Tailoring Gradient Methods for Differentially-Private Distributed Optimization

Yongqiang Wang, Angelia Nedić

***Abstract*—Decentralized optimization is gaining increased traction due to its widespread applications in large-scale machine learning and multi-agent systems. The same mechanism that enables its success, i.e., information sharing among participating agents, however, also leads to the disclosure of individual agents' private information, which is unacceptable when sensitive data are involved. As differential privacy is becoming a de facto standard for privacy preservation, recently results have emerged integrating differential privacy with distributed optimization. However, directly incorporating differential privacy design in existing distributed optimization approaches significantly compromises optimization accuracy. In this paper, we propose to redesign and tailor gradient methods for differentially-private distributed optimization, and propose two differential-privacy oriented gradient methods that can ensure both rigorous ϵ-differential privacy and optimality. The first algorithm is based on static-consensus based gradient methods, and the second algorithm is based on dynamic-consensus (gradient-tracking) based distributed optimization methods and, hence, is applicable to general directed interaction graph topologies. Both algorithms can simultaneously ensure almost sure convergence to an optimal solution and a finite privacy budget, even when the number of iterations goes to infinity. To our knowledge, this is the first time that both goals are achieved simultaneously. Numerical simulations using a distributed estimation problem and experimental results on a benchmark dataset confirm the effectiveness of the proposed approaches.**

I. Introduction

The problem of optimizing a global objective function through the cooperation of multiple agents has gained increased attention in recent years. This is driven by its wide applicability to many engineering and scientific domains, ranging from cooperative control [1], distributed sensing [2], sensor networks [3], to large-scale machine learning [4]. In many of these applications, 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.

The work was supported in part by the National Science Foundation under Grants ECCS-1912702, CCF-2106293, CCF-2106336, CCF-2215088, and CNS-2219487.

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

Plenty of approaches have been reported to solve the above distributed optimization problem since the seminal work of [5], with some of the commonly used approaches including gradient methods (e.g., [6], [7], [8], [9], [10], [11]), distributed alternating direction method of multipliers (e.g., [12]), and distributed Newton methods (e.g., [13]). 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. In general, existing gradient based distributed optimization algorithms can be divided into two categories. The first category combines gradient-descent operations and average-consensus mechanisms (referred to as static-consensus hereafter) by directly concatenating gradient-descent with a consensus operation of individual agents' optimization variables. Typical examples include [6], [14]. Such approaches are simple and efficient in computation since they only require an agent to share one variable in each iteration. However, these approaches are only applicable in balanced graphs (the sum of each agent's in-neighbor coupling weights equal to the sum of its out-neighbor coupling weights). The second category circumvents the balanced-graph restriction by exploiting consensus mechanisms able to track time-varying signals (so-called dynamic consensus, applicable to general directed graphs) to track the global gradient (see, e.g., [9], [10], [11], [15], [16]). It can ensure convergence to an optimal solution under constant stepsizes and, hence, can achieve faster convergence. However, such approaches need every agent to maintain and share an additional gradient-tracking variable besides the optimization variable, which doubles the communication overhead.

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 [12]. In sensor network based localization, the positions of sensor agents should be kept private in sensitive (hostile) environments as well [12], [17]. In fact, without an effective privacy mechanism in place, the results in [12], [17], [18] 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 approaches. 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 [19]. In fact, recent results in [20] 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, recently plenty of efforts have been reported to counteract potential privacy breaches in distributed optimization. One approach resorts to partially homomorphic encryption, which has been employed in both our own prior results [12], [21] and those of others [22]. 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. For example, [19], [23] showed that privacy can be enabled by adding a *constant* uncertain parameter in the projection step or stepsizes. The authors of [24] showed that network structure can be leveraged to construct spatially correlated "structured" noise to cover information. However, since the uncertainties injected by these approaches are correlated, their enabled privacy is restricted: projection based privacy depends on the size of the projection set – a large projection set nullifies privacy protection whereas a small projection set offers strong privacy protection but requires *a priori* knowledge of the optimal solution; "structured" noise based approaches require each agent to have a certain number of neighbors that do not share information with the adversary [25]. Privacy approaches have also been proposed for distributed stochastic optimization using time-varying heterogeneous stepsizes [26] or stochastic quantization effects [27] in our prior work. Differential Privacy (DP) [28] is becoming increasingly popular in privacy protection. It employs uncorrelated noises, and hence can provide strong privacy protection for a participating agent, even when all its neighbors are compromised. As DP is achieving remarkable successes in various applications [29], [30], [31], [32], [33] and becoming a de facto standard for privacy protection, 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 [17], [34], [35], [36], [37], or objective functions [38]. 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 privacy and accuracy. In fact, recently the investigation in [20] indicates that DP-based defense can achieve reasonable privacy protection *"only when the noise variance is large enough to degrade accuracy [20]."*

In this paper, we propose to tailor gradient methods for differentially-private distributed optimization. More specifically, motivated by the observation that persistent DP-noise has to be repeatedly injected in every iteration of gradient based methods to ensure a strong privacy protection, which results in significant reduction in optimization accuracy, 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 sequences such that the consensus and convergence to an optimal solution are ensured even in the presence of persistent DP-noise.

The main contributions are as follows: 1) We propose two gradient-based methods for differentially private distributed optimization. The first one is based on static-consensus combined with a gradient method, which needs every agent to store and share one variable in each iteration. The second one is based on dynamic-consensus (gradient-tracking) combined with an approximate gradient method, which needs every agent to store and share two variables, but it is applicable to general directed graphs; 2) We rigorously prove that both algorithms can ensure almost sure convergence of all agents to the optimal solution even in the presence of persistent DP-noise, which, to our knowledge, has not been achieved before; 3) We prove that both algorithms can ensure rigorous ϵ-DP for participating agents' objective functions, even when all communications are observable to adversaries. More interestingly, both algorithms can ensure a finite privacy budget even when the number of iterations goes to infinity. To our knowledge, this is the first time that almost sure convergence to the optimal solution and rigorous ϵ-DP (with a guaranteed finite privacy budget even when the number of iterations tends to infinity) are achieved simultaneously in distributed optimization; 4) Even without taking privacy into consideration, the two proposed algorithms and theoretical derivations are of interest themselves. We propose a new vector-valued martingale convergence theorem (Lemma 5) as well as its adaptations to distributed optimization problems (Lemmas 6, 8, and 10), which enables us to analyze the consensus-error evolution and optimality-gap evolution under DP-noise simultaneously.

The organization of the paper is as follows. Sec. II gives the problem formulation and some results for a later use. Sec. III presents a static-consensus based gradient method for differentially-private distributed optimization and establishes the almost sure convergence of all agents' iterates to an optimal solution as well as ϵ-DP guarantees. Sec. IV presents a dynamic-consensus based gradient method for differentially-private distributed optimization and establishes the almost sure convergence to an optimal solution as well as ϵ-DP guarantees. Sec. V presents both numerical simulations and experimental results on a benchmark dataset MNIST. Finally, Sec. VI concludes the paper.

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. We also use other vector/matrix norms defined under a certain transformation determined by a matrix W, which will be represented as $\|\cdot\|_W$. 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 consider a network of m agents, interacting on a general directed graph. We describe a directed graph using an ordered pair $\mathcal{G} = ([m], \mathcal{E})$, where $[m] = \{1, 2, \ldots, m\}$ is the set of nodes (agents) and $\mathcal{E} \subseteq [m] \times [m]$ is the edge set of ordered node pairs describing the interaction among agents. For a nonnegative weighting matrix $W = \{w_{ij}\} \in \mathbb{R}^{m\times m}$, we define the induced directed graph as $\mathcal{G}_W = ([m], \mathcal{E}_W)$, where the directed edge (i, j) from agent j to agent i exists, i.e., $(i, j) \in \mathcal{E}_W$ if and only if $w_{ij} > 0$. For an agent $i \in [m]$, its in-neighbor set $\mathbb{N}_i^{\text{in}}$ is defined as the collection of agents j such that $w_{ij} > 0$; similarly, the out-neighbor set $\mathbb{N}_i^{\text{out}}$ of agent i is the collection of agents j such that $w_{ji} > 0$.

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 1. *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 $L > 0$,*
$\|\nabla f_i(u) - \nabla f_i(v)\| \le L\|u - v\|, \quad \forall i \in [m] \text{ and } \forall u, v \in \mathbb{R}^d$

Under Assumption 1, 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 ([39], 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*

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

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

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

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

Proof. From the given relation we have *a.s.*

$$\mathbb{E}\left[v^{k+1}|\mathcal{F}^k\right] \le (1+\alpha^k)v^k - q^k v^k + p^k, \quad \forall k \ge 0 \quad (3)$$

By Lemma 1 with $u^k = q^k v^k$, and $\beta^k = p^k$, it follows that $\sum_{k=0}^{\infty} q^k v^k < \infty$ and $\lim_{k\to\infty} v^k = v$ for a random variable $v \ge 0$ *a.s.* Since $\sum_{k=0}^{\infty} q^k = \infty$, it follows that $\liminf_{k\to\infty} v^k = 0$ *a.s.* This and the fact $v^k \to v$ *a.s.* imply that $\lim_{k\to\infty} v^k = 0$ *a.s.* ∎

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 $z^* \in Z^*$,*

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

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. By letting $z = z^*$ for an arbitrary $z^* \in Z^*$ and defining $\phi^* = \min_{z\in\mathbb{R}^m}\phi(z)$, we obtain *a.s.* for all k:

$$\mathbb{E}\left[\|z^{k+1} - z^*\|^2|\mathcal{F}^k\right] \le (1+\alpha^k)\|z_k - z^*\| - \eta^k(\phi(z^k) - \phi^*) + \beta^k$$

Thus, all the conditions of Lemma 1 are satisfied, yielding

$$\left\{\|z^k - z^*\|\right\} \text{converges for each } z^* \in Z^* \quad a.s. \quad (4)$$

$$\sum_{k=0}^{\infty}\eta^k(\phi(z^k) - \phi^*) < \infty \quad a.s. \quad (5)$$

From (5) and $\sum_{k=0}^{\infty}\eta^k = \infty$ we have $\liminf_{k\to\infty}\phi(z^k) = \phi^*$ *a.s.* Let $\{z^{k_\ell}\}$ be a subsequence such that almost surely

$$\lim_{\ell\to\infty}\phi(z^{k_\ell}) = \liminf_{k\to\infty}\phi(z^k) = \phi^* \quad (6)$$

Relation (4) implies that the sequence $\{z^k\}$ is bounded *a.s.* Thus, we can assume without loss of generality that $\{z^{k_\ell}\}$ converges *a.s.* to some $\tilde{z}$ (for otherwise, we can in turn select a convergent subsequence of $\{z^{k_\ell}\}$). Therefore, by the continuity of ϕ, one has $\lim_{\ell\to\infty}\phi(z^{k_\ell}) = \phi(\tilde{z})$ *a.s.*, which in combination with (6) implies that $\tilde{z} \in Z^*$ *a.s.* By letting $z^* = \tilde{z}$ in (4), we see that z^k converges to $\tilde{z}$ *a.s.* ∎

Lemma 4. *Let $\{v^k\}$ be a nonnegative sequence, and $\{\alpha^k\}$ and $\{\beta^k\}$ be positive 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 $K \ge 0$ such that $v^{k+1} \le (1-\alpha^k)v^k + \beta^k$ holds for all $k \ge K$, then we always have $v^k \le C\frac{\beta^k}{\alpha^k}$ for all k, where C is some constant.*

Proof. The derivation follows the same line of reasoning in Lemma 4 of [40] and is omitted here. ∎

B. On differential privacy

We consider Laplace noise for DP. For a constant $\nu > 0$, $\text{Lap}(\nu)$ 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 [41], 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 $j \in [m]$, $j \neq i$;*
- *the different objective functions f_i and f'_i have similar behaviors around θ^*, the solution of $\mathcal{P}$. More specifically, there exits some $\delta > 0$ such that for all v and v' in $B_\delta(\theta^*) \triangleq \{u : u \in \mathbb{R}^d, \|u-\theta^*\| < \delta\}$, we have $\nabla f_i(v) = \nabla f'_i(v')$.*

Remark 1. *In Definition 1, since the change of an objective function from f_i to f'_i in the second condition can be arbitrary, additional restrictions have to imposed to ensure rigorous DP in distributed optimization. Different from [41] which restricts all gradients to be uniformly bounded, we add the third condition, which, as shown later, allows us to ensure rigorous DP while maintaining provable convergence to an optimal solution.*

Given a distributed optimization problem $\mathcal{P}$, we represent an iterative distributed optimization algorithm as a mapping $\mathcal{R}_\mathcal{P}(x^0) : x^0 \mapsto \mathcal{O}$, where x^0 is the initial state and $\mathcal{O}$ is the observation sequence (the sequence of all shared messages). Under a fixed distributed optimization algorithm, for a given distributed optimization problem $\mathcal{P}$, observation sequence $\mathcal{O}$, and initial state x^0, we denote the corresponding internal state at iteration k as $\mathcal{A}_{\mathcal{P},\mathcal{O},x^0}[k]$.

Definition 2. *(ϵ-DP [41]). 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*

$$\mathbb{P}[\mathcal{R}_\mathcal{P}(x^0) \in \mathcal{O}_s] \leq e^\epsilon \mathbb{P}[\mathcal{R}_{\mathcal{P}'}(x^0) \in \mathcal{O}_s] \tag{7}$$

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. Static-consensus gradient methods for differentially-private distributed optimization

In this section, we tailor a static-consensus based distributed gradient method to construct a differentially-private distributed method with almost sure convergence to an optimal solution. The agent interaction strength is captured by a weight matrix $W = \{w_{ij}\}$, where $w_{ij} > 0$ if there is a link from agent j to agent i, and $w_{ij} = 0$ otherwise. We let $w_{ii} \triangleq -\sum_{j\in\mathbb{N}_i^{\rm in}} w_{ij}$ for all $i \in [m]$, where $\mathbb{N}_i^{\rm in}$ is the in-neighbor set of agent i. We make the following assumption on W:

Assumption 2. *The matrix $W = \{w_{ij}\} \in \mathbb{R}^{m\times m}$ is symmetric and satisfies $\mathbf{1}^T W = \mathbf{0}^T$, $W\mathbf{1} = \mathbf{0}$, $\|I + W - \frac{\mathbf{1}\mathbf{1}^T}{m}\| < 1$.*

Assumption 2 ensures that the interaction graph induced by W is balanced and connected, i.e., there is a path from each agent to every other agent.

To achieve a strong DP, independent DP-noise should be injected repeatedly 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: DP-oriented static-consensus based 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 $k = 0, 1, 2, \ldots$ **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 $i \in \mathbb{N}_j^{\rm out}$.

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

$$x_i^{k+1} = x_i^k + \sum_{j\in\mathbb{N}_i^{\rm in}} \gamma^k w_{ij}(x_j^k + \zeta_j^k - x_i^k) - \lambda^k \nabla f_i(x_i^k) \tag{8}$$

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, to be specified later in Assumption 3.

A. Convergence analysis

We have to 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, we have 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. By multiplying the given relation for $\mathbf{v}^{k+1}$ with π and using $\pi^T V^k \leq \pi^T$ and the nonnegativity of $\mathbf{v}^k$, we obtain

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

Since $\pi > 0$, we have $\pi_{\min} = \min_i\{\pi_i\} > 0$, which yields $\mathbf{1}^T\mathbf{v}^k = \frac{1}{\pi_{\min}}\pi_{\min}\mathbf{1}^T\mathbf{v}^k \leq \frac{1}{\pi_{\min}}\pi^T\mathbf{v}^k$, where the inequality holds since $\mathbf{v}^k \geq 0$. So, one obtains

$$\mathbb{E}\left[\pi^T\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq \left(1 + a^k\frac{\pi^T\mathbf{1}}{\pi_{\min}}\right)\pi^T\mathbf{v}^k + b^k\pi^T\mathbf{1} - \pi^T H^k\mathbf{u}^k$$

By our assumption, $\pi^T H^k\mathbf{u}^k \geq 0$ holds for all k *a.s.* Thus, the preceding relation implies that the conditions of Lemma 1 are satisfied with $v^k = \pi^T\mathbf{v}^k$, $\alpha^k = a^k\pi^T\mathbf{1}/\pi_{\min}$ and $\beta^k = b^k\pi^T\mathbf{1}$. So by Lemma 1, it follows that $\lim_{k\to\infty}\pi^T\mathbf{v}^k$ exists *a.s.* Consequently, $\{\pi^T\mathbf{v}^k\}$ is bounded *a.s.*, and since $\{\mathbf{v}^k\}$ is nonnegative and $\pi > 0$, it follows that $\{\mathbf{v}^k\}$ is also bounded *a.s.* By Lemma 1, we have $\sum_{k=0}^{\infty}\pi^T H^k\mathbf{u}^k < \infty$ *a.s.* ■

Based on Lemma 3 and Lemma 5, we can prove the following general convergence results for static-consensus based distributed algorithms for problem (1).

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^*,*

$$\begin{aligned}
&\begin{bmatrix} \mathbb{E}\left[\|\bar{x}^{k+1} - \theta^*\|^2|\mathcal{F}^k\right] \\ \mathbb{E}\left[\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2|\mathcal{F}^k\right] \end{bmatrix} \\
&\leq \left(\begin{bmatrix} 1 & \frac{\gamma^k}{m} \\ 0 & 1-\kappa\gamma^k \end{bmatrix} + a^k\mathbf{1}\mathbf{1}^T\right)\begin{bmatrix} \|\bar{x}^k - \theta^*\|^2 \\ \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \end{bmatrix} \\
&\quad + b^k\mathbf{1} - c^k\begin{bmatrix} F(\bar{x}^k) - F(\theta^*) \\ 0 \end{bmatrix}, \quad \forall k \geq 0
\end{aligned} \tag{9}$$

where $\bar{x}^k = \frac{1}{m}\sum_{i=1}^m x_i^k$, $\mathcal{F}^k = \{x_i^\ell,\, i \in [m],\, 0 \leq \ell \leq k\}$, 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 $k \geq 0$. 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 assuming that persistent DP-noise satisfies the following assumption.

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*

$$\sum_{k=0}^{\infty}(\gamma^k)^2\max_{i\in[m]}(\sigma_i^k)^2 < \infty \tag{10}$$

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 2. *Given that γ^k decreases with time, (10) 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 (10). Allowing $\{\sigma_i^k\}$ to increase with time is key to enabling the strong ϵ-DP in Theorem 2.*

Theorem 1. *Under Assumptions 1, 2, and 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 3. *Communication imperfections can be modeled as channel noises [7], [42], which can be regarded as the DP-noise here. Therefore, Algorithm 1 can also counteract such communication imperfections in distributed optimization.*

Remark 4. *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. From Lemma 4, the relationship in (23) in the appendix implies that $\sum_{i=1}^m\|x_i^k - \bar{x}^k\|^2$ decreases to zero no slower than $\mathcal{O}((\frac{\lambda^k}{\gamma^k})^2)$, and hence we have x_i^k converging to $\bar{x}^k$ no slower than $\mathcal{O}(\frac{\lambda^k}{\gamma^k})$ (note β^k is on the order of $\frac{(\lambda^k)^2}{\gamma^k}$ from the proof of Theorem 1). Moreover, when $F(\cdot)$ is strongly convex, (25) implies that $\bar{x}^k$ converges to θ^* no slower than $\mathcal{O}((\frac{\lambda^k}{\gamma^k})^{0.5})$ using Lemma 4. Therefore, the convergence of every x_i^k to θ^*, 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 θ^*, should be no slower than $\mathcal{O}((\frac{\lambda^k}{\gamma^k})^{0.5})$. Moreover, from the proof of the theorem, it can be seen that the decreasing speed of $\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2$ (where $\otimes$ is the Kronecker product) increases with an increase in $|\nu|$, which corresponds to the spectral radius of W. Therefore, the decreasing speed of $\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2$ to zero increases with an increase in the spectral radius of W defined in Assumption 2.*

B. Privacy analysis

Similar to [41], we define the sensitivity of an algorithm to problem (1) as follows:

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

$$\Delta^k \triangleq \sup_{\mathcal{O}\in\mathbb{O}}\left\{\sup_{x^k\in\mathcal{A}_{\mathcal{P},\mathcal{O},x^0}[k],\, x'^k\in\mathcal{R}_{\mathcal{P}',\mathcal{O},x^0}[k]}\|x^k - x'^k\|_1\right\} \tag{11}$$

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 $\sum_{k=1}^T\frac{\Delta^k}{\nu^k} \leq \epsilon$, then Algorithm 1 is ϵ-differentially private for iterations from $k = 0$ to $k = T$.*

Proof. The lemma can be obtained following the same line of reasoning of Lemma 2 in [41]. ■

Theorem 2. *Under Assumptions 1 and 2, 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) *For any finite number of iterations T, Algorithm 1 is ϵ-differentially private with the cumulative privacy budget bounded by $\epsilon \leq \sum_{k=1}^T\frac{C\varsigma^k}{\nu^k}$ where $\varsigma^k \triangleq \sum_{p=1}^{k-1}(\Pi_{q=p}^{k-1}(1-\bar{w}\gamma^q))\lambda^{p-1} + \lambda^{k-1}$, $\bar{w} \triangleq \min_i\{|w_{ii}|\}$, and $C \triangleq \max_{i\in[m],0\leq k\leq T-1}\{\|\nabla f_i(x_i^k) - \nabla f_i'(x_i'^k)\|_1\}$*

(note that C is always finite since the algorithm ensures convergence in both $\mathcal{P}$ and $\mathcal{P}'$);

2) *The cumulative privacy budget is finite for $T \to \infty$ when the sequence $\{\frac{\lambda^k}{\nu^k}\}$ is summable.*

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

To prove the statements on privacy, we first analyze the sensitivity of Algorithm 1. 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 depends on $\|x^k - x'^k\|_1$ 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.

Because the initial conditions, objective functions, and observations of $\mathcal{P}$ and $\mathcal{P}'$ are identical for $j \neq i$, we have $x_j^k = x_j'^k$ for all $j \neq i$ and k. Therefore, $\|x^k - x'^k\|_1$ is always equal to $\|x_i^k - x_i'^k\|_1$.

According to Algorithm 1, we can arrive at

$$x_i^{k+1} - x_i'^{k+1} = (1 + w_{ii}\gamma^k)(x_i^k - x_i'^k) - \lambda^k(g_i^k - g_i'^k),$$

where we have represented $\nabla f_i(x_i^k)$ and $\nabla f'_i(x_i'^k)$ as g_i^k and $g_i'^k$, respectively, for notational simplicity. Note that we have also used the definition $w_{ii} \triangleq -\sum_{j\in\mathbb{N}_i} w_{ij}$ and the fact that the observations $x_j^k + \zeta_j^k$ and $x_j'^k + \zeta_j'^k$ are the same.

Hence, the sensitivity Δ^k satisfies

$$\Delta^{k+1} \leq (1 - |w_{ii}|\gamma^k)\Delta^k + \lambda^k\|g_i^k - g_i'^k\|_1.$$

which, implies the first statement by iteration using Lemma 7.

For the infinity horizon result in the second statement, we exploit the fact that our algorithm ensures convergence in both $\mathcal{P}$ and $\mathcal{P}'$. This means that $\|g_i^k - g_i'^k\|_1 = 0$ will be satisfied when k is large enough using the third condition in Definition 1 (see Remark 4 for convergence rate analysis). Furthermore, the ensured convergence also means that $\|g_i^k - g_i'^k\|_1$ is always bounded. Hence, there always exists some constant C such that the sequence $\{\|g_i^k - g_i'^k\|_1\}$ is upper bounded by the sequence $\{C\gamma^k\}$.

Therefore, according to Lemma 4, there always exists a constant $\bar{C}$ such that $\Delta^k \leq \bar{C}\lambda^k$ holds. Using Lemma 7, we can easily obtain $\epsilon \leq \sum_{k=1}^{T} \frac{\bar{C}\lambda^k}{\nu^k}$. Hence, ϵ will be finite even when T tends to infinity if the sequence $\{\frac{\lambda^k}{\nu^k}\}$ is summable, i.e., $\sum_{k=0}^{\infty} \frac{\lambda^k}{\nu^k} < \infty$. ∎

Different from [41] 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 [41] 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 5. *In Theorem 2, to ensure that the privacy budget 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.*

IV. Gradient-tracking based methods for differentially private distributed optimization

In this section, we propose a DP-oriented gradient-tracking based distributed algorithm for general directed graphs and prove that it can ensure convergence to an optimal solution even under persistent DP-noise. In gradient-tracking based algorithms, every agent $i \in [m]$ maintains and updates two iterates, x_i^k and y_i^k, where y_i^k is an estimate of the "joint agent" descent direction. These two iterates are exchanged with local neighbors in two different communication networks, namely, $\mathcal{G}_R$ and $\mathcal{G}_C$, which are, respectively, induced by matrices $R \in \mathbb{R}^{m\times m}$ and $C \in \mathbb{R}^{m\times m}$; that is (i, j) is a directed link in the graph $\mathcal{G}_R$ if and only if $R_{ij} > 0$ and, similarly, (i, j) is a directed link in $\mathcal{G}_C$ if and only if $C_{ij} > 0$. We make the following assumption on R and C. Note that, $\mathcal{R}_{A^T}$ is identical to $\mathcal{R}_A$ with the directions of edges reversed.

Assumption 4. *The matrices $R, C \in \mathbb{R}^{m\times m}$ have nonnegative off-diagonal entries ($R_{ij} \geq 0$ and $C_{ij} \geq 0$ for all $i \neq j$). The induced graphs $\mathcal{G}_R$ and $\mathcal{G}_{C^T}$ satisfy*

1) *$\mathcal{G}_R$ and $\mathcal{R}_{C^T}$ each contain at least one spanning tree;*
2) *There exists at least one node that is a root of spanning trees for both $\mathcal{G}_R$ and $\mathcal{R}_{C^T}$.*

Remark 6. *The assumption on $\mathcal{G}_R$ and $\mathcal{R}_{C^T}$ is weaker than requiring that both induced graphs of R and C to be strongly connected, which is assumed in most of the existing works.*

Algorithm 2: DP-oriented gradient-tracking based distributed optimization

Parameters: Stepsizes λ^k, α^k and weakening factors γ_1^k, γ_2^k. Every agent i maintains two states x_i^k and y_i^k, which are initialized with a random point $x_i^0 \in \mathbb{R}^d$ and $y_i^0 = \nabla f_i(x_i^0)$.

for $k = 0, 1, 2, \cdots$ **do**

a) Every agent j injects zero-mean DP-noises ζ_j^k and ξ_j^k to its states x_j^k and y_j^k, respectively.

b) Agent i pushes $C_{li}(y_i^k + \xi_i^k)$ to each agent $l \in \mathbb{N}_{C,i}^{\text{out}}$, and it pulls $x_j^k + \zeta_j^k$ from each $j \in \mathbb{N}_{R,i}^{\text{in}}$, where the subscript R or C in neighbor sets indicates the neighbors with respect to the graphs induced by these matrices.

c) Agent i chooses $\gamma_1^k > 0$ and $\gamma_2^k > 0$ satisfying $1 + \gamma_1^k R_{ii} > 0$ and $1 + \gamma_2^k C_{ii} > 0$ with

$$R_{ii} = -\sum_{j\in\mathbb{N}_{R,i}^{\text{in}}} R_{ij}, \quad C_{ii} = -\sum_{j\in\mathbb{N}_{C,i}^{\text{out}}} C_{ji} \tag{12}$$

Then, agent i updates its states as follows:

$$\begin{aligned} x_i^{k+1} &= (1+\gamma_1^k R_{ii})x_i^k + \gamma_1^k \sum_{j\in\mathbb{N}_{R,i}^{\text{in}}} R_{ij}(x_j^k+\zeta_j^k) - \lambda^k y_i^k \\ y_i^{k+1} &= (1-\alpha^k+\gamma_2^k C_{ii})y_i^k + \gamma_2^k \sum_{j\in\mathbb{N}_{C,i}^{\text{in}}} C_{ij}(y_j^k+\xi_j^k) \\ &\quad + \nabla f_i(x_i^{k+1}) - (1-\alpha^k)\nabla f_i(x_i^k) \end{aligned} \tag{13}$$

d) **end**

Note that the definition of R_{ii} and C_{ii} in (12) ensures that $R=\{R_{ij}\}$ has zero row sums and $C=\{C_{ij}\}$ has zero column sums.

A. Convergence analysis

We will prove that, when the two sequences $\{\gamma_1^k\}$ and $\{\gamma_2^k\}$ are designed appropriately, all agents' x-iterates generated by Algorithm 2 converge to an optimal solution *a.s.*, as long as the injected noises ζ_j^k and ξ_j^k have zero-mean and $\gamma_1^k(\gamma_2^k)$ bounded variances, to be specified later in Assumption 5. To this end, we first extend Lemma 2 to vectors.

Lemma 8. *Let $\{\mathbf{v}^k\}\subset\mathbb{R}^d$ be a sequence of non-negative random vectors and $\{b^k\}$ be a sequence of nonnegative random scalars such that $\sum_{k=0}^\infty b^k<\infty$ a.s. and*

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right]\le V^k\mathbf{v}^k+b^k\mathbf{1},\quad \forall k\ge 0\quad a.s.$$

where $\{V^k\}$ is a sequence of non-negative matrices and $\mathcal{F}^k=\{\mathbf{v}^\ell,b^\ell;0\le\ell\le k\}$. Assume that there exist a vector $\pi>0$ and a deterministic scalar sequence $\{a^k\}$ satisfying $a^k\in(0,1)$, $\sum_{k=0}^\infty a^k=\infty$, and $\pi^TV^k\le(1-a^k)\pi^T$ for all $k\ge0$. Then, we have $\lim_{k\to\infty}\mathbf{v}^k=0$ a.s.

Proof. By multiplying the given relation for $\mathbf{v}^{k+1}$ with π and using $\pi^TV^k\le(1-a^k)\pi^T$, we obtain the following relation due to the nonnegativity of the vectors $\mathbf{v}^k$:

$$\mathbb{E}\left[\pi^T\mathbf{v}^{k+1}|\mathcal{F}^k\right]\le(1-a^k)\pi^T\mathbf{v}^k+b^k\pi^T\mathbf{1},\quad\forall k\ge0\ a.s.$$

Since $\sum_{k=0}^\infty a^k=\infty$, and $\sum_{k=0}^\infty b^k<\infty$ *a.s.*, the conditions of Lemma 2 are satisfied with $v^k=\pi^T\mathbf{v}^k$, $\alpha^k=0$, $q^k=a^k$, and $p^k=b^k\pi^T\mathbf{1}$, implying *a.s.* $\lim_{k\to\infty}\pi^T\mathbf{v}^k=0$. $\{\mathbf{v}^k\}$ being nonnegative and $\pi>0$ imply $\lim_{k\to\infty}\mathbf{v}^k=0$ *a.s.* ∎

We now proceed to analyze the convergence of Algorithm 2. Defining $(\zeta_w^k)^T=\left[(\zeta_{w1}^k)^T,\cdots,(\zeta_{wm}^k)^T\right]$ with $\zeta_{wi}\triangleq\sum_{j\in\mathbb{N}_{R,i}^{\text{in}}}R_{ij}\zeta_j^k$ and $(\xi_w^k)^T=\left[(\xi_{w1}^k)^T,\cdots,(\xi_{wm}^k)^T\right]$ with $\xi_{wi}\triangleq\sum_{j\in\mathbb{N}_{C,i}^{\text{in}}}C_{ij}\xi_j^k$, we write the dynamics of Algorithm 2 in the following more compact form:

$$\begin{aligned} x^{k+1} &= \left((I+\gamma_1^kR)\otimes I_d\right)x^k+\gamma_1^k\zeta_w^k-\lambda^ky^k \\ y^{k+1} &= \left(((1-\alpha^k)I+\gamma_2^kC)\otimes I_d\right)y^k+\gamma_2^k\xi_w^k+g^{k+1} \\ &\quad -(1-\alpha^k)g^k \end{aligned} \tag{14}$$

where we used $g^{k+1}=\nabla f(x^{k+1})$ for notational simplicity.

Lemma 9. *[43] (or Lemma 1 in [16]) Under Assumption 4, for every k, the matrix $I+\gamma_1^kR$ has a unique nonnegative left eigenvector u^T (associated with eigenvalue 1) satisfying $u^T\mathbf{1}=m$, and the matrix $(1-\alpha^k)I+\gamma_2^kC$ has a unique nonnegative right eigenvector v (associated with eigenvalue $1-\alpha^k$) satisfying $\mathbf{1}^Tv=m$.*

According to Lemma 3 in [16], we know that the spectral radius of $R^k\triangleq I+\gamma_1^kR-\frac{\mathbf{1}u^T}{m}$ is equal to $1-\gamma_1^k|\nu_R|<1$ where ν_R is an eigenvalue of R. Furthermore, there exists a vector norm $\|x\|_R\triangleq\|\tilde{R}x\|_2$ (where $\tilde{R}$ is determined by R [16]) such that $\|R^k\|_R<1$ is arbitrarily close to the spectral radius of R^k, i.e., $1-\gamma_1^k|\nu_R|<1$ (also see [44]). Without loss of generality, we represent this norm as $\|R^k\|_R=1-\gamma_1^k\rho_R<1$. Similarly, we have that the spectral radius of $C^k\triangleq(1-\alpha^k)I+\gamma_2^kC-\frac{v\mathbf{1}^T}{m}$ is equal to $1-\alpha^k-\gamma_2^k|\nu_C|<1$ where ν_C is an eigenvalue of C. Furthermore, there exists a vector norm $\|x\|_C\triangleq\|\tilde{C}x\|_2$ (where $\tilde{C}$ is determined by C [16]) such that $\|C^k\|_C<1$ is arbitrarily close to the spectral radius of C^k, i.e., $1-\alpha^k-\gamma_2^k|\nu_C|<1$. Without loss of generality, we bound this norm as $\|C^k\|_C\le1-\gamma_2^k\rho_C<1$.

Defining $\bar{x}^k=\frac{1}{m}\sum_{i=1}^m u_ix_i^k$ and $\bar{y}^k=\frac{1}{m}\sum_{i=1}^m y_i^k$, we have

$$\begin{aligned} \bar{x}^{k+1} &= \bar{x}^k+\gamma_1^k\bar{\zeta}_w^k-\lambda^k\frac{(u\otimes I_d)^T}{m}y^k \\ \bar{y}^{k+1} &= (1-\alpha^k)\bar{y}^k+\gamma_2^k\bar{\xi}_w^k+\bar{g}^{k+1}-(1-\alpha^k)\bar{g}^k \end{aligned} \tag{15}$$

with $\bar{\zeta}_w^k=\frac{1}{m}\sum_{i=1}^m u_i\zeta_{wi}^k$, $\bar{\xi}_w^k=\frac{1}{m}\sum_{i=1}^m\xi_{wi}^k$, and $\bar{g}^k=\frac{1}{m}\sum_{i=1}^m g_i^k$.

From (15), we can further obtain

$$\begin{aligned} \bar{x}^{k+1} &= \bar{x}^k+\gamma_1^k\bar{\zeta}_w^k-\lambda^k\frac{(u\otimes I_d)^T}{m}\left(y^k-(v\otimes I_d)\bar{y}^k\right) \\ &\quad -\lambda^k\frac{(u\otimes I_d)^T}{m}(v\otimes I_d)\bar{y}^k \end{aligned} \tag{16}$$

Using the relationship $\lambda^k\frac{(u\otimes I_d)^T}{m}(v\otimes I_d)\bar{y}^k=\lambda^k\frac{u^Tv}{m}\bar{y}^k$, we can rewrite (16) as follows

$$\bar{x}^{k+1}=\bar{x}^k-\lambda^k\frac{(u\otimes I_d)^T}{m}\left(y^k-(v\otimes I_d)\bar{y}^k\right)-\lambda^k\frac{u^Tv}{m}\bar{y}^k+\gamma_1^k\bar{\zeta}_w^k \tag{17}$$

In what follows, we use F^* to denote the optimal value of the problem in (1), i.e., $F^*=\min_{\theta\in\mathbb{R}^d}F(\theta)$.

Next, we provide a generic convergence result for dynamic-consensus (gradient-tracking) based distributed algorithms for problem (1). To this end, we need a measure under the $\|\cdot\|_R$ norm for the distance between all $x_1^k,x_2^k,\cdots,x_m^k$ and $\bar{x}^k$. Following [16], we define a matrix norm for all x iterates $\mathbf{x}^k\triangleq\left[x_1^k,x_2^k,\cdots,x_m^k\right]^T\in\mathbb{R}^{m\times d}$:

$$\|\mathbf{x}^k\|_R=\left\|\left[\|\mathbf{x}_{(1)}^k\|_R,\|\mathbf{x}_{(2)}^k\|_R,\cdots,\|\mathbf{x}_{(d)}^k\|_R\right]\right\|_2 \tag{18}$$

where the subscript 2 denotes the 2−norm and $\mathbf{x}_{(i)}^k$ denotes the ith column of $\mathbf{x}^k$. Defining $\bar{\mathbf{x}}^k$ as $\left[\bar{x}^k,\bar{x}^k,\cdots,\bar{x}^k\right]^T\in\mathbb{R}^{m\times d}$, one can easily see that $\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|_R$ measures the distance between all x_i^k and their average $\bar{x}^k$. Similarly, we define a matrix norm $\|\cdot\|_C$ for $\mathbf{y}^k\triangleq\left[y_1^k,y_2^k,\cdots,y_m^k\right]^T\in\mathbb{R}^{m\times d}$:

$$\|\mathbf{y}^k\|_C=\left\|\left[\|\mathbf{y}_{(1)}^k\|_C,\|\mathbf{y}_{(2)}^k\|_C,\cdots,\|\mathbf{y}_{(d)}^k\|_C\right]\right\|_2 \tag{19}$$

and use $\|\mathbf{y}^k-\text{diag}(v)\bar{\mathbf{y}}^k\|_C$ (with $\text{diag}(v)=\text{diag}(v_1,\ldots,v_m)$ and $\bar{\mathbf{y}}^k\triangleq\left[\bar{y}^k,\ldots,\bar{y}^k\right]^T\in\mathbb{R}^{m\times d}$)

to measure the distance between all y iterates and their v-weighted average $\bar{y}^k$.

Lemma 10. *Assume that the objective function $F(\cdot)$ is differentiable and that the problem (1) has an optimal solution. Suppose that a distributed algorithm generates sequences $\{x_i^k\} \subseteq \mathbb{R}^d$ and $\{y_i^k\} \subseteq \mathbb{R}^d$ under coupling matrices R and C, respectively, such that the following relation holds a.s. for some sufficiently large integer $T \geq 0$ and for all $k \geq T$:*

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq \left(V^k + a^k \mathbf{1}\mathbf{1}^T\right)\mathbf{v}^k + b^k\mathbf{1} - H^k\begin{bmatrix} \|\nabla F(\bar{x}^k)\|^2 \\ \|\bar{y}^k\|^2 \end{bmatrix} \tag{20}$$

where $\mathcal{F}^k = \{x_i^\ell, y_i^\ell; 0 \leq \ell \leq k,\ i \in [m]\}$ and

$$\mathbf{v}^k = \begin{bmatrix} \mathbf{v}_1^k \\ \mathbf{v}_2^k \\ \mathbf{v}_3^k \end{bmatrix} \triangleq \begin{bmatrix} F(\bar{x}^k) - F^* \\ \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_R^2 \\ \|\mathbf{y}^k - \mathrm{diag}(v)\bar{\mathbf{y}}^k\|_C^2 \end{bmatrix},$$

$$V^k = \begin{bmatrix} 1 & \kappa_1\lambda^k & \kappa_2\lambda^k \\ 0 & 1-\kappa_3\gamma_1^k & 0 \\ 0 & 0 & 1-\kappa_4\gamma_2^k \end{bmatrix}, \quad H^k = \begin{bmatrix} \kappa_5\lambda^k & \kappa_6\lambda^k - \kappa_7(\lambda^k)^2 \\ 0 & -\kappa_8\frac{(\lambda^k)^2}{\gamma_1^k} \\ 0 & -\kappa_9\frac{(\lambda^k)^2}{\gamma_2^k} \end{bmatrix}$$

with $\kappa_i > 0$ for all $1 \leq i \leq 9$ and $\kappa_3, \kappa_4 \in (0,1)$, while the nonnegative scalar sequences $\{a^k\}$, $\{b^k\}$ and positive sequences $\{\lambda^k\}$, $\{\gamma_1^k\}$, $\{\gamma_2^k\}$ satisfy $\sum_{k=0}^\infty a^k < \infty$ a.s., $\sum_{k=0}^\infty b^k < \infty$ a.s., $\sum_{k=0}^\infty \lambda^k = \infty$, $\sum_{k=0}^\infty \gamma_i^k = \infty$, $\sum_{k=0}^\infty (\gamma_i^k)^2 < \infty$, $\sum_{k=0}^\infty \frac{(\lambda^k)^2}{\gamma_i^k} < \infty$, $\lim_{k\to\infty} \lambda^k/\gamma_i^k = 0$ for $\imath \in \{1, 2\}$, and $\lim_{k\to\infty} \gamma_1^k/\gamma_2^k < \infty$. Then, we have:

(a) $\lim_{k\to\infty} F(\bar{x}^k)$ exists a.s. and

$$\lim_{k\to\infty} \|x_i^k - \bar{x}^k\| = \lim_{k\to\infty} \|y_i^k - v_i\bar{y}^k\| = 0,\ \forall i \quad a.s.$$

(b) $\liminf_{k\to\infty} \|\nabla F(\bar{x}^k)\| = 0$ holds a.s. Moreover, if the function $F(\cdot)$ has bounded level sets, then $\{\bar{x}^k\}$ is bounded and every accumulation point of $\{\bar{x}^k\}$ is an optimal solution a.s., and $\lim_{k\to\infty} F(x_i^k) = F^*$ a.s. for all $i \in [m]$.

Proof. See Appendix C. ∎

Remark 7. *In Lemma 10(b), the bounded level set condition can be replaced with any other condition ensuring that the sequence $\{\bar{x}^k\}$ is a.s. bounded.*

Lemma 10 is critical for establishing convergence properties of the gradient tracking-based distributed algorithm together with suitable conditions on the DP-noise injected by the agents. We make the following assumption on the noise:

Assumption 5. *For every $i \in [m]$, the noise sequences $\{\zeta_i^k\}$ and $\{\xi_i^k\}$ are zero-mean independent random variables, and independent of $\{x_i^0; i \in [m]\}$. Also, for every k, the noise collection $\{\zeta_j^k, \xi_j^k; j \in [m]\}$ is independent. The noise variances $(\sigma_{\zeta,i}^k)^2 = \mathbb{E}\left[\|\zeta_i^k\|^2\right]$ and $(\sigma_{\xi,i}^k)^2 = \mathbb{E}\left[\|\xi_i^k\|^2\right]$ and their attenuation stepsizes γ_1^k and γ_2^k are such that*

$$\sum_{k=0}^\infty (\gamma_1^k)^2 \max_{i\in[m]}(\sigma_{\zeta,i}^k)^2 < \infty, \sum_{k=0}^\infty (\gamma_2^k)^2 \max_{j\in[m]}(\sigma_{\xi,j}^k)^2 < \infty. \tag{21}$$

The initial random vectors satisfy $\mathbb{E}\left[\|x_i^0\|^2\right] < \infty$, $\forall i \in [m]$.

Remark 8. *Given that γ_1^k, γ_2^k, and λ^k decrease with time, (21) can be satisfied even when $\{\sigma_i^k\}$ increases with time. For example, under $\lambda^k = \mathcal{O}(\frac{1}{k})$, $\gamma_1^k = \mathcal{O}(\frac{1}{k^{0.9}})$, $\gamma_2^k = \mathcal{O}(\frac{1}{k^{0.7}})$, an increasing $\{\sigma_i^k\}$ with increasing rate no faster than $\mathcal{O}(k^{0.15})$ still satisfies the summable condition in (21).*

Assumption 6. *The gradients of all individual objective functions are bounded, i.e., there exists a constant C such that $\|\nabla f_i(x)\|_1 \leq C$ holds for all $x \in \mathbb{R}^p$ and $1 \leq i \leq m$.*

Theorem 3. *Let Assumptions 1, 4, 5, and 6 hold. If nonnegative sequences $\{\gamma_1^k\}$, $\{\gamma_2^k\}$, $\{\alpha^k\}$, and $\{\lambda^k\}$ satisfy $\sum_{k=0}^\infty \gamma_i^k = \infty$, $\sum_{k=0}^\infty (\gamma_i^k)^2 < \infty$, $\sum_{k=0}^\infty \alpha^k = \infty$, $\sum_{k=0}^\infty \lambda^k = \infty$, $\sum_{k=0}^\infty \frac{(\lambda^k)^2}{\gamma_i^k} < \infty$, $\lim_{k\to\infty} \lambda^k/\gamma_i^k = 0$ for $\imath \in \{1, 2\}$, $\lim_{k\to\infty} \lambda^k/\alpha^k < \infty$, $\sum_{k=0}^\infty \frac{(\alpha^k)^2}{\gamma_2^k} < \infty$ and $\sum_{k=0}^\infty \frac{(\gamma_1^k)^2}{\gamma_2^k} < \infty$, then, the results of Lemma 10 hold for Algorithm 2.*

Proof. The goal is to establish the relationship in (20), with the σ-field $\mathcal{F}^k = \{x_i^\ell, y_i^\ell; 0 \leq \ell \leq k,\ i \in [m]\}$. Since the derivation is lengthy, we omit it here due to space limitations. The detailed proof can be found in [45]. ∎

Remark 9. *In networked systems, usually communication imperfections can be modeled as channel noises [42], which can be regarded as a special case of the DP noise considered here. Therefore, Algorithm 2 can also be used to counteract such communication imperfections in distributed optimization.*

Remark 10. *Because the evolution of x_i^k to the optimal solution satisfies the conditions in Lemma 10, which are in turn derived based on Lemma 2, we can leverage Lemma 10 and Lemma 2 to characterize the convergence speed. More specifically, in the proof of Lemma 10 in the appendix, (44) and the relationship $\tilde{\pi}^T\tilde{V}^k = (1 - \alpha\gamma_1^k)\tilde{\pi}^T$ imply that $\mathbf{v}_2^k \triangleq \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_R^2$ and $\mathbf{v}_3^k \triangleq \|\mathbf{y}^k - \mathrm{diag}(v)\bar{\mathbf{y}}^k\|_C^2$ decay to zero with a rate no slower than $\mathcal{O}(\frac{\gamma_1^k}{\gamma_2^k})$. Furthermore, (42) implies that $\lambda^k\|\nabla F(\bar{x}^k)\|^2$ decays to zero with a rate no slower than $\mathcal{O}(\frac{1}{k})$, i.e., $\|\nabla F(\bar{x}^k)\|^2$ decays to zero with a rate no slower than $\mathcal{O}(\frac{1}{k\lambda^k})$. Moreover, from the proof in Lemma 10 (specifically (44) and the paragraph below it), we know that the decreasing speed of $\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_R^2$, $\|\mathbf{y}^k - \mathrm{diag}(v)\bar{\mathbf{y}}^k\|_C^2$ increases with an increase in α, which in turn increases with an increase in κ_3 and κ_4. Further noting that κ_3 and κ_4 correspond to the spectral radius of R and C, respectively, we have that the convergence speed increases with an increase in the spectral radius of R and C defined in Assumption 4 (see (12) for diagonal entries).*

B. Privacy analysis

Theorem 4. *Under Assumptions 1, 4 and 6, if $F(\cdot)$ has bounded level sets, nonnegative sequences $\{\lambda^k\}$, $\{\alpha^k\}$, $\{\gamma_1^k\}$, and $\{\gamma_2^k\}$ satisfy the conditions in Theorem 3, and all elements of ζ_i^k and ξ_i^k are drawn independently from Laplace distribution $\mathrm{Lap}(\nu^k)$ with $(\sigma_{\zeta,i}^k)^2 = (\sigma_{\xi,i}^k)^2 = 2(\nu^k)^2$ satisfying Assumption 5, then all agents will converge a.s. to an optimal solution. Moreover,*

1) *For any finite number of iterations T, Algorithm 1 is ϵ-differentially private with the cumulative privacy budget bounded by $\epsilon \leq \sum_{k=1}^{T} \frac{2C(\varsigma_x^k+\varsigma_y^k)}{\nu^k}$ where $\varsigma_x^k \triangleq \sum_{p=1}^{k-1}(\Pi_{q=p}^{k-1}(1-\bar{R}\gamma_1^q))\lambda^{p-1}\varsigma_y^{p-1} + \lambda^{k-1}\varsigma_y^{k-1}$, $\varsigma_y^k \triangleq \sum_{p=1}^{k-1}(\Pi_{q=p}^{k-1}(1-\alpha^q-\bar{C}\gamma_2^q))(2-\alpha^{p-1}) + 2 - \alpha^{k-1}$, $\bar{R} \triangleq \min_i\{|R_{ii}|\}$, $\bar{C} \triangleq \min_i\{|C_{ii}|\}$, and C is from Assumption 6;*
2) *The cumulative privacy budget is finite for $T \to \infty$ when the sequence $\{\frac{\lambda^k}{\nu^k}\}$ is summable.*

Proof. Since the convergence follows Theorem 3, we only consider the privacy statements.

Following Definition 3 and the argument in the proof of Theorem 2, we know that the sensitivity of Algorithm 2 is determined by $x_i^k - x_i'^k$ and $y_i^k - y_i'^k$, which, according to Algorithm 2, have the following dynamics:

$$\begin{aligned} x_i^{k+1} - x_i'^{k+1} =&(1-\gamma_1^k|R_{ii}|)(x_i^k - x_i'^k) - \lambda^k(y_i^k - y_i'^k),\\ y_i^{k+1} - y_i'^{k+1} =&(1-\alpha^k-\gamma_2^k|C_{ii}|)(y_i^k - y_i'^k)\\ &+ (g_i^{k+1} - g_i'^{k+1}) - (1-\alpha^k)(g_i^k - g_i'^k), \end{aligned}$$

where we have represented $\nabla f_i(x_i^k)$ and $\nabla f_i'(x_i'^k)$ as g_i^k and $g_i'^k$, respectively, for notational simplicity. Note that we have also used the fact that the observations $x_j^k + \zeta_j^k$ (resp. $y_j^k + \xi_j^k$) and $x_j'^k + \zeta_j'^k$ (resp. $y_j'^k + \xi_j'^k$) are the same.

Hence, $y_i^k - y_i'^k$ satisfies

$$\begin{aligned} &\|y_i^{k+1} - y_i'^{k+1}\|_1\\ &\leq (1-\alpha^k-\gamma_2^k|C_{ii}|)\|y_i^k - y_i'^k\|_1\\ &\quad + \|g_i^{k+1} - g_i'^{k+1}\|_1 + (1-\alpha^k)\|g_i^k - g_i'^k\|_1\\ &\leq (1-\alpha^k-\gamma_2^k|C_{ii}|)\|y_i^k - y_i'^k\|_1 + (2-\alpha^k)2C, \end{aligned}$$

where C is from Assumption 6.

By iteration, we have that $\|y_i^k - y_i'^k\|_1$ is always bounded by $2C\varsigma_y^k$ in the first privacy statement.

One can also see that $x_i^k - x_i'^k$ satisfies

$$\begin{aligned} \|x_i^{k+1} - x_i'^{k+1}\|_1 \leq &(1-\gamma_1^k|R_{ii}|)\|x_i^k - x_i'^k\|_1\\ &+ \lambda^k\|y_i^k - y_i'^k\|_1. \end{aligned}$$

Hence, using iteration, we obtain that $\|x_i^k - x_i'^k\|_1$ is always bounded by $2C\varsigma_x^k$ in the first privacy statement.

Therefore, the sensitivity at iteration k is no larger than $2C(\varsigma_x^k+\varsigma_y^k)$ in the first privacy statement, and, hence, we have the first privacy statement on the cumulative privacy budget for a finite number of iterations.

On the infinite time horizon, we follow the argument in the proof of Theorem 2. More specifically, we can prove that the sequence $\{\|g_i^k - g_i'^k\|_1\}$ can be bounded by the sequence $\{C_g\gamma_2^k\lambda^k\}$ (with C_g some constant) using the third condition in Definition 1 and the guaranteed convergence. Hence, according to Lemma 4, there always exists a constant $\bar{C}_y$ such that $\|y_i^k - y_i'^k\|_1 \leq \bar{C}_y\lambda^k$ holds. Still using Lemma 4, we can prove that there always exists a constant $\bar{C}_x'$ such that $\|x_i^k - x_i'^k\|_1 \leq \bar{C}_x'\frac{(\lambda^k)^2}{\gamma_1^k}$ holds. Given that λ^k decreases faster than γ_1^k, we have $\|x_i^k - x_i'^k\|_1 \leq \bar{C}_x\lambda^k$ for some constant $\bar{C}_x$. Therefore, on the infinite time horizon, the sensitivity is on the order of λ^k. Hence, we have the result on the cumulative privacy budget when $T \to \infty$ in the second statement. ■

Remark 11. *Since we use the standard ϵ-DP framework, we characterize the cumulative privacy budget directly. Under relaxed (approximate) ϵ-DP frameworks, such as (ϵ, δ)-DP [46], zero-concentrated DP [47], or Rényi DP [48], advanced composition theories in [46], [47], [48] can be exploited to characterize the cumulative privacy budget.*

V. Numerical Experiments

A. Evaluation using distributed estimation

We first evaluate the performance of the two proposed algorithms 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 [26].

We consider a network of $m = 5$ sensors interacting on the graph depicted in Fig. 1. In the evaluation, we set $s = 3$ and $d = 2$. To evaluate the performance of the proposed Algorithm 1, we ignored the directions of edges in Fig. 1 in the selection of coupling weights and 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. 2. For comparison, we also ran the existing static-consensus based distributed gradient descent (DGD) approach in [6] under the same noise, and the differential-privacy approach for distributed optimization (PDOP) in [17] 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. 2, respectively. It is clear that the proposed algorithm has a comparable convergence speed but much better optimization accuracy.

We also evaluated Algorithm 2 which is applicable to general directed graphs. More specifically, still using the topology in Fig. 1, we selected R and C matrices according to Assumption 4 and set the stepsize and diminishing sequences as $\lambda^k = \frac{0.02}{1+0.1k}$, $\alpha^k = \frac{0.02}{1+0.1k}$ $\gamma_1^k = \frac{1}{1+0.1k^{0.9}}$, and $\gamma_2^k = \frac{1}{1+0.1k^{0.7}}$, respectively. We injected Laplace noises ζ_i^k and ξ_i^k (both have parameter $\nu^k = 1 + 0.1k^{0.1}$) on all shared x_i^k and y_i^k respectively to enable DP, and it can be verified that the parameters satisfy the conditions in Theorem

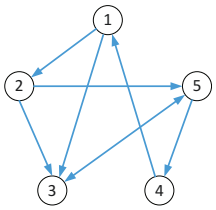

Fig. 1. The interaction topology of the network

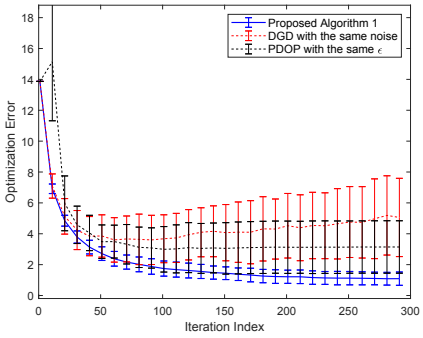

Fig. 2. Comparison of Algorithm 1 with existing distributed gradient descent algorithm (DGD) in [6] (under the same noise) and the differential-privacy approach for decentralized optimization PDOP in [17] (under the same privacy budget) using the distributed estimation problem

3 and Theorem 4. 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. 3. For comparison, we also ran the conventional dynamic-consensus based Push-Pull method in [16] under the same noise and the PDOP based differential-privacy approach for distributed optimization. Because the PDOP based approach requires the stepsize to decay with a geometric rate, we set the stespize of Push-Pull to 0.95^k and used a geometrically decaying noise such that it has the same privacy budget as our approach. The evolution of the average optimization error and variance of Push Pull (with the same noise as our approach) and PDOP-privacy based Push Pull (with the same privacy budget as our approach) are depicted by the red and black curves/error bars in Fig. 3, respectively. It is clear that the proposed algorithm has a comparable convergence speed but gained significant

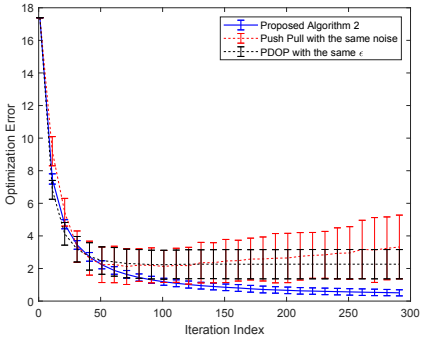

Fig. 3. Comparison of Algorithm 2 with existing dynamic-consensus based distributed gradient algorithm (Push Pull) in [16] (under the same noise) and the PDOP-based differential-privacy approach in [17] for Push Pull (under the same privacy budget) using the distributed estimation problem

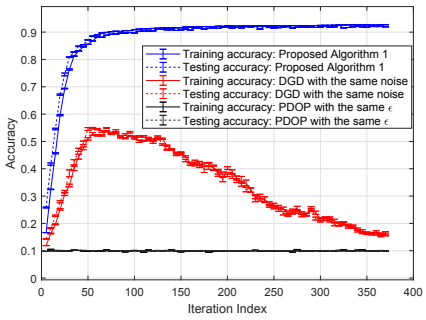

Fig. 4. Comparison of Algorithm 1 with existing distributed gradient descent algorithm (DGD) in [6] (under the same noise) and the differential-privacy approach for decentralized optimization PDOP in [17] (under the same privacy budget) using the MNIST image classification problem

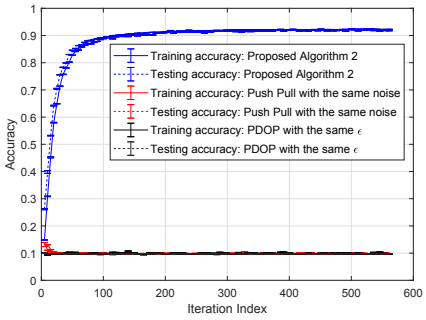

Fig. 5. Comparison of Algorithm 2 with existing dynamic-consensus based distributed gradient algorithm (Push Pull) in [16] (under the same noise) and the PDOP-based differential-privacy approach in [17] for Push Pull (under the same privacy budget) using the MNIST image classification problem

improvement in optimization accuracy.

B. Evaluation using image classification on MNIST

We also used decentralized training of a convolutional neural network (CNN) to evaluate the performance of our proposed algorithms. More specially, we consider five agents which collaboratively train a CNN using the MNIST dataset [49] under the topology in Fig. 1. The MNIST data set is a large benchmark database of handwritten digits widely used for training and testing in the field of machine learning [50]. Each agent has a local copy of the CNN. The CNN has 2 convolutional layers with 32 filters with each followed by a max pooling layer, and then two more convolutional layers with 64 filters each followed by another max pooling layer and a dense layer with 512 units. Each agent has access to a portion of the MNIST dataset, which was further divided into two subsets for training and validation, respectively. To evaluate the proposed Algorithm 1, We set the stepsize as $\lambda^k = \frac{1}{1+0.01k}$ and the weakening factor γ^k as $\frac{1}{1+0.01k^{0.9}}$. The Laplace noise parameter was set to $\nu^k = 1 + 0.01k^{0.3}$ to enable ϵ-DP. The evolution of the training and testing accuracies averaged over 50 runs are illustrated by the solid and dashed blue curves in Fig. 4. To compare the convergence performance of our algorithm with the conventional distributed gradient descent algorithm under DP-noise, we also implemented the distributed gradient descent (DGD) algorithm in [6] to train the same CNN using stepsize $\frac{1}{1+0.01k}$ under the same Laplace noise. The results are illustrated by the solid and dotted red curves in Fig. 4. It can be seen that the proposed algorithm has much

better robustness to DP-noise. Moreover, to compare with the existing DP approach for distributed optimization, we also implemented the DP approach PDOP in [41] on DGD under the same privacy budget ϵ. PDOP uses geometrically decaying stepsizes and noises to ensure a finite privacy budget. However, such fast-decaying stepizes turned out to be unable to train the complex CNN model (see training and testing accuracies in solid and dashed black curves in Fig. 4, respectively under $\lambda^k = 0.95^k$ and $\nu^k = 0.98^k$). These comparisons corroborate the advantage of the proposed Algorithm 1.

To show the influence of DP-noise on the final optimization accuracy, we also scaled the noise by 0.5 and 2 respectively and obtained the training and testing accuracies. To compare the strength of enabled privacy protection, we ran the DLG attack model proposed in [20], which is the most powerful inference algorithm reported to date in terms of reconstructing exact raw data from shared gradient/model updates. The attacker was assumed to be able to observe all messages shared among the agents. The training/testing accuracies under different levels of DP-noise and DLG attacker's inference errors are summarized in Table 1. It can be seen that there is a trade-off between privacy and accuracy under a fixed iteration number $20,000$.

Using the same interaction topology, CNN network, and MNIST dataset, we also evaluated the performance of the proposed Algorithm 2 under DP-noise. The parameters of Algorithm 2 were set as $\lambda^k = \frac{1}{1+0.01k}$, $\alpha^k = \frac{0.1}{1+0.01k}$, $\gamma_1^k = \frac{1}{1+0.01k^{0.9}}$, and $\gamma_2^k = \frac{1}{1+0.01k^{0.7}}$. The Laplace noise parameter was set as $\nu^k = \frac{1}{1+0.01k^{0.1}}$. The evolution of the training and testing accuracies averaged over 50 runs are illustrated by the solid and dashed blue curves in Fig. 5. For comparison, we also implemented the dynamic-consensus based Push Pull algorithm in [16] to train the same CNN using stepsize 0.02 under the same Laplace noise. The results are illustrated by the solid and dotted red curves in Fig. 5. It can be seen that the same amount of noise, which is tolerable to our proposed Algorithm 2, completely prevents the Push Pull algorithm from training the CNN model. Moreover, we also applied PDOP based DP approach in [41] to Push Pull, which uses geometrically decaying stepsizes and noises to ensure a finite privacy budget. However, under the same privacy budget, the fast-decaying stepize for Push Pull turned out to be unable to train the complex CNN model either (see Fig. 5 for training and testing accuracies in solid and dashed black curves, respectively, under $\lambda^k = 0.95^k$ and $\nu^k = 0.98^k$). These comparisons corroborate the advantage of the proposed Algorithm 2.

To show the influence of DP-noise on the final optimization accuracy and the strength of enabled privacy, we also scaled the noise by 0.5 and 2 respectively and obtained the training/testing accuracies as well as DLG attacker's inference errors. The results are given in Table 1, which shows a trade-off between privacy and accuracy under a fixed iteration number $20,000$.

VI. Conclusions, Discussions, and Future Work

Although DP is becoming the de facto standard for publicly sharing information, its direct incorporation into distributed optimization leads to significant reduction in optimization accuracy due to the need to iteratively and repeatedly inject independent noises. This paper proposes two DP-oriented gradient based distributed optimization algorithms that ensure both ϵ-DP and optimization accuracy. Specifically, the two algorithms can ensure almost sure convergence of all agents to the optimal solution even in the presence of persistent DP noise. Both algorithms are also proven able to ensure ϵ-DP with a finite cumulative privacy budget, even when the number of iterations goes to infinity. The simultaneous achievement of both provable convergence to the accurate solution and rigorous ϵ-DP with guaranteed finite cumulative privacy budget, to our knowledge, has not been reported before in distributed optimization. Numerical simulations and experimental results using a benchmark dateset confirm that both algorithms have a better accuracy compared with their respective existing counterparts, while maintaining a comparable convergence speed.

It is worth noting that our algorithms' simultaneous achievement of both provable convergence to the optimal solution and ϵ-DP does not contradict the fundamental theory and limitations of DP in [28]. Firstly, our convergence guarantee (almost sure convergence) is obtained in the stochastic sense, which is different from deterministic convergence under no DP noise. More specifically, when the number of implementations tends to infinity, the concept of almost sure convergence still allows for a finite number of implementations that do not converge to the optimal solution. Secondly, according to the DP theory, conventional query mechanisms on a dataset can achieve ϵ-DP only by sacrificing query accuracies, but 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 [28], 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 examples, as we can perform non-private machine learning." Actually, under Valiant's model of machine learning (PAC), [28] notes that a model parameter (called function in [28]) is PAC learnable if and only if it is PAC learnable under DP (see page 221 of [28]). Thirdly, the achievement of ϵ-DP does incur utility cost. More specifically, in terms of Algorithm 1, 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 the convergence speed is determined by $\mathcal{O}((\frac{\lambda^k}{\gamma^k})^{0.5})$ according to Remark 3, we arrive at the conclusion that a faster decreasing

TABLE I
TRAINING/TESTING ACCURACIES (AFTER $20,000$ ITERATIONS) AND DLG ATTACKER'S INFERENCE ERRORS UNDER DIFFERENT LEVELS OF DP-NOISE

	Algorithm 1			Algorithm 2		
Noise Level[a]	$\times 0.5$	$\times 1$	$\times 2$	$\times 0.5$	$\times 1$	$\times 2$
Training Accuracy	0.951	0.925	0.859	0.924	0.921	0.910
Testing Accuracy	0.951	0.929	0.861	0.926	0.922	0.913
Final DLG Error	310.2	350.3	412.5	301.1	336.7	389.7

[a]Considering the noise in Fig. 4 and Fig. 5 as the base level for Algorithm 1 and Algorithm 2, respectively.

$\{\gamma^k\}$ corresponds to a stronger privacy level but a slower convergence speed. The same conclusion can be drawn for Algorithm 2. In future work, we will systematically quantify the cost of achieving DP in distributed optimization under the constraint of provable convergence to the optimal solution. Furthermore, we also plan to investigate if gradually reducing communication frequency can enable rigorous DP.

ACKNOWLEDGEMENT

The authors would like to thanks Ben Liggett for the help in numerical experiments. They would also like to thank the associate editor and anonymous reviewers, whose comments helped improve the paper.

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 (9) it follows *a.s.* that for all $k \geq 0$,

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq \left(\begin{bmatrix} 1 & \frac{\gamma^k}{m} \\ 0 & 1-\kappa\gamma^k \end{bmatrix} + a^k \mathbf{1}\mathbf{1}^T\right)\mathbf{v}^k + b^k\mathbf{1} \quad (22)$$

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 (22) 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 (22) we have the following relation *a.s.* for the second element of $\mathbf{v}^k$:

$$\mathbb{E}\left[\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2|\mathcal{F}^k\right] \leq (1+a^k-\kappa\gamma^k)\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + \beta^k \quad (23)$$

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.*

$$\sum_{k=0}^\infty \kappa\gamma^k \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 < \infty, \lim_{k\to\infty} \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 = 0 \quad (24)$$

It remains to show that $\|\bar{x}^k - \theta^*\|^2 \to 0$ *a.s.* For this, we consider relation (9) and focus on the first element of $\mathbf{v}^k$, for which we obtain *a.s.* for all $k \geq 0$:

$$\mathbb{E}\left[\|\bar{x}^{k+1} - \theta^*\|^2|\mathcal{F}^k\right] \leq (1+a^k)\|\bar{x}^k - \theta^*\|^2 + \left(\frac{\gamma^k}{m} + a^k\right)\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + b^k - c^k(F(\bar{x}^k) - F(\theta^*)) \quad (25)$$

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 (24), 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

$$x^{k+1} = (I + \gamma^k W \otimes I_d)x^k + \gamma^k\zeta_w^k - \lambda^k g^k \quad (26)$$

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

From (26) we can obtain the following relationship for the average vector $\bar{x}^k = \frac{1}{m}\sum_{i=1}^m x_i^k$

$$\bar{x}^{k+1} = \bar{x}^k + \gamma^k\bar{\zeta}_w^k - \frac{\lambda^k}{m}\sum_{i=1}^m g_i^k \quad (27)$$

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

Using (27) 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

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

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 $k\geq 0$,

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{x}^{k+1}-\theta^*\right\|^2|\mathcal{F}^k\right]\\&\leq\left\|\bar{x}^k-\theta^*\right\|^2-\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k,\bar{x}^k-\theta^*\right\rangle\\&\quad+\frac{2}{m^2}(\lambda^k)^2\left\|\sum_{i=1}^m g_i^k\right\|^2+\frac{2}{m}(\gamma^k)^2\sum_{i=1}^m w_{ii}^2(\sigma_i^k)^2\end{aligned}\tag{28}$$

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

$$\begin{aligned}\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k,\bar{x}^k-\theta^*\right\rangle=&\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k-\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle\\&+\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle\end{aligned}\tag{29}$$

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

$$\begin{aligned}&\lambda^k\left\langle g_i^k-\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle\geq-L\lambda^k\|x_i^k-\bar{x}^k\|\|\bar{x}^k-\theta^*\|\\&\geq-\frac{\gamma^k}{2}\|x_i^k-\bar{x}^k\|^2-\frac{L^2(\lambda^k)^2}{2\gamma^k}\|\bar{x}^k-\theta^*\|^2\end{aligned}\tag{30}$$

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

$$\begin{aligned}\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle\nabla f_i(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle&=2\lambda^k\left\langle\nabla F(\bar{x}^k),\bar{x}^k-\theta^*\right\rangle\\&\geq 2\lambda^k(F(\bar{x}^k)-F(\theta^*))\end{aligned}\tag{31}$$

Combining (29), (30), and (31) leads to

$$\begin{aligned}\frac{2\lambda^k}{m}\sum_{i=1}^m\left\langle g_i^k,\bar{x}^k-\theta^*\right\rangle\geq&-\frac{\gamma^k}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2\\&-\frac{L^2(\lambda^k)^2}{\gamma^k}\|\bar{x}^k-\theta^*\|^2+2\lambda^k(F(\bar{x}^k)-F(\theta^*))\end{aligned}\tag{32}$$

We next estimate the second last term in (28):

$$\begin{aligned}\frac{1}{m^2}\left\|\sum_{i=1}^m g_i^k\right\|^2=&\frac{1}{m^2}\left\|\sum_{i=1}^m\left(g_i^k-\nabla f_i(\theta^*)\right)\right\|^2\\&\leq\frac{L^2}{m}\sum_{i=1}^m\left\|x_i^k-\theta^*\right\|^2=\frac{L^2}{m}\|x^k-x^*\|^2\end{aligned}\tag{33}$$

Further using the inequality

$$\begin{aligned}\|x^k-x^*\|^2&\leq\|x^k-\mathbf{1}\otimes\bar{x}^k+\mathbf{1}\otimes\bar{x}^k-x^*\|^2\\&\leq 2\|x^k-\mathbf{1}\otimes\bar{x}^k\|^2+2\|\mathbf{1}\otimes\bar{x}^k-x^*\|^2\\&\leq 2\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2+2m\|\bar{x}^k-\theta^*\|^2\end{aligned}\tag{34}$$

we have from (33) that

$$\frac{1}{m^2}\left\|\sum_{i=1}^m g_i^k\right\|^2\leq\frac{2L^2}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2+2L^2\|\bar{x}^k-\theta^*\|^2\tag{35}$$

Substituting (32) and (35) into (28) yields

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{x}^{k+1}-\theta^*\right\|^2|\mathcal{F}^k\right]\leq&\left\|\bar{x}^k-\theta^*\right\|^2+\frac{\gamma^k}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2\\&+L^2(\lambda^k)^2\left(\frac{1}{\gamma^k}+4\right)\|\bar{x}^k-\theta^*\|^2-2\lambda^k(F(\bar{x}^k)-F(\theta^*))\\&+\frac{4L^2(\lambda^k)^2}{m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2+\frac{2(\gamma^k)^2}{m}\sum_{i=1}^m w_{ii}^2(\sigma_i^k)^2\end{aligned}\tag{36}$$

Part II: Next we analyze $\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|^2$. Using (26) and (27), 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

$$x^{k+1}-\mathbf{1}\otimes\bar{x}^{k+1}=\hat{W}_k x^k+\gamma^k\Xi\zeta_w^k-\lambda^k\Xi g^k\tag{37}$$

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 (37), 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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\|x^{k+1} - \mathbf{1}\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k\right]\\
&\leq \left\|\hat{W}_k(x^k - \mathbf{1}\otimes\bar{x}^k) - \lambda^k\Xi g^k\right\|^2 + m(\gamma^k)^2\sum_{i=1}^m\sum_{j\in\mathbb{N}_i^{\rm in}} w_{ij}^2(\sigma_j^k)^2\\
&\leq \left(\|\hat{W}_k(x^k - \mathbf{1}\otimes\bar{x}^k)\| + \|\lambda^k\Xi g^k\|\right)^2 + m(\gamma^k)^2\max_{j\in[m]}(\sigma_j^k)^2 C_W
\end{aligned}$$

where $C_W = \sum_{i=1}^m\sum_{j\in\mathbb{N}_i^{\rm in}} w_{ij}^2$. 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 2), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|x^{k+1} - \mathbf{1}\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k\right]\\
&\leq (1-\gamma^k|\nu|)^2(1+\epsilon)\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2\\
&\quad + (1+\epsilon^{-1})(\lambda^k)^2\|g^k\|^2 + m(\gamma^k)^2\max_{j\in[m]}(\sigma_j^k)^2 C_W
\end{aligned} \tag{38}$$

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 (34), 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 (38) 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

$$\begin{aligned}
&\mathbb{E}\left[\|x^{k+1} - \mathbf{1}\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k\right]\\
&\leq \left(1 - \gamma^k|\nu| + \frac{4L^2(\lambda^k)^2}{|\nu|\gamma^k}\right)\|x^k - \mathbf{1}\otimes\bar{x}^k\|^2\\
&+ \frac{4mL^2(\lambda^k)^2}{|\nu|\gamma^k}\|\bar{x}^k - \theta^*\|^2 + \frac{4(\lambda^k)^2m^2}{|\nu|\gamma^k}\|\nabla f(x^*)\|^2\\
&+ m(\gamma^k)^2\max_{j\in[m]}(\sigma_j^k)^2 C_W
\end{aligned} \tag{39}$$

By combining (36) and (39), 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 $C_W = \sum_{i=1}^m\sum_{j\in\mathbb{N}_i^{\rm in}} w_{ij}^2$.

C. Proof of Lemma 10

Since the results of Lemma 5 are asymptotic, they remain valid when the starting index is shifted from $k = 0$ to $k = T$, for an arbitrary $T \geq 0$. So the idea is to show that the conditions in Lemma 5 are satisfied for all $k \geq T$ (for some large enough $T \geq 0$).

(a) Because $\kappa_i > 0$ for all $1 \leq i \leq 9$, for $\pi = [\pi_1, \pi_2, \pi_3]^T$ to satisfy $\pi^T V \leq \pi^T$ and $\pi^T H^k \geq 0$, we only need to show that the following inequalities can be true

$$\begin{aligned}
&\kappa_1\lambda^k\pi_1 + (1 - \kappa_3\gamma_1^k)\pi_2 \leq \pi_2,\\
&\kappa_2\lambda^k\pi_1 + (1 - \kappa_4\gamma_2^k)\pi_3 \leq \pi_3,\\
&\left(\kappa_6\lambda^k - \kappa_7(\lambda^k)^2\right)\pi_1 - \kappa_8\frac{(\lambda^k)^2}{\gamma_1^k}\pi_2 - \kappa_9\frac{(\lambda^k)^2}{\gamma_2^k}\pi_3 \geq 0
\end{aligned} \tag{40}$$

The first inequality is equivalent to $\pi_2 \geq \frac{\kappa_1\lambda^k}{\kappa_3\gamma_1^k}\pi_1$. Given that $\lim_{k\to\infty}\lambda^k/\gamma_1^k = 0$ holds and γ_1^k as well as λ^k are positive according to the assumption, it can easily be seen that for any given $\pi_1 > 0$, we can always find a $\pi_2 > 0$ satisfying the relationship when k is larger than some $T \geq 0$.

The second inequality is equivalent to $\pi_3 \geq \frac{\kappa_2\lambda^k}{\kappa_4\gamma_2^k}\pi_1$. Given that $\lim_{k\to\infty}\lambda^k/\gamma_2^k = 0$ holds and γ_2^k as well as λ^k are positive according to the assumption, it can easily be seen that for any given $\pi_1 > 0$, we can always find a $\pi_3 > 0$ satisfying the relationship when k is larger than some $T \geq 0$.

The third inequality is equivalent to $\pi_1 \geq \frac{\kappa_7\lambda^k}{\kappa_6}\pi_1 + \frac{\kappa_8\lambda^k}{\kappa_6\gamma_1^k}\pi_2 + \frac{\kappa_9\lambda^k}{\kappa_6\gamma_2^k}\pi_3$. Since the right hand side converges to zero according to our assumptions on λ^k, γ_1^k and γ_2^k, we can always find a constant π_1 satisfying this inequality for $k \geq T$. Thus, we can always find a vector π satisfying all inequalities in (40) for $k \geq T$ for some large enough $T \geq 0$, and hence the conditions in Lemma 5 are satisfied.

By Lemma 5, it follows that for the three entries of $\mathbf{v}^k$, i.e., $\mathbf{v}_1^k$, $\mathbf{v}_2^k$, and $\mathbf{v}_3^k$, we have that

$$\lim_{k\to\infty}\pi_1\mathbf{v}_1^k + \pi_2\mathbf{v}_2^k + \pi_3\mathbf{v}_3^k \tag{41}$$

exists *a.s.* and $\sum_{k=0}^\infty \pi^T H^k\mathbf{u}^k < \infty$ holds *a.s.* with $\mathbf{u}^k = [\|\nabla F(\bar{x}^k)\|^2, \ \|\bar{y}^k\|^2]^T$. Since $\pi^T H^k = \left[\kappa_5\lambda^k\pi_1, \ (\kappa_6\lambda^k - \kappa_7(\lambda^k)^2)\pi_1 - \kappa_8\frac{(\lambda^k)^2}{\gamma_1^k}\pi_2 - \kappa_9\frac{(\lambda^k)^2}{\gamma_2^k}\pi_3\right]$ and $(\lambda^k)^2$, $\frac{(\lambda^k)^2}{\gamma_1^k}$, and $\frac{(\lambda^k)^2}{\gamma_2^k}$ are summable, one has

$$\sum_{k=0}^\infty\lambda^k\|\nabla F(\bar{x}^k)\|^2 < \infty, \quad \sum_{k=0}^\infty\lambda^k\|\bar{y}^k\|^2 < \infty, \quad a.s. \tag{42}$$

Hence, it follows that

$$\|\nabla F(\bar{x}^k)\| < \Delta_1, \quad \|\bar{y}^k\| < \Delta_2 \quad a.s. \tag{43}$$

for some random scalars $\Delta_1 > 0$ and $\Delta_2 > 0$ due to the assumption $\sum_{k=0}^\infty\lambda^k = \infty$.

Now, we focus on proving that both $\mathbf{v}_2^k = \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_R^2$ and $\mathbf{v}_3^k = \|\mathbf{y}^k - \mathrm{diag}(v)\bar{\mathbf{y}}^k\|_C^2$ converge *a.s.* to 0. The idea is to show that we can apply Lemma 8. By focusing on the second and third elements of $\mathbf{v}^k$, i.e., $\mathbf{v}_2^k$ and $\mathbf{v}_3^k$, from (20) we have

$$\begin{bmatrix}\mathbf{v}_2^{k+1}\\ \mathbf{v}_3^{k+1}\end{bmatrix} \leq \left(\tilde{V}^k + a^k\mathbf{1}\mathbf{1}^T\right)\begin{bmatrix}\mathbf{v}_2^k\\ \mathbf{v}_3^k\end{bmatrix} + \hat{b}^k\mathbf{1} + \begin{bmatrix}\hat{c}^k\\ \hat{c}^k\end{bmatrix}$$

where $\hat{b}^k = b^k + a^k(F(\bar{x}^k) - F^*)$, $\hat{c}^k = \max\left\{\kappa_8 \frac{(\lambda^k)^2}{\gamma_1^k}\|\bar{y}^k\|^2, \kappa_9 \frac{(\lambda^k)^2}{\gamma_2^k}\|\bar{y}^k\|^2\right\}$, and $\tilde{V}^k = \begin{bmatrix} 1-\kappa_3\gamma_1^k & 0 \\ 0 & 1-\kappa_4\gamma_2^k \end{bmatrix}$, which can be rewritten as

$$\begin{bmatrix} \mathbf{v}_2^{k+1} \\ \mathbf{v}_3^{k+1} \end{bmatrix} \leq \tilde{V}^k \begin{bmatrix} \mathbf{v}_2^k \\ \mathbf{v}_3^k \end{bmatrix} + \tilde{b}^k \mathbf{1} \tag{44}$$

where $\tilde{b}^k = b^k + \hat{c}^k + a^k\left(F(\bar{x}^k) - F^* + \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_R^2 + \|\mathbf{y}^k - \mathrm{diag}(v)\bar{\mathbf{y}}^k\|_C^2\right)$.

To apply Lemma 8, noting that γ_1^k and γ_2^k are not summable, we show that the equation $\tilde{\pi}^T\tilde{V}^k = (1-\alpha\gamma_1^k)\tilde{\pi}^T$ has a solution in $\tilde{\pi} = [\pi_2, \pi_3]$ with $\pi_2, \pi_3 > 0$ and $\alpha \in (0,1)$. From $\tilde{\pi}^T\tilde{V}^k = (1-\alpha\gamma_1^k)\tilde{\pi}^T$, one has

$$(1-\kappa_3\gamma_1^k)\pi_2 \leq (1-\alpha\gamma_1^k)\pi_2, \quad (1-\kappa_4\gamma_2^k)\pi_3 \leq (1-\alpha\gamma_1^k)\pi_3$$

which can be simplified as $\alpha \leq \kappa_3$, $\alpha \leq \frac{\kappa_4\gamma_2^k}{\gamma_1^k}$.

Given $\lim_{k\to\infty}\gamma_1^k/\gamma_2^k < \infty$ according to our assumption, it can be seen that $\frac{\kappa_4\gamma_2^k}{\gamma_1^k}$ is positive, and hence, such an $\alpha \in (0,1)$ and $\tilde{\pi} > 0$ can always be found.

We next prove that the condition $\sum_{k=0}^{\infty}\tilde{b}^k < 0$ *a.s.* of Lemma 8 is also satisfied. Indeed, the condition can be met because: (1) b^k, a^k, $\frac{(\lambda^k)^2}{\gamma_1^k}$, and $\frac{(\lambda^k)^2}{\gamma_2^k}$ are all summable according to the assumption of the lemma; and (2) $\|\bar{y}^k\|$ (see (43)) and $F(\bar{x}^k) - F^*$, $\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_R^2$, $\|\mathbf{y}^k - \mathrm{diag}(v)\bar{\mathbf{y}}^k\|_C^2$ are all bounded *a.s.* due to the existence of the limit in (41). Thus, all the conditions of Lemma 8 are satisfied, so it follows that $\lim_{k\to\infty}\|x_i^k - \bar{x}^k\| = 0$ and $\lim_{k\to\infty}\|y_i^k - v_i\bar{y}^k\| = 0$ *a.s.* Moreover, in view of the existence of the limit in (41) and the facts that $\pi_1 > 0$ and $v_1^k = F(\bar{x}^k) - F(\theta^*)$, it follows that $\lim_{k\to\infty}F(\bar{x}^k)$ exists *a.s.*

(b) Since $\sum_{k=0}^{\infty}\lambda^k\|\nabla F(\bar{x}^k)\|^2 < \infty$ holds *a.s.* (see (42)), from $\sum_{k=0}^{\infty}\lambda^k = \infty$, it follows that we have $\liminf_{k\to\infty}\|\nabla F(\bar{x}^k)\| = 0$ *a.s.*

Now, if the function $F(\cdot)$ has bounded level sets, then the sequence $\{\bar{x}^k\}$ is *a.s.* bounded since $\lim_{k\to\infty}F(\bar{x}^k)$ exists *a.s.* (as shown in part (a)). Thus, $\{\bar{x}^k\}$ *a.s.* has accumulation points. Let $\{\bar{x}^{k_i}\}$ be a sub-sequence such that $\lim_{i\to\infty}\|\nabla F(\bar{x}^{k_i})\| = 0$ *a.s.* Without loss of generality, we may assume that $\{\bar{x}^{k_i}\}$ is *a.s.* convergent, for otherwise we would choose a sub-sequence of $\{\bar{x}^{k_i}\}$. Let $\lim_{i\to\infty}\bar{x}^{k_i} = \hat{x}$. Then, by the continuity of the gradient $\nabla F(\cdot)$, it follows $\nabla F(\hat{x}) = 0$, implying that $\hat{x}$ is an optimal point. Since $F(\cdot)$ is continuous, we have $\lim_{i\to\infty}F(\bar{x}^{k_i}) = F(\hat{x}) = F^*$. By part (a), $\lim_{k\to\infty}F(\bar{x}^k)$ exists *a.s.*, so we must have $\lim_{k\to\infty}F(\bar{x}^k) = F^*$ *a.s.*

Finally, by part (a), we have $\lim_{k\to\infty}\|x_i^k - \bar{x}^k\|^2 = 0$ *a.s.* for every i. Thus, each $\{x_i^k\}$ has the same accumulation points as the sequence $\{\bar{x}^k\}$ *a.s.*, implying by the continuity of the function $F(\cdot)$ that $\lim_{k\to\infty}F(x_i^k) = F^*$ *a.s.* for all i.

REFERENCES

[1] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[2] J. A. Bazerque and G. B. Giannakis, "Distributed spectrum sensing for cognitive radio networks by exploiting sparsity," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1847–1862, 2009.

[3] C. Zhang and Y. Wang, "Distributed event localization via alternating direction method of multipliers," *IEEE Transactions on Mobile Computing*, vol. 17, no. 2, pp. 348–361, 2017.

[4] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning," in *Proceedings of the 50th annual Allerton Conference on Communication, Control, and Computing*, 2012, pp. 1543–1550.

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

[6] 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.

[7] 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.

[8] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

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

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

[26] Y. Wang and H. V. Poor, "Decentralized stochastic optimization with inherent privacy protection," *IEEE Transactions on Automatic Control*, vol. 68, no. 4, pp. 2293–2308, 2023.

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

[28] 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.

[29] S. Han, U. Topcu, and G. J. Pappas, "Differentially private distributed constrained optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 50–64, 2016.
[30] 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.
[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] Y. Wang and T. Başar, "Decentralized nonconvex optimization with guaranteed privacy and accuracy," *Automatica*, vol. 150, p. 110858, 2023.
[38] 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.
[39] B. Polyak, "Introduction to optimization," *Optimization software Inc., Publications Division, New York*, vol. 1, 1987.
[40] K. L. Chung, "On a stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 463–483, 1954.
[41] 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.
[42] 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.
[43] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.
[44] Y. Wang, "The spectral radius of a square matrix can be approximated by its" weighted" spectral norm," *arXiv preprint arXiv:2304.10421*, 2023.
[45] Y. Wang and A. Nedic, "Tailoring gradient methods for differentially-private distributed optimization," *arXiv preprint arXiv:2202.01113*, 2022.
[46] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1376–1385.
[47] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference*. Springer, 2016, pp. 635–658.
[48] I. Mironov, "Rényi differential privacy," in *The 30th Computer Security Foundations Symposium*. IEEE, 2017, pp. 263–275.
[49] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1994.
[50] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

Yongqiang Wang was born in Shandong, China. He received dual B.S. degrees in electrical engineering & automation and computer science & technology from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2004, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2009. From 2007-2008, he was with the University of Duisburg-Essen, Germany, as a visiting student. He was a Project Scientist at the University of California, Santa Barbara before joining Clemson University, SC, USA, where he is currently an Associate Professor. His current research interests include distributed control, optimization, and learning, with an emphasis on privacy protection. He currently serves as an associate editor for *IEEE Transactions on Automatic Control* and *IEEE Transactions on Control of Network Systems*.

Angelia Nedić holds a Ph.D. from Moscow State University, Moscow, Russia, in Computational Mathematics and Mathematical Physics (1994), and a Ph.D. from Massachusetts Institute of Technology, Cambridge, USA in Electrical and Computer Science Engineering (2002). She has worked as a senior engineer in BAE Systems North America, Advanced Information Technology Division at Burlington, MA. She is a recipient (jointly with her co-authors) of the Best Paper Award at the Winter Simulation Conference 2013 and the Best Paper Award at the International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt) 2015 (with co-authors). Her current interest is in large-scale optimization, games, control and information processing in networks.