for publicly sharing information, its direct incorporation into distributed optimization leads to a trade-off between privacy and optimization accuracy. This paper proposes a distributed optimization algorithm that ensures both ϵ-DP and optimization accuracy. The simultaneous achievement of both provable convergence to the optimal solution and rigorous ϵ-DP with guaranteed finite cumulative privacy budget, to our knowledge, has not been reported before in distributed optimization. Numerical simulation results confirm the effectiveness of the proposed algorithm.

It is worth noting that our simultaneous achievement of both provable convergence to the optimal solution and ϵ-DP

does not contradict the fundamental theory and limitations of DP in [24]. Firstly, according to the DP theory, conventional query mechanisms on a dataset can achieve ϵ-DP only by sacrificing query accuracies. However, the distributed optimization algorithm does not correspond to a simple query mechanism on the optimal solution. Instead, what are queried in every iteration of distributed optimization are individual objective functions (gradients), and revealing the precise optimal solution is not equivalent to revealing accurate objective functions (the actual query target). In fact, in the language of machine learning, distributed optimization can be viewed as the empirical risk minimization problem, and the obtained optimal solution corresponds to the optimal model parameter in machine learning. On pages 216-218 of [24], the authors explicitly state that "the constraint of privacy is not necessarily at odds with the goals of machine learning, both of which aim to extract information from the distribution from which the data was drawn, rather than from individual data points," and "we are often able to perform private machine learning nearly as accurately, with nearly the same number of samples, as we can perform non-private machine learning." Secondly, the achievement of ϵ-DP does incur utility cost. More specifically, in order to reduce ϵ to enhance privacy, we can use a faster-increasing $\{\nu^k\}$ according to Theorem 2, which requires $\{\gamma^k\}$ to decrease faster according to Assumption 3. Given that $\{\gamma^k\}$ cannot decrease faster than $\mathcal{O}(\frac{1}{k})$, and the convergence speed is determined by $\mathcal{O}(\frac{1}{k\gamma^k})$ according to Remark 3, we arrive at the conclusion that a faster decreasing $\{\gamma^k\}$ corresponds to a stronger privacy level but a slower convergence speed.

APPENDIX

A. Proof of Lemma 6

Let θ^* be an arbitrary but fixed optimal solution of problem (1). Then, we have $F(\bar{x}^k) - F(\theta^*) \geq 0$ for all k. Hence, by letting $\mathbf{v}^k = \left[\|\bar{x}^k - \theta^*\|^2,\ \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2\right]^T$, from relation (6) it follows *a.s.* that for all $k \geq 0$,

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq \left(\begin{bmatrix} 1 & \frac{\gamma^k}{m} \\ 0 & 1-\kappa\gamma^k \end{bmatrix} + a^k \mathbf{1}\mathbf{1}^T\right)\mathbf{v}^k + b^k\mathbf{1}. \quad (9)$$

Consider the vector $\pi = [1, \frac{1}{m\kappa}]^T$ and note $\pi^T \begin{bmatrix} 1 & \frac{\gamma^k}{m} \\ 0 & 1-\kappa\gamma^k \end{bmatrix} = \pi^T$. Thus, relation (9) satisfies all conditions of Lemma 5. So it follows that $\lim_{k\to\infty} \pi^T\mathbf{v}^k$ exists *a.s.*, and that the sequences $\{\|\bar{x}^k - \theta^*\|^2\}$ and $\{\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2\}$ are bounded *a.s.* From (9) we have the following relation *a.s.* for the second element of $\mathbf{v}^k$:

$$\mathbb{E}\left[\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2|\mathcal{F}^k\right] \leq (1+a^k-\kappa\gamma^k)\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + \beta^k,$$

where $\beta^k = a^k\left(\|\bar{x}^k - \theta^*\|^2 + \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2\right)$. Since $\sum_{k=0}^\infty a^k < \infty$ *a.s.* by our assumption, and the sequences $\{\|\bar{x}^k - \theta^*\|^2\}$ and $\{\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2\}$ are bounded *a.s.*, it follows that $\sum_{k=0}^\infty \beta^k < \infty$ *a.s.* Thus, the preceding relation satisfies the conditions of Lemma 2 with $v^k = \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$, $q^k = \kappa\gamma^k$, and $p^k = \beta^k$ due to our assumptions $\sum_{k=0}^\infty b^k < \infty$ *a.s.* and $\sum_{k=0}^\infty \gamma^k = \infty$. So one yields *a.s.*

$$\sum_{k=0}^\infty \kappa\gamma^k \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 < \infty,\ \lim_{k\to\infty} \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 = 0. \quad (10)$$

It remains to show that $\|\bar{x}^k - \theta^*\|^2 \to 0$ *a.s.* For this, we consider relation (6) and focus on the first element of $\mathbf{v}^k$, for which we obtain *a.s.* for all $k \geq 0$:

$$\mathbb{E}\left[\|\bar{x}^{k+1} - \theta^*\|^2|\mathcal{F}^k\right] \leq (1+a^k)\|\bar{x}^k - \theta^*\|^2 + \left(\frac{\gamma^k}{m} + a^k\right)\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + b^k - c^k(F(\bar{x}^k) - F(\theta^*)). \quad (11)$$

The preceding relation satisfies Lemma 3 with $\phi = F$, $z^* = \theta^*$, $z^k = \bar{x}^k$, $\alpha^k = a^k$, $\eta^k = c^k$, and $\beta^k = (\frac{\gamma^k}{m} + a^k)\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + b^k$. By our assumptions, the sequences $\{a^k\}$ and $\{b^k\}$ are summable *a.s.*, and $\sum_{k=0}^\infty c^k = \infty$. In view of (10), it follows that $\sum_{k=0}^\infty \beta^k < \infty$ *a.s.* Hence, all the conditions of Lemma 3 are satisfied and, consequently, $\{\bar{x}^k\}$ converges *a.s.* to some optimal solution.

B. Proof of Theorem 1

The basic idea is to apply Lemma 6 to the quantities $\mathbb{E}\left[\|\bar{x}^{k+1} - \theta^*\|^2|\mathcal{F}^k\right]$ and $\mathbb{E}\left[\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2|\mathcal{F}^k\right]$. We divide the proof into two parts to analyze $\|\bar{x}^{k+1} - \theta^*\|^2$ and $\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2$, respectively.

Part I: We first analyze $\|\bar{x}^{k+1} - \theta^*\|^2$. For the sake of notational simplicity, we represent $\nabla f_i(x_i^k)$ as g_i^k. Stacking x_i^k and g_i^k into augmented vectors $(x^k)^T = \left[(x_1^k)^T, \cdots, (x_m^k)^T\right]$ and $(g^k)^T = \left[(g_1^k)^T, \cdots, (g_m^k)^T\right]$, respectively, we can write the dynamics of Algorithm 1 as

$$x^{k+1} = (I + \gamma^k W \otimes I_d)x^k + \gamma^k\zeta_w^k - \lambda^k g^k, \quad (12)$$

where $\otimes$ denotes the Kronecker product, and $(\zeta_w^k)^T = \left[(\zeta_{w1}^k)^T, \cdots, (\zeta_{wm}^k)^T\right]$ with $\zeta_{wi}^k \triangleq \sum_{j\in\mathbb{N}_i} w_{ij}\zeta_j^k$.

From (12) we can obtain the following relationship for the average vector $\bar{x}^k = \frac{1}{m}\sum_{i=1}^m x_i^k$:

$$\bar{x}^{k+1} = \bar{x}^k + \gamma^k\bar{\zeta}_w^k - \frac{\lambda^k}{m}\sum_{i=1}^m g_i^k, \quad (13)$$

where $\bar{\zeta}_w^k = \frac{1}{m}\sum_{i=1}^m \zeta_{wi}^k = \frac{1}{m}\sum_{i=1}^m \sum_{j\in\mathbb{N}_i} w_{ij}\zeta_j^k = -\frac{\sum_{i=1}^m w_{ii}\zeta_i^k}{m}$ (note $w_{ii} \triangleq -\sum_{j\in\mathbb{N}_i} w_{ij}$).

Using (13) and the preceding relation, we relate $\bar{x}^k$ to an optimal solution:

$$\bar{x}^{k+1} - \theta^* = \bar{x}^k - \theta^* - \frac{1}{m}\sum_{i=1}^m \left(\lambda^k g_i^k + \gamma^k w_{ii}\zeta_i^k\right),$$

which further implies

$$\begin{aligned}\left\|\bar{x}^{k+1}-\theta^*\right\|^2=&\left\|\bar{x}^k-\theta^*\right\|^2-\frac{2}{m}\sum_{i=1}^m\left\langle\lambda^k g_i^k+\gamma^k w_{ii}\zeta_i^k,\bar{x}^k-\theta^*\right\rangle\\&+\frac{1}{m^2}\left\|\sum_{i=1}^m\left(\lambda^k g_i^k+\gamma^k w_{ii}\zeta_i^k\right)\right\|^2\\\le&\left\|\bar{x}^k-\theta^*\right\|^2-\frac{2}{m}\sum_{i=1}^m\left\langle\lambda^k g_i^k+\gamma^k w_{ii}\zeta_i^k,\bar{x}^k-\theta^*\right\rangle\\&+\frac{2}{m^2}\left\|\sum_{i=1}^m\lambda^k g_i^k\right\|^2+\frac{2}{m^2}\left\|\sum_{i=1}^m\gamma^k w_{ii}\zeta_i^k\right\|^2.\end{aligned}$$

Taking the conditional expectation, given $\mathcal{F}^k=\{x^0,\ldots,x^k\}$, and using the assumption that the noise ζ_i^k is with zero mean and variance $(\sigma_i^k)^2$ conditionally on x_i^k (see Assumption 3), from the preceding relation we obtain *a.s.* for all $k\ge 0$,

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{x}^{k+1}-\theta^*\right\|^2|\mathcal{F}^k\right]\\&\le\left\|\bar{x}^k-\theta^*\right\|^2-\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k,\bar{x}^k-\theta^*\right\rangle\\&\quad+\frac{2}{m^2}(\lambda^k)^2\left\|\sum_{i=1}^m g_i^k\right\|^2+\frac{2}{m}(\gamma^k)^2\sum_{i=1}^m w_{ii}^2(\sigma_i^k)^2.\end{aligned}\tag{14}$$

We next estimate the inner product term, for which we have

$$\begin{aligned}\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k,\bar{x}^k-\theta^*\right\rangle=&\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k-\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle\\&+\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle.\end{aligned}\tag{15}$$

Recalling that $g_i^k=\nabla f_i(x_i^k)$, by the Lipschitz continuous property of $\nabla f_i(\cdot)$, we have

$$\begin{aligned}&\lambda^k\left\langle g_i^k-\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle\ge-L\lambda^k\|x_i^k-\bar{x}^k\|\|\bar{x}^k-\theta^*\|\\&\ge-\frac{\gamma^k}{2}\|x_i^k-\bar{x}^k\|^2-\frac{L^2(\lambda^k)^2}{2\gamma^k}\|\bar{x}^k-\theta^*\|^2.\end{aligned}\tag{16}$$

By the convexity of $F(\cdot)$, we have

$$\begin{aligned}\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle&=2\lambda^k\left\langle\nabla F(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle\\&\ge 2\lambda^k(F(\bar{x}^k)-F(\theta^*)).\end{aligned}\tag{17}$$

Combining (15), (16), and (17) leads to

$$\begin{aligned}\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k,\bar{x}^k-\theta^*\right\rangle\ge&-\frac{\gamma^k}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2\\&-\frac{L^2(\lambda^k)^2}{\gamma^k}\|\bar{x}^k-\theta^*\|^2+2\lambda^k(F(\bar{x}^k)-F(\theta^*)).\end{aligned}\tag{18}$$

We next estimate the second last term in (14):

$$\begin{aligned}\frac{1}{m^2}\left\|\sum_{i=1}^m g_i^k\right\|^2=&\frac{1}{m^2}\left\|\sum_{i=1}^m\left(g_i^k-\nabla f_i(\theta^*)\right)\right\|^2\\&\le\frac{L^2}{m}\sum_{i=1}^m\left\|x_i^k-\theta^*\right\|^2=\frac{L^2}{m}\|x^k-x^*\|^2.\end{aligned}\tag{19}$$

Further using the inequality

$$\begin{aligned}\|x^k-x^*\|^2&\le\|x^k-\mathbf{1}\otimes\bar{x}^k+\mathbf{1}\otimes\bar{x}^k-x^*\|^2\\&\le 2\|x^k-\mathbf{1}\otimes\bar{x}^k\|^2+2\|\mathbf{1}\otimes\bar{x}^k-x^*\|^2\\&\le 2\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2+2m\|\bar{x}^k-\theta^*\|^2,\end{aligned}\tag{20}$$

we have from (19) that

$$\frac{1}{m^2}\left\|\sum_{i=1}^m g_i^k\right\|^2\le\frac{2L^2}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2+2L^2\|\bar{x}^k-\theta^*\|^2.\tag{21}$$

Substituting (18) and (21) into (14) yields

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{x}^{k+1}-\theta^*\right\|^2|\mathcal{F}^k\right]\le&\left\|\bar{x}^k-\theta^*\right\|^2+\frac{\gamma^k}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2\\&+L^2(\lambda^k)^2\left(\frac{1}{\gamma^k}+4\right)\|\bar{x}^k-\theta^*\|^2-2\lambda^k(F(\bar{x}^k)-F(\theta^*))\\&+\frac{4L^2(\lambda^k)^2}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2+\frac{2(\gamma^k)^2}{m}\sum_{i=1}^m w_{ii}^2(\sigma_i^k)^2.\end{aligned}\tag{22}$$

Part II: Next we analyze $\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|^2$. Using (12) and (13), we obtain

$$\begin{aligned}x^{k+1}-\mathbf{1}\otimes\bar{x}^{k+1}=&(I+\gamma^k W\otimes I_d)x^k-\mathbf{1}\otimes\bar{x}^k\\&+\gamma^k\left(\zeta_w^k-\frac{1}{m}\sum_{i=1}^m\mathbf{1}\otimes\zeta_{w,i}^k\right)-\lambda^k\left(g^k-\frac{1}{m}\sum_{i=1}^m\mathbf{1}\otimes g_i^k\right).\end{aligned}$$

Noting $\mathbf{1}\otimes\bar{x}^k=\frac{1}{m}\left(\mathbf{1}\mathbf{1}^T\otimes I_d\right)x^k$, $\sum_{i=1}^m\mathbf{1}\otimes\zeta_{w,i}^k=\left(\mathbf{1}\mathbf{1}^T\otimes I_d\right)\zeta_w^k$, and $\sum_{i=1}^m\mathbf{1}\otimes g_i^k=\left(\mathbf{1}\mathbf{1}^T\otimes I_d\right)g^k$, we can rewrite the preceding equality as

$$x^{k+1}-\mathbf{1}\otimes\bar{x}^{k+1}=\hat{W}_k x^k+\gamma^k\Xi\zeta_w^k-\lambda^k\Xi g^k,\tag{23}$$

with $\hat{W}_k\triangleq\left(I+\gamma^k W-\frac{1}{m}\mathbf{1}\mathbf{1}^T\right)\otimes I_d$ and $\Xi\triangleq\left(I-\frac{1}{m}\mathbf{1}\mathbf{1}^T\right)\otimes I_d$.

Since $\left(I+\gamma^k W-\frac{1}{m}\mathbf{1}\mathbf{1}^T\right)\mathbf{1}=0$ holds and we always have $(A\otimes B)(C\otimes D)=(AC)\otimes(BD)$, it follows that

$$\hat{W}_k\left(\mathbf{1}\otimes\bar{x}^k\right)=\left(\left(I+\gamma^k W-\frac{\mathbf{1}\mathbf{1}^T}{m}\right)\times\mathbf{1}\right)\otimes\left(I_d\times\bar{x}^k\right)=0.$$

By subtracting $\hat{W}_k\left(\mathbf{1}\otimes\bar{x}^k\right)=0$ from the right hand side of (23), we obtain

$$x^{k+1}-\mathbf{1}\otimes\bar{x}^{k+1}=\hat{W}_k\left(x^k-\mathbf{1}\otimes\bar{x}^k\right)+\gamma^k\Xi\zeta_w^k-\lambda^k\Xi g^k,$$

which further leads to

$$\begin{aligned}&\|x^{k+1}-\mathbf{1}\otimes\bar{x}^{k+1}\|^2\\&=\|\hat{W}_k(x^k-\mathbf{1}\otimes\bar{x}^k)-\lambda^k\Xi g^k\|^2+\|\gamma^k\Xi\zeta_w^k\|^2\\&\quad+2\left\langle\hat{W}_k(x^k-\mathbf{1}\otimes\bar{x}^k)-\lambda^k\Xi g^k,\gamma^k\Xi\zeta_w^k\right\rangle\\&\le\|\hat{W}_k(x^k-\mathbf{1}\otimes\bar{x}^k)-\lambda^k\Xi g^k\|^2+m(\gamma^k)^2\sum_{i=1}^m\sum_{j\in\mathbb{N}_i}w_{ij}^2\|\zeta_j^k\|^2\\&\quad+2\left\langle\hat{W}_k(x^k-\mathbf{1}\otimes\bar{x}^k)-\lambda^k\Xi g^k,\gamma^k\Xi\zeta_w^k\right\rangle,\end{aligned}$$

where the inequality follows from $\|\Xi\| = 1$ and the definition $\zeta_{wi}^k \triangleq \sum_{j\in\mathbb{N}_i} w_{ij}\zeta_j^k$. Taking the conditional expectation with respect to $\mathcal{F}^k = \{x^0, \ldots, x^k\}$ and using Assumption 3 yield

$$\mathbb{E}\left[\|x^{k+1} - \mathbf{1}\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k\right] \leq \left(\|\hat{W}_k(x^k - \mathbf{1}\otimes\bar{x}^k)\| + \|\lambda^k\Xi g^k\|\right)^2 + m(\gamma^k)^2\max_{j\in[m]}(\sigma_j^k)^2 C_W,$$

where $C_W = \sum_{i=1}^m \sum_{j\in\mathbb{N}_i} w_{ij}^2$. Using the fact $\|\Xi\| = 1$ and $\|\hat{W}_k\| = \|I + \gamma^k W - \frac{1}{m}\mathbf{1}\mathbf{1}^T\| = 1 - \gamma^k|\nu|$ where $-\nu$ is some non-zero eigenvalue of W (see Assumption 1), we obtain

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1} - \mathbf{1}\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k\right] &\leq (1-\gamma^k|\nu|)^2(1+\epsilon)\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2 \\ &\quad + (1+\epsilon^{-1})(\lambda^k)^2\|g^k\|^2 + m(\gamma^k)^2\max_{j\in[m]}(\sigma_j^k)^2 C_W\end{aligned} \tag{24}$$

for any $\epsilon > 0$, where we used $(a+b)^2 \leq (1+\epsilon)a^2 + (1+\epsilon^{-1})b^2$ valid for any scalars a, b, and $\epsilon > 0$.

We next focus on estimating the term involving the gradient g^k in the preceding inequality. Noting $g^k = m\nabla f(x^k)$ and that $f(\cdot)$ has Lipschitz continuous gradients (with Lipschitz constant $\frac{L}{m}$), we have

$$\begin{aligned}\|g^k\|^2 &= m^2\|\nabla f(x^k) - \nabla f(x^*) + \nabla f(x^*)\|^2 \\ &\leq 2m^2\|\nabla f(x^k) - \nabla f(x^*)\|^2 + 2m^2\|\nabla f(x^*)\|^2 \\ &\leq 2L^2\|x^k - x^*\|^2 + 2m^2\|\nabla f(x^*)\|^2.\end{aligned}$$

Since $x^* = \mathbf{1}\otimes\theta^*$, using the relationship in (20), we obtain

$$\|g^k\|^2 \leq 4L^2(\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2 + m\|\bar{x}^k - \theta^*\|^2) + 2m^2\|\nabla f(x^*)\|^2.$$

Finally, substituting the preceding relation back in (24) yields

$$\begin{aligned}&\mathbb{E}\left[\|x^{k+1} - \mathbf{1}\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k\right] \\ &\leq (1-\gamma^k|\nu|)^2(1+\epsilon)\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2 \\ &+ 4(1+\epsilon^{-1})L^2(\lambda^k)^2(\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2 + m\|\bar{x}^k - \theta^*\|^2) \\ &+ 2(1+\epsilon^{-1})(\lambda^k)^2m^2\|\nabla f(x^*)\|^2 + m(\gamma^k)^2\max_{j\in[m]}(\sigma_j^k)^2 C_W.\end{aligned}$$

By letting $\epsilon = \frac{\gamma^k|\nu|}{1-\gamma^k|\nu|}$ and consequently $1+\epsilon = (1-\gamma^k|\nu|)^{-1}$ and $1 + \epsilon^{-1} = (\gamma^k|\nu|)^{-1}$, we arrive at

$$\begin{aligned}&\mathbb{E}\left[\|x^{k+1} - \mathbf{1}\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k\right] \\ &\leq \left(1 - \gamma^k|\nu| + \frac{4L^2(\lambda^k)^2}{|\nu|\gamma^k}\right)\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2 \\ &+ \frac{4mL^2(\lambda^k)^2}{|\nu|\gamma^k}\|\bar{x}^k - \theta^*\|^2 + \frac{4(\lambda^k)^2m^2}{|\nu|\gamma^k}\|\nabla f(x^*)\|^2 \\ &+ m(\gamma^k)^2\max_{j\in[m]}(\sigma_j^k)^2 C_W.\end{aligned} \tag{25}$$

By combining (22) and (25), and using Assumption 3, we have $\mathbb{E}\left[\|\bar{x}^{k+1} - \theta^*\|^2|\mathcal{F}^k\right]$ and $\mathbb{E}\left[\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2|\mathcal{F}^k\right]$ satisfying the conditions of Lemma 6 with $\kappa = |\nu|$, $c^k = 2\lambda^k$, $a^k = \max\{L^2(\lambda^k)^2\left(\frac{1}{\gamma^k} + 4\right), \frac{4mL^2(\lambda^k)^2}{|\nu|\gamma^k}\}$, and $b^k = (\gamma^k)^2\max\{\frac{2}{m}\sum_{i=1}^m w_{ii}^2(\sigma_i^k)^2, \frac{4(\lambda^k)^2m^2}{|\nu|\gamma^k}\|\nabla f(x^*)\|^2 + m\max_{j\in[m]}(\sigma_j^k)^2 C_W\}$ where $C_W = \sum_{i=1}^m\sum_{j\in\mathbb{N}_i} w_{ij}^2$.

REFERENCES

[1] Y. Wang and A. Nedic, "Tailoring gradient methods for differentially-private distributed optimization," *arXiv preprint arXiv:2202.01113*, 2022.

[2] J. N. Tsitsiklis, "Problems in decentralized decision making and computation." MIT, Tech. Rep., 1984.

[3] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[4] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[5] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of asynchronous distributed gradient methods over stochastic networks," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 434–448, 2017.

[6] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2017.

[7] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 315–320, 2018.

[8] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push-pull gradient methods for distributed optimization in networks," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 1–16, 2021.

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

[10] C. Zhang, M. Ahmad, and Y. Wang, "ADMM based privacy-preserving decentralized optimization," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 565–580, 2019.

[11] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed Newton method for network utility maximization–I: Algorithm," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2162–2175, 2013.

[12] Z. Huang, S. Mitra, and N. Vaidya, "Differentially private distributed optimization," in *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, 2015, pp. 1–10.

[13] D. A. Burbano-L, J. George, R. A. Freeman, and K. M. Lynch, "Inferring private information in wireless sensor networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 4310–4314.

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

[15] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 774–14 784.

[16] C. Zhang and Y. Wang, "Enabling privacy-preservation in decentralized optimization," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 2, pp. 679–689, 2018.

[17] N. M. Freris and P. Patrinos, "Distributed computing over encrypted data," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 1116–1122.

[18] Y. Lu and M. Zhu, "Privacy preserving distributed optimization using homomorphic encryption," *Automatica*, vol. 96, pp. 314–325, 2018.

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

[20] S. Gade and N. H. Vaidya, "Private optimization on networks," in *American Control Conference*. IEEE, 2018, pp. 1402–1409.

[21] H. Gao, Y. Wang, and A. Nedić, "Dynamics based privacy preservation in decentralized optimization," *Automatica*, 2023.

[22] Y. Wang and A. Nedic, "Decentralized gradient methods with time-varying uncoordinated stepsizes: Convergence analysis and privacy design," *arXiv preprint arXiv:2205.10934*, 2022.

[23] Y. Wang and T. Başar, "Quantization enabled privacy protection in decentralized stochastic optimization," *IEEE Transactions on Automatic Control*, 2022.

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

[25] S. Han, U. Topcu, and G. J. Pappas, "Differentially private distributed constrained optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 50–64, 2016.

[26] M. T. Hale and M. Egerstedt, "Cloud-enabled differentially private multiagent optimization with constraints," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 4, pp. 1693–1706, 2017.
[27] Y. Wang, Z. Huang, S. Mitra, and G. E. Dullerud, "Differential privacy in linear distributed control systems: Entropy minimizing mechanisms and performance tradeoffs," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 118–130, 2017.
[28] X. Zhang, M. M. Khalili, and M. Liu, "Recycled admm: Improving the privacy and accuracy of distributed algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1723–1734, 2019.
[29] J. He, L. Cai, and X. Guan, "Differential private noise adding mechanism and its application on consensus algorithm," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4069–4082, 2020.
[30] J. Cortés, G. E. Dullerud, S. Han, J. Le Ny, S. Mitra, and G. J. Pappas, "Differential privacy in control and network systems," in *IEEE 55th Conference on Decision and Control (CDC)*, 2016, pp. 4252–4272.
[31] Y. Xiong, J. Xu, K. You, J. Liu, and L. Wu, "Privacy preserving distributed online optimization over unbalanced digraphs via subgradient rescaling," *IEEE Transactions on Control of Network Systems*, 2020.
[32] T. Ding, S. Zhu, J. He, C. Chen, and X.-P. Guan, "Differentially private distributed optimization via state and direction perturbation in multi-agent systems," *IEEE Transactions on Automatic Control*, 2021.
[33] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private distributed convex optimization via functional perturbation," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 395–408, 2016.
[34] B. Polyak, "Introduction to optimization," *Optimization software Inc., Publications Division, New York*, vol. 1, 1987.
[35] Z. Huang, S. Mitra, and N. Vaidya, "Differentially private distributed optimization," in *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, New York, NY, USA, 2015.
[36] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, 2011.
[37] S. Kar and J. M. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 355–369, 2008.