Decentralized Gradient Methods with Time-varying Uncoordinated Stepsizes: Convergence Analysis and Privacy Design

Yongqiang Wang, Angelia Nedić

***Abstract*—Decentralized optimization enables a network of agents to cooperatively optimize an overall objective function without a central coordinator and is gaining increased attention in domains as diverse as control, sensor networks, data mining, and robotics. However, the information sharing among agents in decentralized optimization also discloses agents' information, which is undesirable or even unacceptable when involved data are sensitive. This paper proposes two gradient based decentralized optimization algorithms that can protect participating agents' privacy without compromising optimization accuracy or incurring heavy communication/computational overhead. Both algorithms leverage a judiciously designed mixing matrix and time-varying uncoordinated stepsizes to enable privacy, one using diminishing stepsizes while the other using non-diminishing stepsizes. In both algorithms, when interacting with any one of its neighbors, a participating agent only needs to share *one* message in each iteration, which is in contrast to most gradient-tracking based algorithms requiring every agent to share two messages (an optimization variable and a gradient-tracking variable) under non-diminishing stepsizes. Furthermore, both algorithms can guarantee the privacy of a participating agent even when all information shared by the agent are accessible to an adversary, a scenario in which most existing accuracy-maintaining privacy approaches will fail to protect privacy. Simulation results confirm the effectiveness of the proposed algorithms.**

I. Introduction

Distributed optimization is gaining increased attention across disciplines due to its fundamental importance and vast applications in areas ranging from cooperative control [1], distributed sensing [2], multi-agent systems [3], sensor networks [4], to large-scale machine learning [5]. In many of these applications, the problem can be formulated in the following general form, in which a network of m agents cooperatively solve a common optimization problem through on-node computation and local communication:

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

where θ is common to all agents but $f_i : \mathbb{R}^d \to \mathbb{R}$ is a local objective function private to agent i. We denote an optimal solution to this problem by θ^*, which we assume to be finite.

Since the 1980s, the above decentralized optimization problem has been intensively studied. To date, various algorithms have been proposed. Some of the commonly used algorithms include decentralized gradient methods (e.g., [6], [7], [8], [9]), distributed alternating direction method of multipliers (e.g., [10], [11]), and distributed Newton methods (e.g., [12]). We focus on the gradient based approach due to its simplicity in computation, which is particularly appealing when agents have limited computational capabilities.

Over the past decade, plenty of gradient based algorithms have been developed for decentralized optimization. Early results combine consensus and gradient method by directly concatenating gradient based step with a consensus operation of the optimization variable. Typical examples include [6], [13]. However, to find an exact optimal solution, these approaches have to use a diminishing stepsize, which slows down the convergence. To guarantee both a fast convergence speed and exact optimization result, algorithms have been proposed to replace the local gradient in decentralized gradient methods with an auxiliary variable which tracks the gradient of the global objective function. Typical examples include Aug-DGM [8], DIGing [14], AsynDGM [15], AB [9], Push-Pull [16], [17] and ADD-OPT [18], etc. While these algorithms can converge to an exact optimal solution under a fixed stepsize, they have to exchange both the optimization variable and the additional auxiliary variable in every iteration.

All aforementioned algorithms explicitly share optimization variables and/or gradients in every iteration, which becomes a problem in applications involving sensitive data. For example, in the rendezvous problem where a group of agents use decentralized optimization to cooperatively find an optimal assembly point, participating agents may want to keep their initial positions private in unfriendly environments [11]. In fact, without an effective privacy mechanism in place, the results in [11], [19], [20] show that a participating agent's position can be easily inferred by an adversary in decentralized-optimization based rendezvous and parameter estimation. Another case underscoring the importance of privacy preservation in decentralized optimization is machine learning where training data may contain sensitive information such as medical records and salary information [21], [22]. In fact, recent results in [23] show that without a privacy mechanism, an adversary can precisely recover the raw data (pixel-wise accurate for images and token-wise matching for texts) through shared gradients.

Recently results have been reported on privacy-preserving decentralized optimization [11], [19], [24], [25], [26], [27], [28], [29], [30], [31]. For example, differential-privacy based approaches have been proposed to obscure information in decentralized optimization by injecting noise to exchanged messages [19], [24], [25] or objective functions [32]. However, the added noise in differential privacy also unavoidably compromises the accuracy of optimization results. To enable privacy protection without sacrificing optimization accuracy,

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

partially homomorphic encryption has been employed in both our own prior results [11], [26], and others' [27], [28]. However, such approaches incur heavy communication and computation overhead. Employing the structural properties of decentralized optimization, results have also been reported on privacy protection in decentralized optimization without using differential privacy or encryption. For example, [21], [33] showed that privacy can be enabled by adding a *constant* uncertain parameter in the projection step or stepsizes. The authors of [29] showed that network structure can be leveraged to construct spatially correlated "structured" noise to cover information. Although these approaches can ensure convergence to an exact optimal solution, 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 requires *a priori* knowledge of the optimal solution; "structured" noise based approach requires each agent to have a certain number of neighbors which do not share information with the adversary. In fact, such a structure constraint is required in most privacy solutions with guaranteed optimization accuracy, including encryption based privacy approaches [11].

Inspired by our recent results that privacy can be enabled in consensus by manipulating inherent dynamics [34], [35], [36], we propose to enable privacy in decentralized gradient methods by judiciously manipulating the inherent dynamics of information mixing and gradient operations. More specifically, leveraging a judiciously designed mixing matrix and time-varying uncoordinated stepsizes, we propose two privacy-preserving decentralized gradient based algorithms, one with diminishing stepsizes and the other one with non-diminishing stepsizes. Not only do our algorithms maintain the accuracy of decentralized optimization, they also enable privacy even when an adversary has access to all messages shared by a participating agent. This is in contrast to most existing accuracy-guaranteed privacy approaches for decentralized optimization which cannot protect an agent against adversaries having access to all shared messages. Furthermore, even in the non-diminishing stepsize case, our algorithm only requires a participating agent to share one variable with any one of its neighboring agents in each iteration, which is extremely appealing when communication bandwidth is limited. In fact, to our knowledge, our algorithm is the first privacy-preserving decentralized gradient based algorithm that uses non-diminishing stepsizes to reach accurate optimization results but requires each participating agent to share only one message with a neighboring agent in every iteration. Note that most existing gradient-tracking based algorithms (e.g., [8], [9], [15], [16], [17], [18], [37], [38], [39], [40], [41]) require an agent to share two messages (the optimization variable and a gradient-tracking variable) in every iteration.

The main contributions are as follows: 1) We propose two accuracy-guaranteed decentralized gradient based algorithms that can protect the privacy of participating agents even when all shared messages are accessible to an adversary, a scenario which fails existing accuracy-guaranteed privacy-preserving approaches for decentralized optimization; 2) The two inherently privacy-preserving algorithms are efficient in communication/computation in that they are encryption-free and only require a participating agent to share *one* message with a neighboring agent in every iteration. This is significant in that, as a comparison, existing gradient-tracking based decentralized optimization algorithms require a participating agent to share both the optimization variable and a gradient-tracking variable in every iteration[1]. In fact, the sharing of the additional gradient-tracking variable will lead to privacy breaches, as detailed in Sec. IV-B; 3) Even without considering privacy, to our knowledge, our convergence analysis is the first to characterize decentralized gradient methods in the presence of time-varying heterogeneity in stepsizes, which is in contrast to existing results only addressing constant or fixed heterogeneity in stepsizes [8], [33], [42].

The organization of the paper is as follows. Sec. II gives the problem formulation. Sec. III presents PDG-DS, an inherently privacy-preserving decentralized gradient algorithm with proven converge to the exact optimization solution under diminishing uncoordinated stepsizes. Sec. IV presents PDG-NDS, an inherently privacy-preserving decentralized gradient algorithm with proven converge to the exact optimization solution under non-diminishing and time-varying uncoordinated stepsizes. Sec. V gives simulation results and comparison with existing works. Finally Sec. VI concludes the paper.

Notations: I_d denotes identity matrix of dimension d. $\mathbf{1}_d$ denotes a d dimensional column vector will all entries equal to 1 and we omit the dimension when clear from the context. For a vector x, x_i denotes the ith element. For two matrices A and B with the same dimension, we say $A < B$ (resp. $A \leq B$) if all entries of $A - B$ are negative (resp. non-positive). A^T denotes the transpose of A and $\langle \cdot, \cdot \rangle$ denotes the inner product. $\| \cdot \|$ denotes the Euclidean norm for a vector or the induced Euclidean norm for a matrix. Matrix A is column-stochastic (resp. row-stochastic) when its entries are nonnegative and elements in every column (resp. row) add up to one. A is doubly stochastic when it is both column-stochastic and row-stochastic. $\otimes$ represents the Kronecker product.

II. Problem Formulation

We consider a network of m agents. The agents interact on an undirected graph, which can be described by a weight matrix $W = \{w_{ij}\}$. More specifically, if agents i and j can interact with each other, then w_{ij} is positive. Otherwise, w_{ij} will be zero. We assume that an agent is always able to affect itself, i.e., $w_{ii} > 0$ for all $1 \leq i \leq m$. The neighbor set $\mathbb{N}_i$ of agent i is defined as the set of agents $\{j | w_{ij} > 0\}$. So the neighbor set of agent i always includes itself.

Assumption 1. *$W = \{w_{ij}\} \in \mathbb{R}^{m \times m}$ satisfies $\mathbf{1}^T W = \mathbf{1}^T$, $W\mathbf{1} = \mathbf{1}$, and $\eta = \|W - \frac{\mathbf{1}\mathbf{1}^T}{m}\| < 1$.*

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

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

[1]Some gradient-tracking based algorithms may be transformed to a one-variable form like EXTRA [7]; however, such an implementation becomes infeasible when the stepsizes or coupling weights are uncoordinated to enable privacy. See Remark 7 and Sec. IV-B for detailed explanations.

where $x_i \in \mathbb{R}^d$ is the local estimate of agent i about the optimization solution and $x = [x_1^T, x_2^T, \cdots, x_m^T]^T \in \mathbb{R}^{md}$.

We make the following standard assumption on objective functions:

Assumption 2. *Problem (1) has at least one optimal solution θ^*. Every f_i has Lipschitz continuous gradients, i.e., for some $L > 0$, $\|\nabla f_i(u) - \nabla f_i(v)\| \leq L\|u - v\|, \quad \forall i$ and $\forall u, v \in \mathbb{R}^d$. Every f_i is convex, i.e., $f_i(u) \geq f_i(v) + \nabla f_i(v)^T(u - v)$, $\forall i$ and $\forall u, v \in \mathbb{R}^d$.*

Under Assumption 2, we know that (2) always has an optimal solution $x^* = [(\theta^*)^T, (\theta^*)^T, \cdots, (\theta^*)^T]^T$.

In decentralized optimization applications, gradients usually carry sensitive information. For example, in decentralized-optimization based rendezvous and localization, disclosing the gradient of an agent amounts to disclosing its (initial) position [11], [19]. In machine learning, gradients are directly calculated from and carry information of raw training data [23]. Therefore, in this paper, we define privacy as preventing disclosing agents' gradients in each iteration.

We consider two potential attacks [43]:

- *Honest-but-curious attacks* are attacks in which a participating agent or multiple participating agents (colluding or not) follows all protocol steps correctly but is curious and collects all received intermediate data to learn the sensitive information about other participating agents.
- *Eavesdropping attacks* are attacks in which an external eavesdropper wiretaps all communication channels to intercept exchanged messages so as to learn sensitive information about sending agents.

III. An inherently privacy-preserving decentralized gradient algorithm with diminishing stepsizes

Conventional decentralized gradient algorithms usually take the following form:

$$x_i^{k+1} = \sum\nolimits_{j \in \mathbb{N}_i} w_{ij} x_j^k - \lambda^k g_i^k \tag{3}$$

where λ^k is a positive scalar denoting the stepsize and g_i^k denotes the gradient of agent i evaluated at x_i^k. It is well-known that under Assumption 1 and Assumption 2, when λ^k satisfies $\sum_{k=0}^{\infty} \lambda^k = \infty$ and $\sum_{k=0}^{\infty} (\lambda^k)^2 < \infty$, all x_i^k will converge to a same optimal solution. However, in (3), agent i has to share x_i^k with all its neighbors. If an adversary has access to x_i^k and the updates that agent i receives from all its neighbors, then the adversary can easily infer g_i^k based on the update rule (3) and publicly known W and λ^k.

Motivated by this observation and inspired by our recent finding that interaction dynamics can be judiciously manipulated to enable privacy [34], [35], [36], we propose the following decentralized gradient algorithm (with per-agent version given in Algorithm PDG-DS) to enable privacy by adapting the stochastic optimization algorithm in [44]:

$$x^{k+1} = (W \otimes I_d)x^k - ((B^k \Lambda^k) \otimes I_d)g^k \tag{4}$$

where $B^k = \{b_{ij}^k\}$ is a column-stochastic nonnegative matrix, $\Lambda^k = \text{diag}[\lambda_1^k, \lambda_2^k, \cdots, \lambda_m^k]$ with $\lambda_i^k \geq 0$ denoting the stepsize of agent i at iteration k and $g^k = [(g_1^k)^T, (g_2^k)^T, \cdots (g_m^k)^T]^T$. Different from [44] which uses a matrix-valued stepsize for each agent, we here require the stepsize λ_j^k for each agent to be a scalar, which is necessary to prove deterministic convergence to an exact optimal solution. It is worth noting that our scalar stepsize here cannot be viewed as a special case of the matrix-valued stepsize in [44] since the stepsize matrix in [44] explicitly requires all diagonal entries to be statistically independent of each other, which prohibits it from having the form of $\lambda_j^k I_d$.

The detailed implementation procedure for individual agents is provided in Algorithm PDG-DS. Compared with the conventional decentralized gradient algorithm, it can be seen that we equip each agent j with two private variables b_{ij}^k and λ_j^k to cover its gradient information g_j^k. The two variables are generated by and only known to agent j. Therefore, the two variables can ensure that a neighboring agent i cannot infer g_j^k based on received information $w_{ij}x_j^k - b_{ij}^k \lambda_j^k g_j^k$ as agent i does not know x_j^k, λ_j^k, or b_{ij}^k. It is worth noting that since agent j determines $b_{ij}^k > 0$ for all $i \in \mathbb{N}_j$ ($b_{ij}^k = 0$ for $i \notin \mathbb{N}_j$), as long as every agent j ensures $\sum_{i \in \mathbb{N}_j} b_{ij}^k = 1$ locally, the column-stochastic condition for matrix B^k will be satisfied.

PDG-DS: Privacy-preserving decentralized gradient method with diminishing stepsizes

Public parameters: W

Private parameters for agent i: $b_{ji}^k \geq 0$, $\lambda_i^k \geq 0$, and x_i^0

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

a) Every agent j computes and sends to agent $i \in \mathbb{N}_j$

$$v_{ij}^k \triangleq w_{ij} x_j^k - b_{ij}^k \lambda_j^k g_j^k \tag{5}$$

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

$$x_i^{k+1} = \sum\nolimits_{j \in \mathbb{N}_i} v_{ij}^k = \sum\nolimits_{j \in \mathbb{N}_i} (w_{ij} x_j^k - b_{ij}^k \lambda_j^k g_j^k) \tag{6}$$

c) **end**

A. Convergence Analysis

We define $\bar{x}^k = \frac{\sum_{i=1}^m x_i^k}{m}$. Because B^k and W are column stochastic, from (4), we can obtain

$$\bar{x}^{k+1} = \bar{x}^k - \tfrac{1}{m} \textstyle\sum_{i=1}^m \lambda_i^k g_i^k \tag{7}$$

To analyze PDG-DS, we first introduce a theorem that applies to general decentralized algorithms for solving (1).

Proposition 1. *Assume that problem (1) is convex and has a solution. Suppose that a distributed algorithm generates sequences $\{x_i^k\} \subseteq \mathbb{R}^d$ such that the following relation is satisfied for any optimal solution θ^* and for all k,*

$$\mathbf{v}^{k+1} \leq \left(\begin{bmatrix} 1 & 1 \\ 0 & \eta \end{bmatrix} + a^k \mathbf{1}\mathbf{1}^T \right) \mathbf{v}^k + b^k \mathbf{1} - c^k \begin{bmatrix} \zeta \\ 0 \end{bmatrix} \tag{8}$$

where $\mathbf{v}^k = \begin{bmatrix} \mathbf{v}_1^k \\ \mathbf{v}_2^k \end{bmatrix} \triangleq \begin{bmatrix} \|\bar{x}^k - \theta^*\|^2 \\ \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \end{bmatrix}$, $\zeta \triangleq \sum_{i=1}^m \big(f_i(\bar{x}^k) - f_i(\theta^*)\big)$, and the scalar sequences $\{a^k\}$, $\{b^k\}$, and $\{c^k\}$ are nonnegative and satisfy $\sum_{k=0}^{\infty} a^k <$

∞, $\sum_{k=0}^{\infty} b^k < \infty$, *and* $\sum_{k=0}^{\infty} c^k = \infty$. *Then, we have* $\lim_{k\to\infty} \|x_i^k - \bar{x}^k\| = 0$ *for all* i *and there exists an optimal solution* $\tilde{\theta}^*$ *such that* $\lim_{k\to\infty} \|\bar{x}^k - \tilde{\theta}^*\| = 0$.

Proof. Since θ^* is an optimal solution of problem (1), we always have $\sum_{i=1}^m \left(f_i(\bar{x}^k) - f_i(\theta^*)\right) \geq 0$.

From (8) it follows that for all $k \geq 0$,

$$\mathbf{v}^{k+1} \leq \left(\begin{bmatrix} 1 & 1 \\ 0 & \eta \end{bmatrix} + a^k \mathbf{1}\mathbf{1}^T \right) \mathbf{v}^k + b^k \mathbf{1} \tag{9}$$

Consider the vector $\pi = [1, \frac{1}{1-\eta}]^T$ and note $\pi^T \begin{bmatrix} 1 & 1 \\ 0 & \eta \end{bmatrix} = \pi^T$. Thus, the sequence $\{\pi^T \mathbf{v}^k\}$ satisfies all conditions of Lemma 3 in the Appendix. Therefore, it follows that $\lim_{k\to\infty} \pi^T \mathbf{v}^k$ exists and that $\{\|\bar{x}^k - \theta^*\|\}$ and $\{\sum_{i=1}^m \|x^k - \bar{x}^k\|^2\}$ are bounded.

We use $M > 0$ to represent an upper bound on $\{\|\bar{x}^k - \theta^*\|\}$ and $\{\sum_{i=1}^m \|x^k - \bar{x}^k\|^2\}$, i.e., $\|\bar{x}^k - \theta^*\| \leq M$ and $\sum_{i=1}^m \|x^k - \bar{x}^k\|^2 \leq M$ hold $\forall k \geq 0$. Thus, for all $k \geq 0$, we have

$$\begin{aligned} &\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2 \leq \eta \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 2a^k M + b^k \\ &= \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 - (1-\eta) \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 2a^k M + b^k \end{aligned} \tag{10}$$

By summing (10) over k and using the fact $\sum_{k=0}^{\infty}(2a^k M + b^k) < \infty$, we obtain

$$(1-\eta) \sum\nolimits_{k=0}^{\infty} \sum\nolimits_{i=1}^m \|x_i^k - \bar{x}^k\|^2 < \infty \tag{11}$$

which implies $\lim_{k\to\infty} \|x_i^k - \bar{x}^k\|^2 = 0$ for all i.

Next, we consider the first element of $\mathbf{v}^k$, i.e., $\|\bar{x}^k - \theta^*\|^2$. From (8) we have

$$\begin{aligned} \|\bar{x}^{k+1} - \theta^*\|^2 &\leq (1+a^k) \left(\|\bar{x}^k - \theta^*\|^2 + \sum\nolimits_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \right) \\ &\quad + b^k - c^k \sum\nolimits_{i=1}^m \left(f_i(\bar{x}^k) - f_i(\theta^*)\right) \\ &\leq \|\bar{x}^k - \theta^*\|^2 + \sum\nolimits_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 2a^k M \\ &\quad + b^k - c^k \sum\nolimits_{i=1}^m \left(f_i(\bar{x}^k) - f_i(\theta^*)\right) \end{aligned}$$

We can see that the preceding relation satisfies the relation in Lemma 5 in the Appendix with $\phi = \sum_{i=1}^m f_i$, $z^* = \theta^*$, $z^k = \bar{x}^k$, $\alpha^k = 0$, $\gamma^k = c^k$, and $\beta^k = \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 2a^k M + b^k$. By our assumption, we have $\sum_{k=0}^{\infty} a^k < \infty$, $\sum_{k=0}^{\infty} b^k < \infty$, and $\sum_{k=0}^{\infty} c^k = \infty$. Thus, in view of relation (11), it follows $\sum_{k=0}^{\infty} \beta^k < \infty$. Hence, all conditions of Lemma 5 in the Appendix are satisfied, and it follows that $\{\bar{x}^k\}$ converges to some optimal solution. ∎

Now, we are in position to prove convergence of PDG-DS.

Theorem 1. *Under Assumption 1 and Assumption 2, if the stepsize of every agent* i *is non-negative and satisfies* $\sum_{k=0}^{\infty} \lambda_i^k = \infty$ *and* $\sum_{k=0}^{\infty} (\lambda_i^k)^2 < \infty$, *and the stepsize heterogeneity satisfies*

$$\sum\nolimits_{k=0}^{\infty} \sum\nolimits_{i,j\in\{1,2,\cdots,m\},\, i\neq j} |\lambda_i^k - \lambda_j^k| < \infty \tag{12}$$

then we have $\lim_{k\to\infty} \|x_i^k - \bar{x}^k\| = 0$ *for all* i, *and there exists an optimal solution* $\tilde{\theta}^*$ *such that* $\lim_{k\to\infty} \|\bar{x}^k - \tilde{\theta}^*\| = 0$ *holds.*

Proof. The basic idea is to show that Proposition 1 applies. So we have to establish necessary relationships for $\|\bar{x}^k - \theta^*\|^2$ and $\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$, which are fulfilled in Step I and Step II below, respectively.

Step I: Relationship for $\|\bar{x}^k - \theta^*\|^2$. Using (7), we have for any optimal solution θ^*

$$\bar{x}^{k+1} - \theta^* = \bar{x}^k - \theta^* - \tfrac{1}{m} \textstyle\sum_{i=1}^m \lambda_i^k g_i^k$$

which further implies

$$\begin{aligned} \left\|\bar{x}^{k+1} - \theta^*\right\|^2 &= \left\|\bar{x}^k - \theta^*\right\|^2 - \tfrac{2}{m} \textstyle\sum_{i=1}^m \langle \lambda_i^k g_i^k, \bar{x}^k - \theta^* \rangle \\ &\quad + \tfrac{1}{m^2} \left\| \textstyle\sum_{i=1}^m \lambda_i^k g_i^k \right\|^2 \end{aligned} \tag{13}$$

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

$$\begin{aligned} \langle \lambda_i^k g_i^k, \bar{x}^k - \theta^* \rangle &= \langle \lambda_i^k (g_i^k - \nabla f_i(\bar{x}^k)), \bar{x}^k - \theta^* \rangle \\ &\quad + \langle \lambda_i^k \nabla f_i(\bar{x}^k), \bar{x}^k - \theta^* \rangle \end{aligned} \tag{14}$$

By the Lipschitz continuous property of ∇f_i, we obtain

$$\begin{aligned} &\langle \lambda_i^k (g_i^k - \nabla f_i(\bar{x}^k)), \bar{x}^k - \theta^* \rangle \geq -L\lambda_i^k \|x_i^k - \bar{x}^k\| \|\bar{x}^k - \theta^*\| \\ &\geq -\tfrac{1}{2}\|x_i^k - \bar{x}^k\|^2 - \tfrac{1}{2} L^2 (\lambda_i^k)^2 \|\bar{x}^k - \theta^*\|^2 \end{aligned} \tag{15}$$

Defining the average stepsize $\bar{\lambda}^k = \frac{\sum \lambda_i^k}{m}$, we have

$$\begin{aligned} &\langle \lambda_i^k \nabla f_i(\bar{x}^k), \bar{x}^k - \theta^* \rangle = \\ &\langle (\lambda_i^k - \bar{\lambda}^k) \nabla f_i(\bar{x}^k), \bar{x}^k - \theta^* \rangle + \langle \bar{\lambda}^k \nabla f_i(\bar{x}^k), \bar{x}^k - \theta^* \rangle \end{aligned} \tag{16}$$

Defining $\lambda^k = [\lambda_1^k, \cdots, \lambda_m^k]^T$ and combining (14)-(16) yield

$$\begin{aligned} &\frac{\sum_{i=1}^m \langle \lambda_i^k g_i^k, \bar{x}^k - \theta^* \rangle}{m} \\ &\geq -\frac{\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2}{2m} - \frac{L^2 \|\lambda^k\|^2 \|\bar{x}^k - \theta^*\|^2}{2m} + \\ &\frac{\sum_{i=1}^m \langle (\lambda_i^k - \bar{\lambda}^k) \nabla f_i(\bar{x}^k), \bar{x}^k - \theta^* \rangle}{m} + \bar{\lambda}^k \langle \nabla F(\bar{x}^k), \bar{x}^k - \theta^* \rangle \\ &\geq -\frac{\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2}{2m} - \frac{L^2 \|\lambda^k\|^2 \|\bar{x}^k - \theta^*\|^2}{2m} + \\ &\frac{\sum_{i=1}^m \langle (\lambda_i^k - \bar{\lambda}^k) \nabla f_i(\bar{x}^k), \bar{x}^k - \theta^* \rangle}{m} + \bar{\lambda}^k (F(\bar{x}^k) - F(\theta^*) \end{aligned} \tag{17}$$

where we used the convexity of $F(\cdot)$ in the last inequality.

Noting $\lambda^k = [\lambda_1^k, \cdots, \lambda_m^k]^T$, we always have

$$\begin{aligned} &\frac{\sum_{i=1}^m \langle (\lambda_i^k - \bar{\lambda}^k) \nabla f_i(\bar{x}^k), \bar{x}^k - \theta^* \rangle}{m} \\ &\geq -\frac{\| \sum_{i=1}^m (\lambda_i^k - \bar{\lambda}^k) \nabla f_i(\bar{x}^k) \| \, \|\bar{x}^k - \theta^*\|}{m} \\ &= -\frac{\| \left((\lambda^k - \bar{\lambda}^k \mathbf{1}_m) \otimes \mathbf{1}_d \right)^T m \nabla f(\mathbf{1} \otimes \bar{x}^k) \| \, \|\bar{x}^k - \theta^*\|}{m} \\ &\geq -\sqrt{d} \|\lambda^k - \bar{\lambda}^k \mathbf{1}_m\| \, \|\nabla f(\mathbf{1} \otimes \bar{x}^k)\| \, \|\bar{x}^k - \theta^*\| \end{aligned} \tag{18}$$

where we used Cauchy-Schwarz inequality and $m\nabla f(\mathbf{1} \otimes \bar{x}^k)) = [(\nabla f_1(\bar{x}^k))^T, \cdots, (\nabla f_m(\bar{x}^k))^T]^T$ in the last equality. Furthermore, $\|\nabla f(\mathbf{1} \otimes \bar{x}^k)\|$ can be bounded by using $\nabla f(x^*) = 0$ at $x^* = \mathbf{1} \otimes \theta^*$ as follows:

$$\begin{aligned} \|\nabla f(\mathbf{1} \otimes \bar{x}^k)\| &= \|\nabla f(\mathbf{1} \otimes \bar{x}^k) - \nabla f(x^*)\| \\ &\leq L \|\mathbf{1} \otimes \bar{x}^k - x^*\| = L\sqrt{m} \, \|\bar{x}^k - \theta^*\| \end{aligned} \tag{19}$$

Combining (17), (18), and (19) yields

$$\begin{aligned}&\frac{\sum_{i=1}^m \langle \lambda_i^k g_i^k, \bar{x}^k - \theta^* \rangle}{m} \\ &\geq -\frac{\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2}{2m} - \frac{L^2 \|\lambda^k\|^2 \|\bar{x}^k - \theta^*\|^2}{2m} - \\ &L\sqrt{md}\|\lambda^k - \bar{\lambda}^k \mathbf{1}_m\| \, \|\bar{x}^k - \theta^*\|^2 + \bar{\lambda}^k (F(\bar{x}^k) - F(\theta^*)\end{aligned} \tag{20}$$

We next estimate the last term in relation (13), for which we use $\frac{1}{m} g^k = \nabla f(x^k)$, the notation $\lambda^k = [\lambda_1^k, \cdots, \lambda_m^k]^T$, and the Cauchy-Schwarz inequality:

$$\frac{\left\|\sum_{i=1}^m \lambda_i^k g_i^k\right\|^2}{m^2} = \left\|(\lambda^k \otimes \mathbf{1})^T \nabla f(x^k)\right\|^2 \leq d\|\lambda^k\|^2 \|\nabla f(x^k)\|^2$$

We then add and subtract $\nabla f(x^*) = 0$ to obtain

$$\begin{aligned}\frac{\left\|\sum_{i=1}^m \lambda_i^k g_i^k\right\|^2}{m^2} &\leq 2d\|\lambda^k\|^2 \|\nabla f(x^k) - \nabla f(x^*)\|^2 \\ &\leq 2d\|\lambda^k\|^2 L^2 \|x^k - x^*\|^2\end{aligned} \tag{21}$$

where the last inequality follows by the Lipschitz continuity of ∇f. Further using the inequality

$$\begin{aligned}\|x^k - x^*\|^2 &\leq \|x^k - \mathbf{1} \otimes \bar{x}^k + \mathbf{1} \otimes \bar{x}^k - x^*\|^2 \\ &\leq 2\|x^k - \mathbf{1} \otimes \bar{x}^k\|^2 + 2\|\mathbf{1} \otimes \bar{x}^k - x^*\|^2 \\ &\leq 2\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 2m\|\bar{x}^k - \theta^*\|^2\end{aligned} \tag{22}$$

we have from (21)

$$\begin{aligned}\frac{\left\|\sum_{i=1}^m \lambda_i^k g_i^k\right\|^2}{m^2} \leq &4d\|\lambda^k\|^2 L^2 \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \\ &+ 4md\|\lambda^k\|^2 L^2 \|\bar{x}^k - \theta^*\|^2\end{aligned} \tag{23}$$

Substituting (20) and (23) into (13) yields

$$\begin{aligned}&\left\|\bar{x}^{k+1} - \theta^*\right\|^2 \leq \left\|\bar{x}^k - \theta^*\right\|^2 \\ &+ \frac{\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2}{m} + \frac{L^2 \|\lambda^k\|^2 \|\bar{x}^k - \theta^*\|^2}{m} + \\ &2L\sqrt{md}\|\lambda^k - \bar{\lambda}^k \mathbf{1}_m\| \|\bar{x}^k - \theta^*\|^2 - 2\bar{\lambda}^k (F(\bar{x}^k) - F(\theta^*) \\ &+ 4d\|\lambda^k\|^2 L^2 \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 4md\|\lambda^k\|^2 L^2 \|\bar{x}^k - \theta^*\|^2\end{aligned} \tag{24}$$

We can group the common terms on the right hand side of the preceding relation and obtain

$$\begin{aligned}&\left\|\bar{x}^{k+1} - \theta^*\right\|^2 \leq \left\|\bar{x}^k - \theta^*\right\|^2 \times \\ &\left(1 + \frac{L^2\|\lambda^k\|^2 + 2Lm\sqrt{md}\|\lambda^k - \bar{\lambda}^k \mathbf{1}\|}{m} + 4md\|\lambda^k\|^2 L^2\right) \\ &+ (\frac{1}{m} + 4d\|\lambda^k\|^2 L^2) \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 - 2\bar{\lambda}^k (F(\bar{x}^k) - F(\theta^*)\end{aligned} \tag{25}$$

Step II: Relationship for $\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$. For the convenience of analysis, we write PDG-DS on per-coordinate expressions. Define for all $\ell = 1, \ldots, d$, and $k \geq 0$,

$$x^k(\ell) = \left[[x_1^k]_\ell, \ldots, [x_m^k]_\ell\right]^T, \; g^k(\ell) \;\; = \left[[g_1^k]_\ell, \ldots, [g_m^k]_\ell\right]^T$$

In this per-coordinate view, (4) and (7) have the following form for all $\ell = 1, \ldots, d$, and $k \geq 0$:

$$\begin{aligned}x^{k+1}(\ell) &= W x^k(\ell) - B^k \Lambda^k g^k(\ell) \\ [\bar{x}^{k+1}]_\ell &= [\bar{x}^k]_\ell - \frac{1}{m} \mathbf{1}^T \Lambda^k g^k(\ell)\end{aligned} \tag{26}$$

From (26), we obtain

$$\begin{aligned}x^{k+1}(\ell) - [\bar{x}^{k+1}]_\ell \mathbf{1} =& W x^k(\ell) - [\bar{x}^k]_\ell \mathbf{1} \\ &- \left(B^k \Lambda^k g^k(\ell) - \frac{1}{m} \mathbf{1}^T \Lambda^k g^k(\ell) \mathbf{1}\right)\end{aligned}$$

Noting that $[\bar{x}^k]_\ell \mathbf{1} = \frac{1}{m} \mathbf{1}\mathbf{1}^T x^k(\ell)$ and $\frac{1}{m} \mathbf{1}^T \Lambda^k g^k(\ell) \mathbf{1} = \frac{1}{m} \mathbf{1}\mathbf{1}^T \Lambda^k g^k(\ell)$, we have

$$x^{k+1}(\ell) - [\bar{x}^{k+1}]_\ell \mathbf{1} = \bar{W} x^k(\ell) - \bar{B}^k \Lambda^k g^k(\ell)$$

where $\bar{W} = W - \frac{\mathbf{1}\mathbf{1}^T}{m}$ and $\bar{B}^k = B^k - \frac{\mathbf{1}\mathbf{1}^T}{m}$.

Noticing $\bar{W}[\bar{x}^k]_\ell \mathbf{1} = \left(W - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right)[\bar{x}^k]_\ell \mathbf{1} = 0$, by subtracting this expression from the right hand side of the preceding relation, we obtain

$$x^{k+1}(\ell) - [\bar{x}^{k+1}]_\ell \mathbf{1} = \bar{W}(x^k(\ell) - [\bar{x}^k]_\ell \mathbf{1}) - \bar{B}^k \Lambda^k g^k(\ell)$$

Taking norm on both sides and using $\eta = \|W - \frac{1}{m}\mathbf{1}\mathbf{1}^T\|$ from Assumption 1, we obtain

$$\|x^{k+1}(\ell) - [\bar{x}^{k+1}]_\ell \mathbf{1}\| \leq \eta \|x^k(\ell) - [\bar{x}^k]_\ell \mathbf{1}\| + \left\|\bar{B}^k\right\| \|\Lambda^k\| \, \|g^k(\ell)\| \tag{27}$$

The column stochastic property of B^k implies

$$\left\|\bar{B}^k\right\| \leq \left\|\bar{B}^k\right\|_F \leq m \tag{28}$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm, yielding

$$\|x^{k+1}(\ell) - [\bar{x}^{k+1}]_\ell \mathbf{1}\| \leq \eta \|x^k(\ell) - [\bar{x}^k]_\ell \mathbf{1}\| + m\|\Lambda^k\| \, \|g^k(\ell)\|$$

By taking squares on both sides and using the inequality $2ab \leq \epsilon a^2 + \epsilon^{-1} b^2$ valid for any a, b, and $\epsilon > 0$, we obtain

$$\begin{aligned}\|x^{k+1}(\ell) - [\bar{x}^{k+1}]_\ell \mathbf{1}\|^2 \leq &\eta^2(1+\epsilon)\|x^k(\ell) - [\bar{x}^k]_\ell \mathbf{1}\|^2 \\ &+ m^2(1+\epsilon^{-1})\|\Lambda^k\|^2 \|g^k(\ell)\|^2\end{aligned}$$

Summing these relations over $\ell = 1, \ldots, d$, and noting $\sum_{\ell=1}^d \|x^k(\ell) - [\bar{x}^k]_\ell \mathbf{1}\|^2 = \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$ and $\sum_{\ell=1}^d \|g^k(\ell)\|^2 = \sum_{i=1}^m \|g_i^k\|^2$, we obtain

$$\begin{aligned}\sum_{i=1}^m \|x_i^{k+1} - \bar{x}^{k+1}\|^2 \leq &\eta^2(1+\epsilon) \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \\ &+ m^2(1+\epsilon^{-1})\|\Lambda^k\|^2 \|g^k\|^2\end{aligned} \tag{29}$$

We next focus on estimating $\|g^k\|^2$. Noting that $g^k = m\nabla f(x^k)$, $\nabla f(x^*) = 0$, and f has Lipschitz continuous gradients, we have

$$\begin{aligned}\|g^k\|^2 &= m^2 \|\nabla f(x^k) - \nabla f(x^*)\|^2 \leq m^2 L^2 \|x^k - x^*\|^2 \\ &\leq 2m^2 L^2 \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 2m^3 L^2 \|\bar{x}^k - \theta^*\|^2\end{aligned} \tag{30}$$

where the last inequality used (22).

Substituting (30) into (29) and grouping terms yield

$$\sum_{i=1}^{m}\|x_i^{k+1}-\bar{x}^{k+1}\|^2 \le 2m^5L^2(1+\epsilon^{-1})\|\Lambda^k\|^2\|\bar{x}^k-\theta^*\|^2 + \left(\eta^2(1+\epsilon)+2m^4L^2(1+\epsilon^{-1})\|\Lambda^k\|^2\right)\sum_{i=1}^{m}\|x_i^k-\bar{x}^k\|^2$$

By letting $\epsilon=\frac{1-\eta}{\eta}$ with $\epsilon>0$, and noting $\eta\in(0,1)$, $1+\epsilon=\eta^{-1}$, $1+\epsilon^{-1}=(1-\eta)^{-1}$, and $\|\Lambda^k\|=\max_i\lambda_i^k\le\|\lambda^k\|$, we arrive at

$$\sum_{i=1}^{m}\|x_i^{k+1}-\bar{x}^{k+1}\|^2 \le 2m^5L^2(1-\eta)^{-1}\|\lambda^k\|^2\|\bar{x}^k-\theta^*\|^2 + \left(\eta+2m^4L^2(1-\eta)^{-1}\|\lambda^k\|^2\right)\sum_{i=1}^{m}\|x_i^k-\bar{x}^k\|^2 \tag{31}$$

Combining (25) and (31), we have

$$\mathbf{v}^{k+1}\le\left(\begin{bmatrix}1&\frac{1}{m}\\0&\eta\end{bmatrix}+A^k\right)\mathbf{v}^k-2\bar{\lambda}^k\begin{bmatrix}(F(\bar{x}^k)-F(\theta^*)\\0\end{bmatrix} \tag{32}$$

where $\mathbf{v}^k=\begin{bmatrix}\|\bar{x}^k-\theta^*\|^2\\\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2\end{bmatrix}$, $A^k=\begin{bmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{bmatrix}$ with $A_{11}\triangleq\frac{L^2\|\lambda^k\|^2+2Lm\sqrt{md}\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|}{m}+4md\|\lambda^k\|^2L^2$, $A_{12}=4d\|\lambda^k\|^2L^2$, $A_{21}=2m^5L^2(1-\eta)^{-1}\|\lambda^k\|^2$, and $A_{22}=2m^4L^2(1-\eta)^{-1}\|\lambda^k\|^2$.

Because $\begin{bmatrix}1&\frac{1}{m}\\0&\eta\end{bmatrix}\le\begin{bmatrix}1&1\\0&\eta\end{bmatrix}$ and $A^k\le a^k\mathbf{1}\mathbf{1}^T$ holds when a^k is set to $a^k=\max\{A_{11},A_{12},A_{21},A_{22}\}$, we can see that (8) in Proposition 1 is satisfied. Further note that under the conditions in the statement, all conditions for $\{a^k\}$, $\{b^k\}$, and $\{c^k\}$ in Proposition 1 are also satisfied (b^k is always 0 here). Therefore, we have the claimed results. ■

Remark 1. *To our knowledge, for decentralized gradient methods with diminishing stepsizes, our result is the first to prove exact convergence under general time-varying stepsize heterogeneity. In fact, the condition in (12) can be satisfied even when the stepsize differences in a finite number of iterations are arbitrarily large. This can enable strong privacy, as detailed in Sec. III-B.*

B. Privacy Analysis

Recall that in Sec. II we identify the gradients of agents as information to be protected in decentralized optimization. In this subsection, we will show that the PDG-DS algorithm can effectively protect the gradients of all participating agents from being inferable by honest-but-curious adversaries and external eavesdroppers. To this end, we first give a privacy metric and our definition of privacy protection.

We define the difference from x to x' as follows (in the log scale, could be positive or negative):

$$\zeta=\log\frac{\|x\|}{\|x'\|} \tag{33}$$

Definition 1. *For a network of m agents in decentralized optimization, the privacy of agent i is preserved if for any finite number of iterations T, its gradient values $g_i^1,\cdots,g_i^T$ always have alternative realizations $\hat{g}_i^1,\cdots,\hat{g}_i^T$ which allow each $\hat{g}_i^k$ ($1\le k\le T$) to have an arbitrarily large difference (could be different for different k) from g_i^k, but lead to the same shared information in inter-agent communications.*

The above privacy definition requires that when an agent's gradient is perturbed by an arbitrary value ζ, its shared information can still be the same, i.e., an alteration to an agent's gradient value is not distinguishable by an adversary having access to all information shared by the agent. Since the alteration in gradient can be arbitrarily large, our privacy definition requires that an adversary cannot even find a range for a protected value, and hence is more stringent than many existing privacy definitions (e.g., [11], [29]) that only require an adversary unable to *uniquely* determine a protected value.

Theorem 2. *In the presence of honest-but-curious or eavesdropping adversaries, PDG-DS can protect the privacy of all participating agents defined in Definition 1.*

Proof. Without loss of generality, we first consider the protection of the gradient of agent i at any single time instant k, and then show that the argument also applies to any finite number of time instants (iterations). When the gradient is g_i^k, we represent the information that agent i shares with neighboring agents when participating PDG-DS as $\mathcal{I}_i$. According to Definition 1, we have to prove that when the gradient is altered from g_i^k to $\hat{g}_i^k=e^{\zeta^k}g_i^k$ with ζ^k difference from g_i^k according to the metric in (33), the corresponding shared information $\hat{\mathcal{I}}_i$ of agent i could be identical to $\mathcal{I}_i$ under any $\zeta^k>0$.

According to Algorithm PDG-DS, agent i shares the following information in decentralized optimization:

$$\mathcal{I}_i=\mathcal{I}_i^{\text{sent}}\bigcup\mathcal{I}_i^{\text{public}}$$

with $\mathcal{I}_i^{\text{sent}}=\left\{v_{ji}^k\triangleq w_{ji}^kx_i^k-b_{ji}^k\lambda_i^kg_i^k|k=1,2,\cdots\right\}$ and $\mathcal{I}_i^{\text{public}}=\left\{W\bigcup\sum_{j\in\mathbb{N}_i}b_{ji}^k=1|k=0,1,\cdots\right\}$. One can obtain that at some iteration k, if the gradient is changed to $\hat{g}_i^k=e^{\zeta^k}g_i^k$, the difference defined in (33) is ζ^k. However, in this case, if we set the stepsize $\hat{\lambda}_i^k$ to $\hat{\lambda}_i^k=e^{-\zeta^k}\lambda_i^k$, then the corresponding shared information will still be v_{ji}^k. Since other parameters are not changed and changing the stepsize from λ_i^k to $\hat{\lambda}_i^k=e^{-\zeta^k}\lambda_i^k$ will not violate the summable stepsize heterogeneity condition in (12) for any given $\zeta^k<\infty$, according to Theorem 1, convergence to the optimal solution will still be guaranteed. Therefore, changes in an agent's gradient can be completely covered by the agent's flexibility in changing its stepsize, which does not affect the convergence. Thus, privacy of any agent's gradient will be protect when running PDG-DS. Given that the summable stepsize heterogeneity condition in (12) allows the stepsize of agent i to change by any finite amount for any finite number of iterations, one can obtain that the privacy of every agent's gradients in any number of iterations can be completely covered by the flexibility in changing the agent's stepsize in these iterations, as long as the number of these iterations is finite. It is worth noting that the perturbation does not violate the convexity and Lipschitz conditions in Assumption 2. This is because in order for an adversary to

check if Assumption 2 is violated, it has to know x_i^k and L, which, however, are not available to adversaries: before convergence, x_i^k is inaccessible to the adversary because the information shared by agent i is $w_{ji}x_i^k - b_{ji}^k\lambda_i^k g_i^k$, avoiding x_i^k from being inferable; L is inaccessible to the adversary either because Assumption 2 only requires all f_i to have finite Lipschitz constants, and agents do not share their Lipschitz constants in the implementation of the algorithm. In fact, even with the gradient g_i^k unchanged, the value of observation $w_{ji}^k x_i^k - b_{ji}^k\lambda_i^k g_i^k$ in Algorithm 1 can be changed by an arbitrary finite value by changing the stepsize λ_i^k. Therefore, before convergence, an adversary cannot use Assumption 2 to confine the change in observed values and further confine the change in the value of g_i^k. After convergence, the perturbation does not violate the convexity and Lipschitz conditions, either. In fact, although x_i^k becomes accessible to the adversary after convergence, gradient is eliminated in adversary's observation (the shared $w_{ji}x_i^k - b_{ji}^k\lambda_i^k g_i^k$ becomes $w_{ji}x_i^k$) because λ_i^k converges to zero. So after convergence, the adversary still cannot use Assumption 2 to confine changes in gradients. ■

Remark 2. *Even after convergence when g_i^k becomes a constant, an adversary still cannot infer gradients from shared messages in PDG-DS. More specifically, when g_i^k converges to a constant value, the stepsize λ_i^k also converges to zero, which completely eliminates the information of g_i^k in observed information (the observed information becomes $w_{ji}x_i^k$ after convergence). This can also be understood intuitively as follows: Even if the adversary can collect $T \to \infty$ observations $w_{ji}x_i^k - b_{ji}^k\lambda_i^k g_i^k$ in the neighborhood of the optimal point and establish a system of T equations to solve for g_i^k (which can be viewed to be approximately time-invariant in the neighborhood of the optimal point), the number of unknowns b_{ji}^k, λ_i^k, and g_i^k in the system of T equations is $3T$ (even if we view λ_i^k and g_i^k approximately as constants in the neighborhood of the optimal point, the number of unknowns is still $T+2$), which makes it impossible for the adversary to solve for g_i^k using the system of T equations established from observations.*

Remark 3. *Different from existing privacy solutions for decentralized optimization that patch a privacy mechanism (e.g., differential-privacy noise or encryption) with a pre-designed decentralized optimization algorithm, our proposed algorithm uses stepsize and coupling coefficients that are inherent to the decentralized optimization algorithm to perturb gradients, and hence has inherent privacy.*

Remark 4. *Existing accuracy-maintaining privacy approaches for decentralized optimization can only protect the privacy of participating agents when the interaction topology meets certain conditions. For example, the approach in [29] assumes that an adversary cannot have access to messages sent on at least one communication channel of an agent to guarantee the privacy of this agent. The approach in [26] requires that an adversary cannot be the only neighbor of a target agent. To the contrary, our PDG-DS can protect the privacy of an agent without any constraint on the interaction topology. In fact, to our knowledge, our algorithm is the first decentralized gradient based algorithm that can guarantee both optimization accuracy and privacy defined in Definition 1 when an adversary has access to all shared information.*

IV. An inherently privacy-preserving decentralized gradient algorithm with non-diminishing stepsizes

Because diminishing stepsizes in decentralized gradient methods may slow down convergence, plenty of efforts have been devoted to developing algorithms that can achieve accurate optimization results under a non-diminishing stepsize. Typical examples include gradient-tracking based algorithms such as Aug-DGM [8], DIGing [14], AsynDGM [15], AB [9], Push-Pull [16], [17], and ADD-OPT [18], etc. However, these algorithms will lead to privacy breaches in implementation. For example, DIGing [14] implements the following update rule (note that under our assumption, $i \in \mathbb{N}_i$):

$$\begin{cases} x_i^{k+1} = \sum_{j\in\mathbb{N}_i} w_{ij}x_j^k - \lambda y_i^k \\ y_i^{k+1} = \sum_{j\in\mathbb{N}_i} w_{ij}(y_j^k + g_j^{k+1} - g_j^k) \end{cases}$$

At iteration $k=0$, agent j sets $y_j^0 = g_j^0$ and sends $w_{ij}(y_i^0 + g_j^1 - g_j^0) = w_{ij}g_j^1$ to its neighboring agent i. At iteration $k=1$, agent j further sends x_j^1 to agent i. Given that w_{ij}s are publicly known, agent i can easily determine the gradient of agent j at x_j^1. Using a similar argument, we can see that other commonly used gradient-tracking based algorithms also have the same issue of leaking agents' gradient information, even when the stepsizes are heterogeneous (see Sec. IV.B for details).

The EXTRA algorithm [7] can also ensure convergence to the exact optimal solution under non-diminishing stepsizes:

$$x_i^{k+2} = \sum_{j\in\mathbb{N}_i} w_{1,ij}x_j^{k+1} - \sum_{j\in\mathbb{N}_i} w_{2,ij}x_j^k - \lambda(g_i^{k+1} - g_i^k)$$

However, since $\lambda, w_{1,ij}, w_{2,ij}$ are publicly known, and an agent i has to share x_i^k directly, one can see that the gradient information of participating agent i will also be disclosed.

Motivated by the observation that the main sources of information leakage in decentralized optimization are constant parameters and the sharing of two messages by every agent in every iteration, we propose the following inherently privacy-preserving decentralized gradient based algorithm which can protect the gradients of participating agents while ensuring convergence to the exact optimal solution under non-diminishing stepsizes (the per-agent version is given in Algorithm PDG-NDS):

$$\begin{aligned} x^{k+2} &= 2(W\otimes I_d)x^{k+1} - (W^2\otimes I_d)x^k \\ &\quad - \left(\left((B^k\Lambda^{k+1})\otimes I_d\right)g^{k+1} - \left((B^k\Lambda^k)\otimes I_d\right)g^k\right) \end{aligned} \tag{34}$$

where $B^k = \{b_{ij}^k\} \in \mathbb{R}^{m\times m}$ is a column-stochastic matrix, $\Lambda^k = \mathrm{diag}[\lambda_1^k, \lambda_2^k, \cdots, \lambda_m^k]$ with $\lambda_i^k \geq 0$ denoting the stepsize of agent i at iteration k, $g^k = [(g_1^k)^T, (g_2^k)^T, \cdots (g_m^k)^T]^T$, and $\otimes$ denotes Kronecker product.

Remark 5. *PDG-NDS requires one agent to share only one variable with every neighboring agent at each iteration. This is different from all existing gradient-tracking based algorithms which have to exchange two variables between two neighboring agents in every iteration (one optimization variable*

and one auxiliary variable tracking the gradient of the global objective function). This difference is key to 1) reduce communication overhead; 2) enable privacy because exchanging the additional gradient-tracking variable will disclose gradient information, as detailed in Sec. IV.B.

PDG-NDS: Privacy-preserving decentralized gradient method with non-diminishing stepsizes

Public parameters: W

Private parameters for agent i: $b_{ji}^k \geq 0$, $\lambda_i^k \geq 0$, and x_i^0

1) At iteration $k = 1$: Agent i shares x_i^0 (randomly selected) with neighbors and updates its state as follows

$$x_i^1 = \sum\nolimits_{j\in\mathbb{N}^i} w_{ij}x_j^0 - \lambda_i^0\nabla f_i(x_i^0)$$

2) **for** $k = 2, 3, \cdots$ **do**

a) Every agent j computes and sends v_{ij}^k (defined in (35)) to all agents $i \in \mathbb{N}_j$ where $\{W^2\}_{ij}$ denotes the (i,j)th entry of matrix W^2:

$$\begin{aligned} v_{ij}^k = {} & 2w_{ij}x_j^{k-1} + \{W^2\}_{ij}x_j^{k-2} \\ & - b_{ij}^{k-2}(\lambda_j^{k-1}g_j^{k-1} - \lambda_j^{k-2}g_j^{k-2}) \end{aligned} \tag{35}$$

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

$$x_i^k = \sum\nolimits_{j\in\mathbb{N}_i} v_{ij}^k \tag{36}$$

c) **end**

A. Convergence analysis

We define an auxiliary variable

$$y^k \triangleq (W \otimes I_d)x^k - x^{k+1} \tag{37}$$

It can be verified that

$$\begin{aligned} y^{k+1} = {} & (W \otimes I_d)y^k + \\ & \left(\left((B^k\Lambda^{k+1}) \otimes I_d\right) g^{k+1} - \left((B^k\Lambda^k) \otimes I_d\right) g^k\right) \end{aligned} \tag{38}$$

Define mean vectors of x_i^k and y_i^k as $\bar{x}^k = \frac{1}{m}\sum_{i=1}^m x_i^k$ and $\bar{y}^k = \frac{1}{m}\sum_{i=1}^m y_i^k$, respectively. Then from (37), we have

$$\bar{x}^{k+1} = \bar{x}^k - \bar{y}^k \tag{39}$$

Further using (38) and the initialization condition $x^1 = Wx^0 - \Lambda^0 g^0$ in PDG-NDS, we have $\bar{y}^0 = \frac{1}{m}\sum_{i=1}^m \lambda_i^0 g_i^0$ and

$$\bar{y}^k = \tfrac{1}{m}\textstyle\sum_{i=1}^m \lambda_i^k g_i^k \tag{40}$$

To prove convergence of our algorithm, we first present two lemmas and one proposition. The proposition applies to general distributed algorithms for solving optimization problem (1).

Lemma 1. *Let $\{\mathbf{v}^k\} \subset \mathbb{R}^d$ and $\{\mathbf{u}^k\} \subset \mathbb{R}^p$ be sequences of non-negative vectors such that*

$$\mathbf{v}^{k+1} \leq (V^k + a^k\mathbf{1}\mathbf{1}^T)\mathbf{v}^k + b^k\mathbf{1} - C^k\mathbf{u}^k, \forall k \geq 0 \tag{41}$$

where $\{V^k\}$ is a sequence of non-negative matrices, and $\{a^k\}$ and $\{b^k\}$ are non-negative scalar sequences satisfying $\sum_{k=0}^\infty a^k < \infty$ and $\sum_{k=0}^\infty b^k < \infty$. Assume that there exists a vector $\pi > 0$ such that $\pi^T V^k \leq \pi^T$ and $\pi^T C^k \geq 0$ hold for all $k \geq 0$. Then, $\lim_{k\to\infty}\pi^T\mathbf{v}^k$ exists, the sequence $\{\mathbf{v}^k\}$ is bounded, and $\sum_{k=1}^\infty \pi^T C^k\mathbf{u}^k < \infty$.

Proof. By multiplying (41) with π^T and using the assumptions $\pi^T V^k \leq \pi^T$ and $\mathbf{v}^k \geq 0$, we obtain for $\forall k \geq 0$

$$\pi^T\mathbf{v}^{k+1} \leq \pi^T\mathbf{v}^k + a^k(\pi^T\mathbf{1})(\mathbf{1}^T\mathbf{v}^k) + b^k\pi^T\mathbf{1} - \pi^T C^k\mathbf{u}^k$$

Since $\pi > 0$, we have $\pi_{\min} = \min_i \pi_i > 0$, and hence $\mathbf{1}^T\mathbf{v}^k = \frac{1}{\pi_{\min}}\pi_{\min}\mathbf{1}^T\mathbf{v}^k \leq \frac{1}{\pi_{\min}}\pi^T\mathbf{v}^k$, where the inequality holds since $\mathbf{v}^k \geq 0$. Therefore,

$$\pi^T\mathbf{v}^{k+1} \leq \left(1 + a^k\frac{\pi^T\mathbf{1}}{\pi_{\min}}\right)\pi^T\mathbf{v}^k + b^k\pi^T\mathbf{1} - \pi^T C^k\mathbf{u}^k, \forall k \geq 0 \tag{42}$$

By our assumption, $\pi^T C^k\mathbf{u}^k \geq 0$ for all k, so (42) implies that the conditions of Lemma 3 in the Appendix are satisfied with $v^k = \pi^T\mathbf{v}^k$, $\alpha^k = a^k\pi^T\mathbf{1}/\pi_{\min}$, and $\beta^k = b^k\pi^T\mathbf{1}$. Thus, by Lemma 3, it follows that $\lim_{k\to\infty}\pi^T v^k$ exists. Consequently, $\{\pi^T v^k\}$ is bounded, and under $\pi > 0$, implying that $\{\mathbf{v}^k\}$ is also bounded. Moreover, by summing the relations in (42), we find $\sum_{k=1}^\infty \pi^T C^k\mathbf{u}^k < \infty$. ∎

Lemma 2. *Let $\{\mathbf{v}^k\} \subset \mathbb{R}^d$ be a sequence of non-negative vectors such that for $\forall k \geq 0$*

$$\mathbf{v}^{k+1} \leq V^k\mathbf{v}^k + b^k\mathbf{1} \tag{43}$$

where $\{V^k\}$ is a sequence of non-negative matrices. Assume that there exist a vector $\pi > 0$ and a scalar sequence $\{\alpha^k\}$ such that $\alpha^k \in (0,1)$, $\sum_{k=0}^\infty \alpha^k = \infty$, $\lim_{k\to\infty} b^k/\alpha^k = 0$, and $\pi^T V^k \leq (1-\alpha^k)\pi^T$ for all $k \geq 0$. Then, $\lim_{k\to\infty}\mathbf{v}^k = 0$.

Proof. We use Lemma 4 in the Appendix to establish the result. By multiplying (43) with π^T and using the assumptions $\pi^T V^k \leq (1-\alpha^k)\pi^T$ and $\mathbf{v}^k \geq 0$, we obtain

$$\pi^T\mathbf{v}^{k+1} \leq (1-\alpha^k)\pi^T\mathbf{v}^k + b^k\pi^T\mathbf{1}, \forall k \geq 0$$

Since $\alpha^k \in (0,1)$, $\sum_{k=0}^\infty \alpha^k = \infty$, $\lim_{k\to\infty} b^k/\alpha^k = 0$, the conditions of Lemma 4 in the Appendix are satisfied with $v^k = \pi^T\mathbf{v}^k$ and $\beta^k = b^k\pi^T\mathbf{1}$. Thus, it follows $\lim_{k\to\infty}\pi^T\mathbf{v}^k = 0$ and further (because $\pi > 0$) $\lim_{k\to\infty}\mathbf{v}^k = 0$. ∎

Proposition 2. *Assume that problem (1) has an optimal solution and that $F(\cdot)$ in (1) is continuously differentiable. Suppose that a distributed algorithm generates sequences $\{x_i^k\} \subseteq \mathbb{R}^d$ and $\{y_i^k\} \subseteq \mathbb{R}^d$ such that the following relation is satisfied for any optimal solution θ^* and for all $k \geq 0$,*

$$\mathbf{v}^{k+1} \leq \left(V + a^k\mathbf{1}\mathbf{1}^T\right)\mathbf{v}^k + b^k\mathbf{1} - C\begin{bmatrix} \|\nabla F(\bar{x}^k)\|^2 \\ \|\bar{y}^k\|^2 \end{bmatrix} \tag{44}$$

where $\nu > 0$,

$$\mathbf{v}^k \triangleq \begin{bmatrix} \nu(F(\bar{x}^k) - F(\theta^*)) \\ \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \\ \sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2 \end{bmatrix}, C = \begin{bmatrix} \gamma^k & \frac{1-\tau^k}{\tau^k} \\ 0 & 0 \\ 0 & -(1-\eta)^2(1-c) \end{bmatrix}$$

$$V = \begin{bmatrix} 1 & 1-\eta & 0 \\ 0 & \eta & \frac{1}{1-\eta} \\ 0 & (1-\eta)^2(1-c)(1-\delta) & c \end{bmatrix}$$

with $\eta, c, \delta \in (0,1)$, while the scalar sequences $\{a^k\}$, $\{b^k\}$, $\{\tau^k\}$, $\{\gamma^k\}$ are nonnegative satisfying $\tau^k \in (0,1)$, $\frac{1-\tau^k}{\tau^k}\delta \geq 1$ for all $k \geq 0$, and $\sum_{k=0}^{\infty} a^k < \infty$, $\sum_{k=0}^{\infty} b^k < \infty$. Then, we have:

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

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

(b) If $\{\gamma^k\}$ satisfies $\sum_{k=0}^{\infty} \gamma^k = \infty$ and $\lim_{k\to\infty} \gamma^k > 0$, then $\lim_{k\to\infty} \|\nabla F(\bar{x}^k)\| = 0$. Moreover, if $\{\bar{x}^k\}$ is bounded, then every accumulation point of $\{\bar{x}^k\}$ is an optimal solution, and $\lim_{k\to\infty} F(x_i^k) = F(\theta^*)$ for all i.

Proof. (a) The idea is to show that Lemma 1 applies. Setting up the equation $\pi^T V = \pi^T$, we have $(1-\eta)\pi_1 + (1-\eta)^2(1-c)(1-\delta)\pi_3 = (1-\eta)\pi_2$ and $\pi_2 = (1-\eta)(1-c)\pi_3$. Dividing the first equation with $1-\eta$, we find

$$\pi_1 + (1-\eta)(1-c)(1-\delta)\pi_3 = \pi_2$$

which in view of $\pi_2 = (1-\eta)(1-c)\pi_3$ implies $\pi_1 + (1-\delta)\pi_2 = \pi_2$, and hence $\pi_1 = \delta\pi_2$.

Thus, for the vector π satisfying $\pi^T = V\pi^T$, we have

$$\pi_1 = \delta\pi_2, \qquad \pi_2 = (1-\eta)(1-c)\pi_3 \tag{45}$$

Hence, we can find such a vector π with $\pi > 0$. We next verify that such a vector also satisfies $\pi^T C > 0$. We have $\pi^T C = [\gamma^k\pi_1, \frac{1-\tau^k}{\tau^k}\pi_1 - (1-\eta)^2(1-c)\pi_3]$, which, under (45), implies the second coordinate of $\pi^T C$ satisfying $[\pi^T C]_2 = \frac{1-\tau^k}{\tau^k}\delta\pi_2 - (1-\eta)\pi_2 = \left(\frac{1-\tau^k}{\tau^k}\delta - 1 + \eta\right)\pi_2$.

The condition $\frac{1-\tau^k}{\tau^k}\delta \geq 1$ implies $[\pi^T C]_2 \geq \eta\pi_2 > 0$. Thus, Lemma 1's conditions are satisfied, and it follows that for the three elements of $\mathbf{v}^k$, i.e., $\mathbf{v}_1^k$, $\mathbf{v}_2^k$, and $\mathbf{v}_3^k$, we have that

$$\lim_{k\to\infty} \pi_1\mathbf{v}_1^k + \pi_2\mathbf{v}_2^k + \pi_3\mathbf{v}_3^k \tag{46}$$

exists and $\sum_{k=0}^{\infty} \pi^T C\mathbf{u}^k < \infty$ holds with $\mathbf{u}^k = [\|\nabla F(\bar{x}^k)\|^2, \ \|\bar{y}^k\|^2]^T$. Since $\pi^T C \geq [\gamma^k\pi_1, \ \eta\pi_2]$, one has

$$\sum_{k=0}^{\infty} \gamma^k \|\nabla F(\bar{x}^k)\|^2 < \infty, \qquad \sum_{k=0}^{\infty} \|\bar{y}^k\|^2 < \infty \tag{47}$$

and hence,

$$\lim_{k\to\infty} \|\bar{y}^k\| = 0 \tag{48}$$

If we had that $\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$ and $\sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2$ are convergent, then it would follow from (46) that the limit $\lim_{k\to\infty} F(\bar{x}^k)$ exists.

Now, we focus on proving that both $\mathbf{v}_2^k = \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$ and $\mathbf{v}_3^k = \sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2$ converge to 0. The idea is to show that we can apply Lemma 2. By focusing on the elements $\mathbf{v}_2^k$ and $\mathbf{v}_3^k$, from (44) we have

$$\begin{bmatrix} \mathbf{v}_2^{k+1} \\ \mathbf{v}_3^{k+1} \end{bmatrix} \leq \left(\tilde{V} + a^k \mathbf{1}\mathbf{1}^T\right) \begin{bmatrix} \mathbf{v}_2^k \\ \mathbf{v}_3^k \end{bmatrix} + \hat{b}^k \mathbf{1} + \begin{bmatrix} 0 \\ \hat{c}^k \end{bmatrix}$$

where $\hat{b}^k = b^k + a^k\nu(F(\bar{x}^k) - F(\theta^*))$, $\hat{c}^k = (1-\eta)^2(1-c)\|\bar{y}^k\|^2$, and $\tilde{V} = \begin{bmatrix} \eta & \frac{1}{1-\eta} \\ (1-\eta)^2(1-c)(1-\delta) & c \end{bmatrix}$.

By separating the first term on the right hand side and bounding the last vector by $(1-\eta)^2(1-c)\|\bar{y}^k\|^2\mathbf{1}$, we obtain

$$\begin{bmatrix} \mathbf{v}_2^{k+1} \\ \mathbf{v}_3^{k+1} \end{bmatrix} \leq \tilde{V} \begin{bmatrix} \mathbf{v}_2^k \\ \mathbf{v}_3^k \end{bmatrix} + \tilde{b}^k \mathbf{1} \tag{49}$$

where

$$\tilde{b}^k = b^k + (1-\eta)^2(1-c)\|\bar{y}^k\|^2 + a^k \times \left(\nu(F(\bar{x}^k) - F(\theta^*)) + \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + \sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2\right) \tag{50}$$

To apply Lemma 2, we show that the equation $\pi^T\tilde{V} = (1-\alpha)\pi^T$ has a solution in $\pi = [\pi_2, \pi_3]^T$ with $[\pi_2, \pi_3] > 0$ and $\alpha \in (0,1)$. Note that, if we have such a solution, then we will let $\alpha^k = \alpha > 0$ for all k, so that the condition $\sum_{k=0}^{\infty} \alpha^k = \infty$ of Lemma 2 will be satisfied. In this case, the condition $\lim_{k\to\infty} \tilde{b}^k/\alpha^k = 0$ of Lemma 2 will also be satisfied. This is because by our assumption on the sequences $\{a^k\}$ and $\{b^k\}$, it follows that $\lim_{k\to\infty} a^k = 0$ and $\lim_{k\to\infty} b^k = 0$. We also have $\lim_{k\to\infty} \|\bar{y}^k\| = 0$ (see (48)). Moreover, in view of relation (46), the sequences $\{F(\bar{x}^k) - F(\theta^*)\}$, $\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2$, and $\sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2$ are bounded. Hence, it follows that $\tilde{b}^k$ defined in (50) will converge to 0 as k tends to infinity. Thus, all the conditions of Lemma 2 will be satisfied.

It remains to show that the system of equations $\pi^T\tilde{V} = (1-\alpha)\pi^T$ has a solution in $\pi = [\pi_2, \pi_3]^T$ with $[\pi_2, \pi_3] > 0$ and $\alpha \in (0,1)$. The system is equivalent to

$$(1-\eta)^2(1-c)(1-\delta)\pi_3 = (1-\eta-\alpha)\pi_2, \ \pi_2 = (1-\eta)(1-c-\alpha)\pi_3$$

which gives $\pi_2 > 0$ with arbitrary $\pi_3 > 0$, and imposes that α satisfies $(1-\eta)(1-c)(1-\delta) = (1-\eta-\alpha)(1-c-\alpha)$, or equivalently,

$$\alpha^2 - (2-\eta-c)\alpha + \delta(1-\eta)(1-c) = 0 \tag{51}$$

Letting $\psi(\alpha) = \alpha^2 - (2-\eta-c)\alpha + \delta(1-\eta)(1-c)$ for all $\alpha \in \mathbb{R}$, we note that $\psi(\cdot)$ is strongly convex and its minimum is attained at $\alpha_0 = \frac{1}{2}(2-\eta-c)$.

For the minimum value we have

$$\begin{aligned} \psi(\alpha_0) &= \tfrac{1}{4}(2-\eta-c)^2 - \tfrac{1}{2}(2-\eta-c)^2 + \delta(1-\eta)(1-c) \\ &= -\tfrac{1}{4}(2-\eta-c)^2 + \delta(1-\eta)(1-c) \\ &= -\tfrac{1}{4}(1-\eta+1-c)^2 + \delta(1-\eta)(1-c) \end{aligned}$$

Since $\delta < 1$, it follows that $\psi(\alpha_0) < -\frac{(1-\eta+1-c)^2}{4} + (1-\eta)(1-c) = -\frac{((1-\eta)-(1-c))^2}{4} \leq 0$. We also have $\psi(0) = \delta(1-\eta)(1-c) > 0$ since $\delta > 0$ and $c, \eta \in (0,1)$. Thus, we have $\psi(0) > 0$ and $\psi(\alpha_0) < 0$, implying that there exists some $\alpha^* \in (0, \alpha_0)$ satisfying $\psi(\alpha^*) = 0$ with $\alpha_0 = \frac{2-\eta-c}{2}$. Since $c, \eta \in (0,1)$, we have $\alpha_0 \in (0,1)$. Hence, (51) has a solution $\alpha^* \in (0,1)$. So there is a vector $\pi > 0$ and $\alpha \in (0,1)$ that satisfy $\pi^T\tilde{V} = (1-\alpha)\pi^T$, and we can apply Lemma 2 with $\alpha_k = \alpha$ for all k. By Lemma 2, we have $\lim_{k\to\infty} \|x_i^k - \bar{x}^k\| = 0$ and $\lim_{k\to\infty} \|y_i^k - \bar{y}^k\| = 0$.

(b) Since $\sum_{k=0}^{\infty} \gamma^k \|\nabla F(\bar{x}^k)\|^2 < \infty$ (see (47)), from $\sum_{k=0}^{\infty} \gamma^k = \infty$ and $\lim_{k\to\infty} \gamma^k > 0$, it follows $\lim_{k\to\infty} \|\nabla F(\bar{x}^k)\| = 0$.

Now, if $\{\bar{x}^k\}$ is bounded, then it has accumulation points. Let $\{\bar{x}^{k_i}\}$ be a sub-sequence such that $\lim_{i\to\infty} \|\nabla F(\bar{x}^{k_i})\| = 0$. Without loss of generality, we may assume that $\{\bar{x}^{k_i}\}$ is 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 continuity of the gradient $\nabla F(\cdot)$, it follows $\nabla F(\hat{x})=0$, implying that $\hat{x}$ is an optimal point. Since F is continuous, it follows $\lim_{i\to\infty}F(\bar{x}^{k_i})=F(\hat{x})=F(\theta^*)$. By part (a), $\lim_{k\to\infty}F(\bar{x}^k)$ exists, so we must have $\lim_{k\to\infty}F(\bar{x}^k)=F(\theta^*)$.

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

Theorem 3. *Under Assumption 1 and Assumption 2, if there exists some $T\geq 0$ such that for all $k\geq T$, the stepsize vector $\lambda^k=[\lambda_1^k,\cdots,\lambda_m^k]^T$ (with all elements non-negative) satisfies*

$$\sum_{k=T}^{\infty}\bar{\lambda}^k=\infty,\sum_{k=T}^{\infty}\|\lambda^{k+1}-\lambda^k\|^2<\infty,\sum_{k=T}^{\infty}\frac{\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2}{\bar{\lambda}^k}<\infty$$

with $\bar{\lambda}^k=\frac{\sum_{i=1}^m\lambda_i^k}{m}$, and

$$\frac{2L}{m\bar{\lambda}^k}(\lambda_{\max}^k)^2\leq 1-\eta,\bar{\lambda}^k\leq\frac{\delta}{1+\delta},\eta+\frac{6m^2L^2}{1-\eta}\|\lambda^{k+1}\|^2\leq c,$$
$$\max\{m^3r^2,m^2\}6L^2\|\lambda^{k+1}\|^2\leq(1-\eta)^3(1-c)(1-\delta)$$

for some $\delta\in(0,1)$, $c\in(0,1)$, then, the results of Proposition 2 hold for the proposed PDG-NDS.

Proof. The idea is to prove that we can establish the relationship in (44). To this end, we divide the derivation into four steps: in Step I, Step II, and Step III, we establish relationships for $\frac{2}{L}\big(F(\bar{x}^{k+1})-F(\theta^*)\big)$, $\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|^2$, and $\sum_{i=1}^m\|y_i^{k+1}-\bar{y}^{k+1}\|^2$, respectively, and in Step IV, we prove that (44) holds. To help the exposition of the main idea, we put Step I, Step II, and Step III in Appendix B, and only give the derivation of Step IV here.

Step IV: We summarize the relationships obtained in Steps I-III in Appendix B and prove the theorem. Defining $\mathbf{v}^k=\big[\frac{2}{L}(F(\bar{x}^{k+1})-F(\theta^*)),\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|^2,\sum_{i=1}^m\|y_i^{k+1}-\bar{y}^{k+1}\|^2\big]^T$, we have the following relations from (65), (68), and (75) in Appendix B:

$$\mathbf{v}^{k+1}\leq(V^k+A^k)\mathbf{v}^k-C^k\begin{bmatrix}\|\nabla F(\bar{x}^k)\|^2\\ \|\bar{y}^k\|^2\end{bmatrix}+B^k \quad (52)$$

where

$$V^k=\begin{bmatrix}1 & \frac{2L}{m\bar{\lambda}^k}(\lambda_{\max}^k)^2 & 0\\ 0 & \eta & \frac{1}{1-\eta}\\ 0 & \frac{6m^2r^2L^2}{1-\eta}\|\lambda^{k+1}\|^2 & \eta+\frac{6m^2L^2}{1-\eta}\|\lambda^{k+1}\|^2\end{bmatrix},$$

$$A^k=\begin{bmatrix}\frac{c_1}{\bar{\lambda}^k}\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2 & 0 & 0\\ 0 & 0 & 0\\ \frac{8m^3L^2}{1-\eta}\|\lambda^{k+1}-\lambda^k\|^2 & \frac{4m^2L^2}{1-\eta}\|\lambda^{k+1}-\lambda^k\|^2 & 0\end{bmatrix},$$

$$C^k=\begin{bmatrix}\frac{\bar{\lambda}^k}{L} & \frac{1-\bar{\lambda}^kL}{\bar{\lambda}^kL}\\ 0 & 0\\ 0 & -\frac{6m^3r^2L^2}{1-\eta}\|\lambda^{k+1}\|^2\end{bmatrix},$$

$$B^k=\begin{bmatrix}\frac{c_1}{\bar{\lambda}^k}\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2\\ 0\\ \frac{8m^2}{1-\eta}\|\lambda^{k+1}-\lambda^k\|^2\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2\end{bmatrix}$$

Now using the conditions of the theorem, we bound the entries in V^k, A^k, C^k, and B^k. It can be seen that

$$A^k\leq a^k\mathbf{1}\mathbf{1}^T,\quad B^k\leq b^k\mathbf{1} \quad (53)$$

hold where

$$a^k=\max\left\{\frac{c_1}{\bar{\lambda}^k}\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2,\frac{8m^3L^2}{1-\eta}\|\lambda^{k+1}-\lambda^k\|^2\right\} \quad (54)$$

$$b^k=\max\left\{\frac{c_1}{\bar{\lambda}^k}\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2,\frac{8m^2}{1-\eta}\|\lambda^{k+1}-\lambda^k\|^2\right\}\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2 \quad (55)$$

Using $\eta+\frac{6m^2L^2}{1-\eta}\|\lambda^{k+1}\|^2\leq c$ with $c\in(0,1)$, $\frac{2L}{m\bar{\lambda}^k}(\lambda_{\max}^k)^2\leq 1-\eta$, and $m^2r^26L^2\|\lambda^{k+1}\|^2\leq m^3r^26L^2\|\lambda^{k+1}\|^2\leq(1-\eta)^3(1-c)(1-\delta)$ from the theorem conditions, we can bound V^k:

$$V^k\leq V\triangleq\begin{bmatrix}1 & 1-\eta & 0\\ 0 & \eta & \frac{1}{1-\eta}\\ 0 & (1-\eta)^2(1-c)(1-\delta) & c\end{bmatrix} \quad (56)$$

Furthermore, we can bound C^k using the condition $\max\{m^3r^2,m^2\}6L^2\|\lambda^{k+1}\|^2\leq(1-\eta)^3(1-c)(1-\delta)$ which implies $\frac{m^3r^26L^2}{1-\eta}\|\lambda^{k+1}\|^2\leq(1-\eta)^2(1-c)$:

$$C^k\geq C\triangleq\begin{bmatrix}\frac{\bar{\lambda}^k}{L} & \frac{1-\bar{\lambda}^kL}{\bar{\lambda}^kL}\\ 0 & 0\\ 0 & -(1-\eta)^2(1-c)\end{bmatrix} \quad (57)$$

Combining (52), (53), (56), and (57) leads to

$$\mathbf{v}^{k+1}=(V+a^k\mathbf{1}\mathbf{1}^T)\mathbf{v}^k-C\begin{bmatrix}\|\nabla F(\bar{x}^k)\|^2\\ \|\bar{y}^k\|^2\end{bmatrix}+b^k\mathbf{1} \quad (58)$$

We note that Proposition 2 applies to the relations in (58) for $k\geq T$, with $\gamma^k=\frac{\bar{\lambda}^k}{L}$, $\tau^k=\bar{\lambda}^kL$, and a^k and b^k given by (54) and (55), respectively. By our assumption $\sum_{k=T}^{\infty}\|\lambda^{k+1}-\lambda^k\|^2<\infty$ and $\sum_{k=T}^{\infty}\frac{\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2}{\bar{\lambda}^k}<\infty$, it follows that $\{a^k\}$ and $\{b^k\}$ are nonegative and summable. The condition $\bar{\lambda}^kL\leq\delta/(1+\delta)$ is equivalent to $\bar{\lambda}^kL+\delta\bar{\lambda}^kL\leq\delta$, implying $1\leq\frac{\delta(1-\bar{\lambda}^kL)}{\bar{\lambda}^kL}$. Thus, with $\tau^k=\bar{\lambda}^kL$, we see that the condition $\frac{1-\tau^k}{\tau^k}\delta\geq 1$ of Proposition 2 is satisfied for all $k\geq T$. Additionally, by our assumption $\sum_{k=T}^{\infty}\bar{\lambda}^k=\infty$ we see that the condition of Proposition 2(b) also holds for $k\geq T$. Since the results of Proposition 2 are asymptotic, the results remain valid when the starting index is shifted from $k=0$ to $k=T$, for an arbitrary $T\geq 0$. ■

Remark 6. *In implementations, to satisfy the condition in the statement of Theorem 3, all agents can be given the same baseline value of stepsize λ. Then, every agent can set its λ_i^k by deviating from the baseline value in a finite number of iterations, the indices of which are private to agent i. As long as the deviation in each of these iterations is finite, the heterogeneity condition of Theorem 3 will be satisfied.*

Remark 7. *It is worth noting that although some gradient-tracking based algorithms can be reduced to the x-variable only form by eliminating the auxiliary variable, such a conversion is infeasible when the stepsizes are heterogeneous and not shared across agents (for the purpose of, e.g., privacy*

preservation). For example, the Aug-DGM algorithm in [8] has the following form

$$\begin{cases} x^{k+1} = W(x^k - \Lambda y^k) \\ y^{k+1} = W(y^k + g^{k+1} - g^k) \end{cases}$$

Although we can eliminate the y variable and convert it to

$$x^{k+2} = (W + W\Lambda W\Lambda^{-1}W^{-1})x^{k+1} - W\Lambda W\Lambda^{-1}x^k - W\Lambda W(g^{k+1} - g^k)$$

we cannot let agent j share $(W + W\Lambda W\Lambda^{-1}W^{-1})_{ij}x_j^{k+1} - (W\Lambda W\Lambda^{-1})_{ij}x_j^k - (W\Lambda W)_{ij}(g_j^{k+1} - g_j^k)$ in each iteration when the stepsizes are not shared across agents. This is because calculating $(W + W\Lambda W\Lambda^{-1}W^{-1})_{ij}$ and $(W\Lambda W\Lambda^{-1})_{ij}$ requires agent j to know all stepsizes Λ, which however, were assumed to be private to individual agents. Therefore, even though some existing gradient-tracking based algorithms can use heterogeneous stepsizes to hide information, they have to exchange two messages between interacting agents. In fact, privacy enabled in this way is quite weak, as detailed in Sec. IV-B.

B. Privacy analysis

Theorem 4. *In the presence of honest-but-curious or eavesdropping adversaries, PDG-NDS can protect the privacy of all participating agents defined in Definition 1.*

Proof. Without loss of generality, we first consider the protection of the gradient of agent i at any single time instant k, and then show that the argument also applies to any finite number of time instants (iterations). When the gradient is g_i^k, we represent the information that agent i shares with others in PDG-DS as $\mathcal{I}_i$. According to Definition 1, we have to prove that at some iteration k, if the gradient is altered from g_i^k to $\hat{g}_i^k = e^{\zeta^k}g_i^k$, the shared information $\hat{\mathcal{I}}_i$ could be identical to $\mathcal{I}_i$ under any $\zeta^k > 0$.

According to Algorithm PDG-NDS, agent i shares the following information in decentralized optimization:

$$\mathcal{I}_i = \mathcal{I}_i^{\text{sent}} \bigcup \mathcal{I}_i^{\text{public}}$$

with $\mathcal{I}_i^{\text{sent}} = \left\{v_{ji}^k | k = 1, 2, \cdots\right\}$, $v_{ji}^k = 2w_{ji}x_i^{k-1} + \{W^2\}_{ji}x_i^{k-2} - b_{ji}^{k-2}(\lambda_i^{k-1}g_i^{k-1} - \lambda_i^{k-2}g_i^{k-2})$, and $\mathcal{I}_i^{\text{public}} = \left\{W \bigcup \sum_{j\in\mathbb{N}_i} b_{ji}^k = 1 | k = 0, 1, \cdots\right\}$.

It can be obtained that when the gradient is altered to $\hat{g}_i^k = e^{\zeta^k}g_i^k$, the difference defined in (33) is ζ^k. However, in this case, if we set the stepsize $\hat{\lambda}_i^k$ to $\hat{\lambda}_i^k = e^{-\zeta^k}\lambda_i^k$, then the corresponding shared information will still be v_{ji}^k. Since other parameters are not changed and changing the stepsize from λ_i^k to $\hat{\lambda}_i^k = e^{-\zeta^k}\lambda_i^k$ at k will not violate the summable stepsize heterogeneity conditions in Theorem 3 for any given $|\zeta^k| < \infty$, convergence to the optimal solution will still be guaranteed. Similarly, if the gradient of agent i is altered at iteration k and iteration $k+1$ to $\hat{g}_i^k = e^{\zeta^k}g_i^k$ and $\hat{g}_i^{k+1} = e^{\zeta^{k+1}}g_i^{k+1}$, respectively, these alterations can be covered by a stepsize alteration of $\hat{\lambda}_i^k = e^{-\zeta^k}\lambda_i^k$ at iteration k and $\hat{\lambda}_i^{k+1} = e^{-\zeta^{k+1}}\lambda_i^{k+1}$ at iteration $k+1$. Given that the convergence condition in the statement of Theorem 3 allows the stepsize of agent i to change by any finite amount for any finite number of iterations, one can obtain that the variations of every agent's gradients in any number of iterations can be completely covered by the flexibility in changing the agent's stepsizes in these iterations, as long as the number of these iterations is finite. Therefore, privacy of the gradient information of any agent will be protected when running PDG-NDS. It is worth noting that the perturbation does not violate the convexity and Lipschitz conditions in Assumption 2. This is because in order for an adversary to check if Assumption 2 is violated, it has to know x_i^k, which, however, is not available to adversaries: before convergence, x_i^k is inaccessible to the adversary because the information shared by agent i is $2w_{ji}x_i^{k-1} + \{W^2\}_{ji}x_i^{k-2} - b_{ji}^{k-2}(\lambda_i^{k-1}g_i^{k-1} - \lambda_i^{k-2}g_i^{k-2})$, avoiding x_i^{k-1} and x_i^{k-2} from being inferrable. In fact, even with the gradient g_i^{k-1} and g_i^{k-2} unchanged, the value of observation $2w_{ji}x_i^{k-1} + \{W^2\}_{ji}x_i^{k-2} - b_{ji}^{k-2}(\lambda_i^{k-1}g_i^{k-1} - \lambda_i^{k-2}g_i^{k-2})$ can be changed by an arbitrary finite value by changing the stepsize λ_i^{k-1} or λ_i^{k-2}. After convergence, the perturbation does not violate the convexity and Lipschitz conditions, either. In fact, although x_i^k becomes accessible to the adversary after convergence, gradient information is eliminated in adversary's observation (the shared information $2w_{ji}x_i^{k-1} + \{W^2\}_{ji}x_i^{k-2} - b_{ji}^{k-2}(\lambda_i^{k-1}g_i^{k-1} - \lambda_i^{k-2}g_i^{k-2})$ becomes $2w_{ji}x_i^{k-1} + \{W^2\}_{ji}x_i^{k-2}$ because λ_i^{k-1} and λ_i^{k-2} converge to the same constant value). ■

Remark 8. *Even after convergence when g_i^k becomes a constant, an adversary still cannot infer gradients from shared messages in PDG-NDS. More specifically, when g_i^k converges to a constant value, the stepsize λ_i^k also converges to a constant value, which completely eliminates the information of g_i^k in observed information (the observed information becomes $2w_{ji}x_i^{k-1} + \{W^2\}_{ji}x_i^{k-2}$ after convergence). This can also be understood intuitively as follows: Even if the adversary can collect $T \to \infty$ observations $2w_{ji}x_i^{k-1} + \{W^2\}_{ji}x_i^{k-2} - b_{ji}^{k-2}(\lambda_i^{k-1}g_i^{k-1} - \lambda_i^{k-2}g_i^{k-2})$ in the neighborhood of the optimal point and establish a system of T equations to solve for g_i^{k-1} and g_i^{k-2} (which can be viewed to be approximately time-invariant in the neighborhood of the optimal point), the number of unknowns b_{ji}^{k-2}, λ_i^{k-1}, λ_i^{k-2}, g_i^{k-1} and g_i^{k-2} in the system of T equations is $5T$ (even if we view λ_i^{k-1} and λ_i^{k-2} to be approximately constant and equal to each other, g_i^{k-1} and g_i^{k-2} to be constant and equal to each other, in the neighborhood of the optimal point, the number of unknowns is still $T + 2$), which makes it impossible for the adversary to solve for g_i^{k-1} or g_i^{k-2} using the system of T equations established from T observations.*

Remark 9. *Similar to PDG-DS, our PDG-NDS algorithm can protect the privacy of every participating agent against honest-but-curious and eavesdropping adversaries without any constraint on the interaction topology.*

We next show that directly making stepsize and coupling matrices time-varying in existing gradient-tracking based algorithms cannot provide the defined privacy. We use the AB algorithm in [9] as an example to show this since it allows a column-stochastic coupling matrix, which allows individual

agents to keep their coupling coefficients private. The AB algorithm has the following form [9]:

$$\begin{cases} x^{k+1} = Rx_j^k - \lambda y^k \\ y^{k+1} = C(y^k + g^{k+1} - g^k) \end{cases}$$

where $R = \{r_{ij}\}$ is row-stochastic and $C = \{c_{ij}\}$ is column-stochastic.

Directly making its stepsize and coupling coefficients time-varying leads to the following algorithm (we also introduce heterogeneity in the stepsize):

$$\begin{cases} x^{k+1} = R^k x_j^k - \Lambda^k y^k \\ y^{k+1} = C^k(y^k + g^{k+1} - g^k) \end{cases}$$

where $R^k = \{r_{ij}^k\}$ should be row-stochastic and $C^k = \{c_{ij}^k\}$ should be column-stochastic. At each iteration k, an agent j will share x_j^k and $c_{ij}^k(y_j^k + g_j^{k+1} - g_j^k)$ with its neighboring agent i. Also, all agents initialize as $y_i^0 = g_i^0$.

Because for all $i \in \mathbb{N}_j$, c_{ij}^k are generated by agent j, it seems that agent j can keep c_{ij}^k confidential and hence uses them to cover shared information $c_{ij}^k(y_j^k + g_j^{k+1} - g_j^k)$. Next, we show that this is not true.

We consider the case where agent i is the only neighbor of agent j. In this case, agent i knows agent j's update rule

$$y_j^{k+1} = c_{jj}^k(y_j^k + g_j^{k+1} - g_j^k) + c_{ji}^k(y_i^k + g_i^{k+1} - g_i^k)$$

Using the fact $c_{jj}^k + c_{ij}^k = 1$, the above update rule can be rewritten as

$$\begin{aligned} y_j^{k+1} - y_j^k &= g_j^{k+1} - g_j^k \\ &\quad - c_{ij}^k(y_j^k + g_j^{k+1} - g_j^k) + c_{ji}^k(y_i^k + g_i^{k+1} - g_i^k) \end{aligned} \tag{59}$$

Note that $c_{ij}^k(y_j^k + g_j^{k+1} - g_j^k)$ is shared with agent i by agent j and $c_{ji}^k(y_i^k + g_i^{k+1} - g_i^k)$ is generated by agent i, hence the last two terms on the right hand side of (59) are known to agent i. We represent $-c_{ij}^k(y_j^k + g_j^{k+1} - g_j^k) + c_{ji}^k(y_i^k + g_i^{k+1} - g_i^k)$ as m_i^k and add (59) from $k = 0$ to t to obtain $y_j^{t+1} = g_j^{t+1} + \sum_{k=0}^t m_i^k$ where we used the relationship $y_j^0 = g_j^0$.

When $t \to \infty$, we have $y_j^t \to 0$ and $x_i^k \to x_j^k$ in the AB algorithm, resulting in $g_j^{t+1} = -\sum_{k=0}^t m_i^k$. Therefore, agent i can infer the gradient of agent j based on its accessible information m_i^k. Note that the above derivation is independent of the evolution of x^k and stepsize Λ^k, so the same privacy leakage will occur even if the stepzies are uncoordinated (not shared across agents) such as in [8], [42].

One may wonder if we can reduce the AB algorithm to the x-variable only form to avoid information leakage. Given that when the stepsizes are heterogeneous and not shared across agents, such reduction is impossible, as detailed in Remark 7, we only consider the homogeneous stepsize case. In fact, after eliminating the y^k variable, the AB algorithm reduces to

$$x^{k+2} = (R^k + C^k)x^{k+1} - C^k R^k x^k - \lambda C^k(g^{k+1} - g^k) \tag{60}$$

Note that for privacy-preserving purposes, agent j should keep c_{ij}^k private and send $(R^k + C^k)_{ij}x_j^{k+1} - (C^k R^k)_{ij}x_j^k - \lambda C^k(g_j^{k+1} - g_j^k)$ to agent i where $(\cdot)_{ij}$ represents the (i,j)th element of a matrix. Given $(C^k R^k)_{ij} = \sum_{p=1}^m c_{ip}^k r_{pj}^k$, agent j has to know all elements of C^k to implement the algorithm in the x-variable only form (60), which contradicts the assumption that the elements of C^k are kept private to cover information. In other words, if the column-stochastic matrix C^k is used to cover information, the AB algorithm cannot be implemented in an x-variable only form. In summary, gradient-tracking based decentralized optimization algorithms cannot be used to enable the privacy defined in this paper even under time-varying coupling weights and heterogeneous stepsizes.

V. Numerical Simulations

We consider the canonical distributed estimation problem where a sensor network of m sensors are used to collectively estimate an unknown parameter $\theta \in \mathbb{R}^d$. Each sensor 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 and w_i is Gaussian noise. The maximum likelihood estimation problem can be formulated as the decentralized optimization problem (1) with each f_i given by $f_i(\theta) = \|z_i - M_i\theta\|^2 + \sigma_i\|\theta\|^2$ where $\sigma_i \geq 0$ is the regularization parameter [15].

We considered a network of $m = 5$ sensors interacting on the graph depicted in Fig. 1. We set $s = 3$ and $d = 2$. To evaluate the performance of our PDG-DS algorithm, we set the stepsize of agent i as $\lambda_i^k = \frac{1-\varrho_i^k/k^2}{k}$ where ϱ_i^k was randomly chosen by agent i from the interval $[0,\ 1]$ for each iteration. Given that different agents i chose ϱ_i^k independently, the stepsizes are heterogeneous across the agents. Each agent i also chose b_{ji}^k for all $j \in \mathbb{N}_i$ randomly and independently of each other under the sum-one condition (to make B^k column-stochastic). The evolution of the optimization error of PDG-DS is given by the dashed magenta line in Fig. 2. When we fixed the B^k matrix to an identity matrix, we also obtained convergence to the optimal solution, which is illustrated by the solid green line in Fig. 2. The evolution of the conventional decentralized gradient algorithm (3) under homogeneous diminishing stepsize $\frac{1}{k}$ is also presented in Fig. 2 by the dotted blue line for comparison. It can be seen that our PDG-DS algorithm has a comparable (in fact faster) convergence speed with the conventional decentralized gradient algorithm which does not take privacy into account. Furthermore, comparing the case with B^k and the case without B^k, we can see that by using the mixing matrix B^k in PDG-DS, we get faster convergence. This is intuitive since the B^k matrix enhances mixing of information across the agents.

We also simulated our PDG-NDS algorithm with non-diminishing stepsizes. More specifically, we set the stepsize of agent i to $0.02(1 - \frac{\varrho_i^k}{k^2})$ with ϱ_i^k randomly chosen by agent i from $[0,\ 1]$ for each iteration. Again, since each agent i chose ϱ_i^k randomly and independently of each other, the stepsizes are heterogeneous across the network. Each agent i also randomly chose b_{ji}^k for all $j \in \mathbb{N}_i$ under the sum-one condition (to make B^k column-stochastic). The evolution of the optimization error of PDG-NDS is illustrated by the solid line in Fig. 3. For the purpose of comparison, we also plotted the optimization results of gradient-tracking algorithms DIGing [14], Push-Pull [16], and ADD-OPT [18] under homogeneous stepsize 0.02.

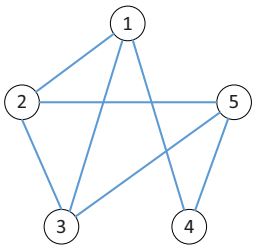

Fig. 1. The interaction topology of the network.

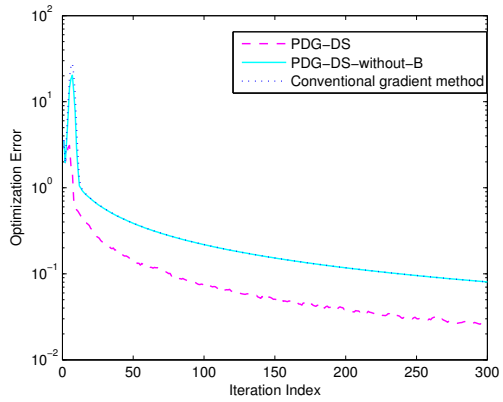

Fig. 2. Comparison of PDG-DS with the conventional decentralized gradient method under diminishing-stepsizes.

It can be seen that PDG-NDS provides similar convergence performance besides enabling privacy protection.

VI. Conclusions

This paper proposes two inherently privacy-preserving decentralized optimization algorithms which can guarantee the privacy of all participating agents without compromising optimization accuracy. This is in distinct difference from differential privacy based approaches which trade optimization accuracy for privacy. The two algorithms are also efficient in communication and computation in that they are encryption-free and only require an agent to share one message with a neighboring agent in every iteration. The two approaches can protect the privacy of every agent even if all information shared by an agent is accessible to an adversary, in which case most existing accuracy-maintaining privacy-preserving decentralized optimization solutions fail to provide privacy protection. In fact, even without considering privacy, the convergence analyses of the two algorithms under time-varying uncoordinated stepsizes are also of interest by themselves since existing results only consider constant or fixed heterogeneity in stepsizes. Numerical simulation results show that both approaches have similar convergence speeds compared with their respective privacy-violating counterparts.

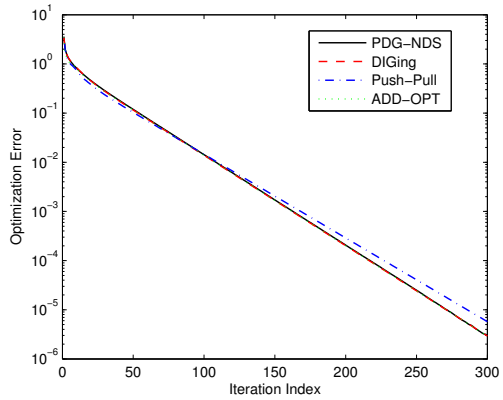

Fig. 3. Comparison of PDG-NDS with some gradient-tracking based decentralized optimization algorithms under non-diminishing-stepsizes.

Appendix A

Lemma 3. *([45], Lemma 11, page 50) Let* $\{v^k\}$, $\{\alpha^k\}$, *and* $\{\beta^k\}$ *be sequences of nonnegative scalars such that* $\sum_{k=0}^{\infty}\alpha^k<\infty$, $\sum_{k=0}^{\infty}\beta^k<\infty$, *and* $v^{k+1}\leq(1+\alpha^k)v^k+\beta^k$ *holds for all* $k\geq 0$. *Then, the sequence* $\{v^k\}$ *is convergent, i.e.,* $\lim_{k\to\infty}v^k=v$ *for some* $v\geq 0$.

Lemma 4. *([45], Lemma 10, page 49) Let* $\{v^k\}$, $\{\alpha^k\}$, *and* $\{\beta^k\}$ *be sequences of nonnegative scalars such that* $\sum_{k=0}^{\infty}\alpha^k=\infty$, $\lim_{k\to\infty}\beta^k/\alpha^k=0$, *and* $v^{k+1}\leq(1-\alpha^k)v^k+\beta^k$ *and* $\alpha^k\leq 1$ *hold for all* k. *Then, the sequence* $\{v^k\}$ *converges to 0, i.e.,* $\lim_{k\to\infty}v^k=0$.

Lemma 5. *[46] Consider a minimization 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 sequence such that for any optimal solution* $z^*\in Z^*$ *and for all* $k\geq 0$,

$$\|z^{k+1}-z^*\|^2\leq(1+\alpha^k)\|z^k-z^*\|^2-\gamma^k\left(\phi(z^k)-\phi(z^*)\right)+\beta^k,$$

where $\alpha_k\geq 0$, $\beta_k\geq 0$, *and* $\gamma_k\geq 0$ *for all* $k\geq 0$, *with* $\sum_{k=0}^{\infty}\alpha^k<\infty$, $\sum_{k=0}^{\infty}\gamma^k=\infty$, *and* $\sum_{k=0}^{\infty}\beta^k<\infty$. *Then, the sequence* $\{z^k\}$ *converges to some optimal solution* $\tilde{z}^*\in Z^*$.

Appendix B

In this section, we establish the relations in Step I, Step II, and Step III of Theorem 3's proof.

Step I: Relationship for $\frac{2}{L}\left(F(\bar{x}^{k+1})-F(\theta^*)\right)$.

Since F is convex with a Lipschitz gradient, we have:

$$F(y)\leq F(x)+\langle\nabla F(x),y-x\rangle+\frac{L}{2}\|y-x\|^2,\ \forall y,x\in\mathbb{R}^d$$

Letting $y=\bar{x}^{k+1}$ and $x=\bar{x}^k$ in the preceding relation and using (39) as well as $F(\bar{x}^k)=\frac{1}{m}\sum_{i=1}^m f_i(\bar{x}^k)$, we obtain

$$F(\bar{x}^{k+1})\leq F(\bar{x}^k)-\frac{1}{m}\sum_{i=1}^m\langle\nabla f_i(\bar{x}^k),\bar{y}^k\rangle+\frac{L}{2}\|\bar{y}^k\|^2$$

Subtracting $F(\theta^*)$ from both sides and multiplying $2\bar{\lambda}^k$ yield

$$\begin{aligned}2\bar{\lambda}^k\left(F(\bar{x}^{k+1})-F(\theta^*)\right)\leq{}&2\bar{\lambda}^k\left(F(\bar{x}^k)-F(\theta^*)\right)\\&-\frac{2\bar{\lambda}^k}{m}\sum_{i=1}^m\langle\nabla f_i(\bar{x}^k),\bar{y}^k\rangle+\bar{\lambda}^kL\|\bar{y}^k\|^2\end{aligned}\tag{61}$$

The term $-\frac{2\bar{\lambda}^k}{m}\sum_{i=1}^m\langle\nabla f_i(\bar{x}^k),\bar{y}^k\rangle$ satisfies

$$-2\left\langle\frac{\bar{\lambda}^k}{m}\sum_{i=1}^m\langle\nabla f_i(\bar{x}^k),\bar{y}^k\right\rangle$$
$$=\left\|\frac{\bar{\lambda}^k}{m}\sum_{i=1}^m\nabla f_i(\bar{x}^k)-\bar{y}^k\right\|^2-\left\|\frac{\bar{\lambda}^k}{m}\sum_{i=1}^m\nabla f_i(\bar{x}^k)\right\|^2-\left\|\bar{y}^k\right\|^2 \tag{62}$$

For the first term on the right hand side of (62), by adding and subtracting $\frac{1}{m}\sum_{i=1}^m\lambda_i^k\nabla f_i(\bar{x}^k)$, we obtain

$$\left\|\tfrac{\bar{\lambda}^k}{m}\textstyle\sum_{i=1}^m\nabla f_i(\bar{x}^k)-\bar{y}^k\right\|^2$$
$$=\left\|\tfrac{1}{m}\textstyle\sum_{i=1}^m(\bar{\lambda}^k-\lambda_i^k)\nabla f_i(\bar{x}^k)+\tfrac{1}{m}\sum_{i=1}^m\lambda_i^k\nabla f_i(\bar{x}^k)-\bar{y}^k\right\|^2$$
$$\le 2\left\|\tfrac{1}{m}\textstyle\sum_{i=1}^m(\bar{\lambda}^k-\lambda_i^k)\nabla f_i(\bar{x}^k)\right\|^2$$
$$+2\left\|\tfrac{1}{m}\textstyle\sum_{i=1}^m\lambda_i^k\left(\nabla f_i(\bar{x}^k)-\nabla f_i(x_i^k)\right)\right\|^2$$

where we used $\bar{y}^k=\frac{1}{m}\sum_{i=1}^m\lambda_i^k\nabla f_i(x_i^k)$ in (40). Using the assumption that each $\nabla f_i(\cdot)$ is Lipschitz continuous with a constant L, we can further rewrite the preceding inequality as

$$\left\|\frac{\bar{\lambda}^k}{m}\sum_{i=1}^m\nabla f_i(\bar{x}^k)-\bar{y}^k\right\|^2\le\frac{2}{m}\sum_{i=1}^m(\bar{\lambda}^k-\lambda_i^k)^2\|\nabla f_i(\bar{x}^k)\|^2$$
$$+\frac{2}{m}\sum_{i=1}^m(\lambda_i^k)^2\|\nabla f_i(\bar{x}^k)-\nabla f_i(x_i^k)\|^2$$
$$\le\frac{2}{m}\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2\sum_{i=1}^m\|\nabla f_i(\bar{x}^k)\|^2+\frac{2L^2}{m}(\lambda_{\max}^k)^2\sum_{i=1}^m\|\bar{x}^k-x_i^k\|^2 \tag{63}$$

We next proceed to analyze $\sum_{i=1}^m\|\nabla f_i(\bar{x}^k)\|^2$.

By Assumption 2, each $\nabla f_i(\cdot)$ is Lipschitz continuous with a constant L, so we have

$$f_i(v)+\langle\nabla f_i(v),u-v\rangle+\frac{1}{2L}\|\nabla f_i(v)-\nabla f_i(u)\|^2\le f_i(u),\forall u,v$$

Letting $v=\theta^*$ and $u=\bar{x}^k$, and summing the resulting relations over $i=1,\ldots,m$ yield $F(\theta^*)+\langle\nabla F(\theta^*),\bar{x}^k-\theta^*\rangle+\frac{\sum_{i=1}^m\|\nabla f_i(\theta^*)-\nabla f_i(\bar{x}^k)\|^2}{2mL}\le F(\bar{x}^k)$. Using $\nabla F(\theta^*)=0$, we have $\sum_{i=1}^m\|\nabla f_i(\theta^*)-\nabla f_i(\bar{x}^k)\|^2\le 2mL(F(\bar{x}^k)-F(\theta^*))$. Thus, it follows

$$\sum_{i=1}^m\|\nabla f_i(\bar{x}^k)\|^2\le\sum_{i=1}^m2\left(\|\nabla f_i(\bar{x}^k)-\nabla f_i(\theta^*)\|^2+\|\nabla f_i(\theta^*)\|^2\right)$$
$$\le 4mL(F(\bar{x}^k)-F(\theta^*))+2\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2 \tag{64}$$

Combining (61), (62), (63), and (64) leads to

$$\frac{2}{L}\left(F(\bar{x}^{k+1})-F(\theta^*)\right)\le\frac{2}{L}\left(F(\bar{x}^k)-F(\theta^*)\right)$$
$$+\frac{c_1}{\bar{\lambda}^k}\|\lambda^k-\bar{\lambda}^k\mathbf{1}\|^2\left(\frac{2}{L}(F(\bar{x}^k)-F(\theta^*))+\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2\right)$$
$$+\frac{2L}{m\bar{\lambda}^k}(\lambda_{\max}^k)^2\sum_{i=1}^m\|\bar{x}^k-x_i^k\|^2$$
$$-\frac{\bar{\lambda}^k}{L}\|\nabla F(\bar{x}^k)\|^2+\frac{\bar{\lambda}^kL-1}{\bar{\lambda}^kL}\|\bar{y}^k\|^2 \tag{65}$$

where $c_1=\max\{4L,\,4/(mL)\}$.

Step II: Relationship for $\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|$ and $\sum_{i=1}^m\|x_i^{k+1}-x_i^k\|^2$.

For the convenience of analysis, we write the iterates of algorithm PDG-NDS on per-coordinate expressions. Define for all $\ell=1,\ldots,d$, and $k\ge0$,

$$x^k(\ell)=([x_1^k]_\ell,\ldots,[x_m^k]_\ell)^T,\ y^k(\ell)=([y_1^k]_\ell,\ldots,[y_m^k]_\ell)^T,$$
$$g^k(\ell)=([g_1^k]_\ell,\ldots,[g_m^k]_\ell)^T.$$

In this per-coordinate view, (37) and (38) has the following form for all $\ell=1,\ldots,d$, and $k\ge0$,

$$x^{k+1}(\ell)=Wx^k(\ell)-y^k(\ell)$$
$$y^{k+1}(\ell)=Wy^k(\ell)+B^k\left(\Lambda^{k+1}g^{k+1}(\ell)-\Lambda^kg^k(\ell)\right) \tag{66}$$

From the definition of $x^{k+1}(\ell)$ in (66), and the relation for the average $\bar{x}^{k+1}$ in (39), we obtain for all $\ell=1,\ldots,d$,

$$x^{k+1}(\ell)-[\bar{x}^{k+1}]_\ell\mathbf{1}=W\left(x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\right)-(y^k(\ell)-[\bar{y}^k]_\ell\mathbf{1})$$

where we use $W\mathbf{1}=\mathbf{1}$. Noting that $[\bar{x}^k]_\ell$ is the average of $x^k(\ell)$, i.e., $\frac{1}{m}\mathbf{1}\mathbf{1}^T\left(x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\right)=0$, we have

$$x^{k+1}(\ell)-[\bar{x}^{k+1}]_\ell\mathbf{1}=\bar{W}\left(x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\right)-(y^k(\ell)-[\bar{y}^k]_\ell\mathbf{1})$$

where $\bar{W}=W-\frac{\mathbf{1}\mathbf{1}^T}{m}$. So it follows

$$\|x^{k+1}(\ell)-[\bar{x}^{k+1}]_\ell\mathbf{1}\|\le\eta\|x^k-\bar{x}^k\mathbf{1}\|+\|y^k(\ell)-[\bar{y}^k]_\ell\mathbf{1}\|$$

with $\eta=\|W-\frac{1}{m}\mathbf{1}\mathbf{1}^T\|<1$. Taking squares on both sides of the preceding relation, and using the inequality $(a+b)^2\le(1+\epsilon)a^2+(1+\epsilon^{-1})b^2$, valid for any scalars a and b, and $\epsilon>0$, we obtain

$$\|x^{k+1}(\ell)-[\bar{x}^{k+1}]_\ell\mathbf{1}\|^2\le\eta^2(1+\epsilon)\|x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\|^2$$
$$+(1+\epsilon^{-1})\|y^k(\ell)-[\bar{y}^k]_\ell\mathbf{1}\|^2$$

By using $\eta\in(0,1)$ and letting $\epsilon=\frac{1-\eta}{\eta}$ which implies $1+\epsilon=\eta^{-1}$ and $1+\epsilon^{-1}=(1-\eta)^{-1}$, we have

$$\|x^{k+1}(\ell)-[\bar{x}^{k+1}]_\ell\mathbf{1}\|^2\le\eta\|x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\|^2$$
$$+(1-\eta)^{-1}\|y^k(\ell)-[\bar{y}^k]_\ell\mathbf{1}\|^2$$

Summing the preceding relations over $\ell=1,\ldots,d$, and noting $\sum_{\ell=1}^d\|x^{k+1}(\ell)-[\bar{x}^{k+1}]_\ell\mathbf{1}\|^2=\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|^2$, $\sum_{\ell=1}^d\|x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\|^2=\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2$, and $\sum_{\ell=1}^d\|y^k(\ell)-[\bar{y}^k]_\ell\mathbf{1}\|^2=\sum_{i=1}^m\|y_i^k-\bar{y}^k\|^2$, we obtain

$$\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|^2\le\eta\sum_{i=1}^m\|x_i^{k+1}-\bar{x}^{k+1}\|^2$$
$$+(1-\eta)^{-1}\sum_{i=1}^m\|y_i^k-\bar{y}^k\|^2 \tag{67}$$

Next we proceed to analyze $\sum_{i=1}^m\|x_i^{k+1}-x_i^k\|^2$. Using (37), we have for every coordinate index $\ell=1,\ldots,d$

$$x^{k+1}(\ell)-x^k(\ell)=Wx^k(\ell)-y^k(\ell)-x^k(\ell)$$
$$=(W-I)x^k(\ell)-y^k(\ell)=(W-I)(x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1})-y^k(\ell)$$

where we used the fact $(W-I)\mathbf{1}=0$. By letting $r=\|W-I\|$, we obtain

$$\|x^{k+1}(\ell)-x^k(\ell)\|\le r\|x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\|+\|y^k(\ell)\|$$
$$\le r\|x^k(\ell)-[\bar{x}^k]_\ell\mathbf{1}\|+\|y^k-[\bar{y}^k]_\ell\mathbf{1}\|+\sqrt{m}|[\bar{y}^k]_\ell|$$

where the last inequality is obtained by adding and subtracting $[\bar{y}^k]_\ell \mathbf{1}$ to $y^k(\ell)$, and using the triangle inequality for the norm. Thus, we have

$$\begin{aligned}\|x^{k+1}(\ell) - x^k(\ell)\|^2 \leq & 3r^2\|x^k(\ell) - [\bar{x}^k]_\ell \mathbf{1}\|^2 + \\ & 3\|y^k - [\bar{y}^k]_\ell \mathbf{1}\|^2 + 3m|[\bar{y}^k]_\ell|^2\end{aligned}$$

By summing over $\ell = 1, \ldots, d$, we obtain

$$\begin{aligned}\sum_{i=1}^m \|x_i^{k+1} - x_i^k\|^2 \leq & 3r^2 \sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \\ & + 3\sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2 + 3m\|\bar{y}^k\|^2\end{aligned} \tag{68}$$

Step III: Relationship for $\sum_{i=1}^m \|y_i^{k+1} - \bar{y}^{k+1}\mathbf{1}\|$.

Using the column stochastic property of B^k, from (38), the ℓth entries of $[\bar{y}^k]_\ell$ satisfy

$$[\bar{y}^{k+1}]_\ell = [\bar{y}^k]_\ell + \frac{1}{m}\mathbf{1}^T\Lambda^{k+1}g^{k+1}(\ell) - \frac{1}{m}\mathbf{1}^T\Lambda^k g^k(\ell)$$

Then, using (66), we obtain for all $\ell = 1, \ldots, d$,

$$\begin{aligned}y^{k+1}(\ell) - [\bar{y}^{k+1}]_\ell \mathbf{1} = & \bar{W}(y^k(\ell) - [\bar{y}^k]_\ell \mathbf{1}) + \\ & \bar{B}^k\Lambda^{k+1}g^{k+1}(\ell) - \bar{B}^k\Lambda^k g^k(\ell)\end{aligned}$$

where $\bar{W} = W - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ and $\bar{B}^k = \left(B^k - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right)$.

By adding and subtracting $\bar{B}^k\Lambda^{k+1}g^k(\ell)$, and taking the Euclidean norm, we find that for all $\ell = 1, \ldots, d$,

$$\begin{aligned}&\|y^{k+1}(\ell) - [\bar{y}^{k+1}]_\ell \mathbf{1}\| \leq \eta\|y^k(\ell) - [\bar{y}^k]_\ell \mathbf{1}\| \\ &+ \tau\left\|\Lambda^{k+1}\left(g^{k+1}(\ell) - g^k(\ell)\right)\right\| + \tau\left\|\left(\Lambda^{k+1} - \Lambda^k\right)g^k(\ell)\right\|\end{aligned} \tag{69}$$

where $\eta = \|\bar{W}\|$ and $\tau = \|\bar{B}^k\|$.

Since we always have $\left\|\bar{B}^k\right\| \leq \left\|\bar{B}^k\right\|_F \leq m$ where $\|\cdot\|_F$ denotes the Frobenius matrix norm, we have

$$\tau \leq m \tag{70}$$

Using the fact that Λ^k is a diagonal matrix for all $k \geq 0$, i.e., $\Lambda^k = \mathrm{diag}(\lambda^k)$, we have

$$\begin{aligned}&\left\|\Lambda^{k+1}\left(g^{k+1}(\ell) - g^k(\ell)\right)\right\| \leq \|\lambda^{k+1}\|\,\|g^{k+1}(\ell) - g^k(\ell)\|, \\ &\left\|\left(\Lambda^{k+1} - \Lambda^k\right)g^k(\ell)\right\| \leq \|\lambda^{k+1} - \lambda^k\|\,\|g^k(\ell)\|\end{aligned} \tag{71}$$

Therefore, combining (69), (70), and (71) leads to

$$\begin{aligned}&\|y^{k+1}(\ell) - [\bar{y}^{k+1}]_\ell \mathbf{1}\| \leq \eta\|y^k(\ell) - [\bar{y}^k]_\ell \mathbf{1}\| + \\ &m\|\lambda^{k+1}\|\,\|g^{k+1}(\ell) - g^k(\ell)\| + m\left\|\lambda^{k+1} - \lambda^k\right\|\,\|g^k(\ell)\|\end{aligned}$$

Thus, by taking squares and using $(a+b)^2 \leq (1+\epsilon)a^2 + (1+\epsilon^{-1})b^2$ holding for any $\epsilon > 0$, we obtain the following inequality by setting $\epsilon = \frac{1-\eta}{\eta}$:

$$\begin{aligned}&\|y^{k+1}(\ell) - [\bar{y}^{k+1}]_\ell \mathbf{1}\|^2 \leq \eta\|y^k(\ell) - [\bar{y}^k]_\ell \mathbf{1}\|^2 + \frac{2m^2}{1-\eta}\times \\ &\left(\|\lambda^{k+1}\|^2\,\|g^{k+1}(\ell) - g^k(\ell)\|^2 + \left\|\lambda^{k+1} - \lambda^k\right\|^2\,\|g^k(\ell)\|^2\right)\end{aligned}$$

By summing these relations over $\ell = 1, \ldots, d$, we find

$$\begin{aligned}&\sum_{i=1}^m \|y_i^{k+1} - \bar{y}^{k+1}\|^2 \leq \eta\sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2 + \frac{2m^2}{1-\eta}\times \\ &\left(\|\lambda^{k+1}\|^2\sum_{i=1}^m \|g_i^{k+1} - g_i^k\|^2 + \left\|\lambda^{k+1} - \lambda^k\right\|^2\sum_{i=1}^m \|g_i^k\|^2\right)\end{aligned} \tag{72}$$

Next we bound $\sum_{i=1}^m \|g_i^{k+1} - g_i^k\|^2$ and $\sum_{i=1}^m \|g_i^k\|^2$.

Since every ∇f_i is Lipschitz continuous with $L > 0$, we have

$$\sum_{i=1}^m \|g_i^{k+1} - g_i^k\|^2 \leq L^2\sum_{i=1}^m \|x_i^{k+1} - x_i^k\|^2$$

which, in combination with (68), leads to

$$\begin{aligned}\sum_{i=1}^m \|g_i^{k+1} - g_i^k\|^2 \leq & 3r^2L^2\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 \\ & + 3L^2\sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2 + 3mL^2\|\bar{y}^k\|^2\end{aligned} \tag{73}$$

Using the Lipschitz continuity of each g_i^k, we obtain

$$\begin{aligned}\sum_{i=1}^m \|g_i^k\|^2 &= \sum_{i=1}^m \|\nabla f_i(x_i^k)\|^2 \\ &\leq 2L^2\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 2\sum_{i=1}^m \|\nabla f_i(\bar{x}^k)\|^2\end{aligned}$$

which, in combination with (64), leads to

$$\begin{aligned}\sum_{i=1}^m \|g_i^k\|^2 \leq & 2L^2\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + 8mL(F(\bar{x}^k) - F(\theta^*)) \\ & + 4\sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2\end{aligned} \tag{74}$$

By substituting (73) and (74) into (72), we obtain

$$\begin{aligned}&\sum_{i=1}^m \|y_i^{k+1} - \bar{y}^{k+1}\|^2 \leq (\eta + \frac{6m^2L^2}{1-\eta}\|\lambda^{k+1}\|^2)\sum_{i=1}^m \|y_i^k - \bar{y}^k\|^2 \\ &+ \frac{6m^2r^2L^2}{1-\eta}\left(\left\|\lambda^{k+1}\right\|^2\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + m\left\|\lambda^{k+1}\right\|^2\|\bar{y}^k\|^2\right) \\ &+ \frac{2m^2}{1-\eta}\left\|\lambda^{k+1} - \lambda^k\right\|^2\left(2L^2\sum_{i=1}^m \|x_i^k - \bar{x}^k\|^2 + \right. \\ &\left. 8mL(F(\bar{x}^k) - F(\theta^*)) + 4\sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2\right)\end{aligned} \tag{75}$$

REFERENCES

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

[2] J. Bazerque and G. Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, 58(3):1847–1862, 2009.

[3] R. Raffard, C. Tomlin, and S. Boyd. Distributed optimization for cooperative agents: Application to formation flight. In *2004 43rd IEEE Conference on Decision and Control (CDC)*, volume 3, pages 2453–2459. IEEE, 2004.

[4] C. Zhang and Y. Wang. Distributed event localization via alternating direction method of multipliers. *IEEE Transactions on Mobile Computing*, 17(2):348–361, 2017.

[5] K. Tsianos, S. Lawlor, and M. Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *50th Annual Allerton Conference on Communication, Control, and Computing*, pages 1543–1550. IEEE, 2012.

[6] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[7] W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[8] J. Xu, S. Zhu, Y. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *54th IEEE Conference on Decision and Control*, pages 2055–2060. IEEE, 2015.

[9] R. Xin and U. Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.

[10] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[11] C. Zhang, M. Ahmad, and Y. Wang. ADMM based privacy-preserving decentralized optimization. *IEEE Transactions on Information Forensics and Security*, 14(3):565–580, 2019.

[12] E. Wei, A. Ozdaglar, and A. Jadbabaie. A distributed newton method for network utility maximization–i: Algorithm. *IEEE Transactions on Automatic Control*, 58(9):2162–2175, 2013.

[13] K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[14] A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[15] J. Xu, S. Zhu, Y. Soh, and L. Xie. Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Transactions on Automatic Control*, 63(2):434–448, 2017.

[16] S. Pu, W. Shi, J. Xu, and A. Nedić. Push-pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 2020.

[17] W. Du, L. Yao, D. Wu, X. Li, G. Liu, and T. Yang. Accelerated distributed energy management for microgrids. In *IEEE Power & Energy Society General Meeting*, pages 1–5. IEEE, 2018.

[18] C. Xi, R. Xin, and U. Khan. Add-opt: Accelerated distributed directed optimization. *IEEE Transactions on Automatic Control*, 63(5):1329–1339, 2017.

[19] Z. Huang, S. Mitra, and N. Vaidya. Differentially private distributed optimization. In *2015 International Conference on Distributed Computing and Networking*, pages 1–10, 2015.

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

[21] 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*, 25(11):2483–2493, 2012.

[22] K. Wei, J. Li, M. Ding, C. Ma, H. Yang, F. Farokhi, S. Jin, T. Quek, and V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[23] L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, pages 14774–14784, 2019.

[24] J. Cortés, G. Dullerud, S. Han, J. Le Ny, S. Mitra, and G. Pappas. Differential privacy in control and network systems. In *IEEE 55th Conference on Decision and Control*, pages 4252–4272. IEEE, 2016.

[25] 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.

[26] C. Zhang and Y. Wang. Enabling privacy-preservation in decentralized optimization. *IEEE Transactions on Control of Network Systems*, 6(2):679–689, 2018.

[27] N. Freris and P. Patrinos. Distributed computing over encrypted data. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1116–1122. IEEE, 2016.

[28] Y. Lu and M. Zhu. Privacy preserving distributed optimization using homomorphic encryption. *Automatica*, 96:314–325, 2018.

[29] S. Gade and N. Vaidya. Private optimization on networks. In *2018 Annual American Control Conference (ACC)*, pages 1402–1409. IEEE, 2018.

[30] Y. Wang and T. Başar. Quantization enabled privacy protection in decentralized stochastic optimization. *IEEE Transactions on Automatic Control*, 68(7):4038 – 4052, 2023.

[31] Y. Wang and A. Nedić. Tailoring gradient methods for differentially-private distributed optimization. *IEEE Transactions on Automatic Control*, 2023.

[32] E. Nozari, P. Tallapragada, and J. Cortés. Differentially private distributed convex optimization via functional perturbation. *IEEE Transactions on Control of Network Systems*, 5(1):395–408, 2016.

[33] Y. Lou, L. Yu, S. Wang, and P. Yi. Privacy preservation in distributed subgradient optimization algorithms. *IEEE transactions on cybernetics*, 48(7):2154–2165, 2017.

[34] M. Ruan, H. Gao, and Y. Wang. Secure and privacy-preserving consensus. *IEEE Transactions on Automatic Control*, 64(10):4035–4049, 2019.

[35] Y. Wang. Privacy-preserving average consensus via state decomposition. *IEEE Transactions on Automatic Control*, 64(11):4711–4716, 2019.

[36] H. Gao and Y. Wang. Algorithm-level confidentiality for average consensus on time-varying directed graphs. *IEEE Transactions on Network Science and Engineering*, 9(2):918–931, 2022.

[37] P. Di Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

[38] A. Daneshmand, G. Scutari, and V. Kungurtsev. Second-order guarantees of distributed gradient algorithms. *SIAM Journal on Optimization*, 30(4):3029–3068, 2020.

[39] R. Xin, S. Pu, A. Nedić, and U. Khan. A general framework for decentralized optimization with first-order methods. *Proceedings of the IEEE*, 108(11):1869–1889, 2020.

[40] M. Bin, I. Notarnicola, L. Marconi, and G. Notarstefano. A system theoretical perspective to gradient-tracking algorithms for distributed quadratic optimization. In *58th Conference on Decision and Control*, pages 2994–2999. IEEE, 2019.

[41] J. Zhang, K. You, and K. Cai. Distributed dual gradient tracking for resource allocation in unbalanced networks. *IEEE Transactions on Signal Processing*, 68:2186–2198, 2020.

[42] A. Nedić, A. Olshevsky, W. Shi, and C. Uribe. Geometrically convergent distributed optimization with uncoordinated step-sizes. In *2017 American Control Conference (ACC)*, pages 3950–3955. IEEE, 2017.

[43] O. Goldreich. *Foundations of Cryptography: volume 2, Basic Applications*. Cambridge University Press, 2001.

[44] Y. Wang and V. Poor. Decentralized stochastic optimization with inherent privacy protection. *IEEE Transactions on Automatic Control*, 68(4):2293–2308, 2023.

[45] B. Polyak. Introduction to optimization. *Optimization software Inc., Publications Division, New York*, 1, 1987.

[46] A. Nedić and A. Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

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.