A Fast Randomized Incremental Gradient Method for Decentralized Nonconvex Optimization

Ran Xin, Usman A. Khan, and Soummya Kar

***Abstract*—In this article, we study decentralized nonconvex finite-sum minimization problems described over a network of nodes, where each node possesses a local batch of data samples. In this context, we analyze a single-timescale randomized incremental gradient method, called GT-SAGA. GT-SAGA is computationally efficient as it evaluates one component gradient per node per iteration and achieves provably fast and robust performance by leveraging node-level variance reduction and network-level gradient tracking. For general smooth nonconvex problems, we show the almost sure and mean-squared convergence of GT-SAGA to a first-order stationary point and further describe regimes of practical significance, where it outperforms the existing approaches and achieves a network topology-independent iteration complexity, respectively. When the global function satisfies the Polyak–Łojaciewisz condition, we show that GT-SAGA exhibits linear convergence to an optimal solution in expectation and describe regimes of practical interest where the performance is network topology independent and improves upon the existing methods. Numerical experiments are included to highlight the main convergence aspects of GT-SAGA in nonconvex settings.**

***Index Terms*—Decentralized nonconvex optimization, incremental gradient methods, variance reduction.**

I. Introduction

IN THIS article, we consider decentralized optimization problems that arise in many control and modern learning applications, where very large scale and geographically distributed nature of data precludes centralized storage and processing. The problem setup is based on n nodes communicating over a network modeled as a directed graph $\mathcal{G} := \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} := \{1, \ldots, n\}$ is the set of node indices and $\mathcal{E}$ is the collection of ordered pairs $(i, r), i, r \in \mathcal{V}$, such that node r sends information to node i. Each node i has access to a local, possibly private, collection of m smooth component functions $\{f_{i,j} : \mathbb{R}^p \to \mathbb{R}\}_{j=1}^m$ that are nonconvex. Each $f_{i,j}$ can be viewed as a cost incurred by the jth data sample at the ith node. The goal of the networked nodes is to agree on a first-order stationary point of the average of all component functions via local computation and communication at each node, i.e.,

$$\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),\ f_i(\mathbf{x}) := \frac{1}{m}\sum_{j=1}^{m} f_{i,j}(\mathbf{x}). \quad (1)$$

Decentralized optimization dates back to the seminal contribution of Tsitsiklis in 1980s [1], where the primary focus was on control and signal estimation problems, and has received a resurgence of interest recently due to its promise in large-scale control and machine learning applications [2], [3].

A. Related Work

First-order methods [4], [5] that rely mainly on gradient information are commonly used to approach large-scale optimization formulations like Problem (1). The works on decentralized first-order methods include several well-known papers on decentralized gradient descent (DGD) [6]–[9]. Although DGD-type methods are effective in homogeneous environments like data centers, its performance degrades significantly when data distributions across the nodes become heterogeneous [10]. Decentralized first-order methods that improve the performance of DGD include, e.g., EXTRA [11], Exact Diffusion/NIDS [12]–[14], DLM [15], and methods based on gradient tracking [16]–[23]; see also general primal-dual frameworks [24]–[27] that unify the aforementioned methods under certain conditions. Some structured formulations have also been considered recently, such as weak convexity [28], coupled constraints [29], and coordinate updates [30].

In decentralized batch gradient methods such as in [9], [11], and [18], each node i computes a full batch gradient $\sum_j \nabla f_{i,j}$ at each iteration. Clearly, batch gradient computation becomes expensive when the local batch size m is large and nodes have limited computational capabilities. Efficient stochastic methods, e.g., [12], [14], [21], [31]–[33], thus, use randomly sampled component gradients from each local batch; however, the convergence of these methods is typically slower compared with

Manuscript received 19 May 2021; accepted 24 September 2021. Date of publication 26 October 2021; date of current version 27 September 2022. The work of Ran Xin and Soummya Kar was supported in part by the National Science Foundation under Grant 1513936. The work of Usman A. Khan was supported in part by the National Science Foundation under Grant 1903972 and Grant 1935555. Recommended by Associate Editor Javad Lavaei. *(Corresponding author: Ran Xin.)*

Ran Xin and Soummya Kar are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15232 USA (e-mail: ranx@andrew.cmu.edu; soummyak@ece.cmu.edu).

Usman A. Khan is with the Department of Electrical and Computer Engineering, Tufts University, Medford, MA 02155 USA (e-mail: khan@ece.tufts.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2021.3122586.

Digital Object Identifier 10.1109/TAC.2021.3122586

their batch gradient counterparts due to the persistent noise incurred by the stochastic gradients. Toward fast convergence with stochastic gradients, popular variance reduction techniques, e.g., [34]–[38], have been adapted to the decentralized settings. For instance, algorithms in [10] and [39]–[43], including **GT-SAGA** [10], [41] considered in this article, are shown to achieve linear rate to the optimal solution for strongly convex problems; however, the applicability of these methods to nonconvex problems remains open. Recent works [44], [45] propose decentralized variance-reduced methods for nonconvex problems. These two methods, however, require periodic batch gradient evaluations across the nodes in addition to component gradient computations at each iteration; this *two-timescale hybrid scheme* imposes practical implementation challenges, such as periodic network synchronizations, especially over large-scale ad hoc networks.

B. Our Contributions

In this article, we analyze **GT-SAGA**, a *single-timescale randomized incremental* gradient method, originally proposed in [41] for strongly convex problems, and show that it achieves fast convergence in nonconvex settings. At the node level, **GT-SAGA** adopts a local SAGA-type [34], [46]–[49] randomized incremental approach to obtain variance-reduced estimates of local batch gradients, by leveraging historical component gradient information. At the network level, **GT-SAGA** employs a gradient tracking mechanism [16], [17] to fuse the local batch gradient estimates, obtained from the local SAGA procedures, to track the global batch gradient. These are the two building blocks that amount to the fast convergence and robustness to heterogeneous data in **GT-SAGA** for nonconvex problems. Compared with the existing two-timescale variance reduced methods [44], [45] for decentralized nonconvex optimization, **GT-SAGA** is single timescale and completely eliminates the need of batch gradient computations and periodic network synchronizations and is, hence, much easier to implement especially in ad hoc settings; see Remarks 1 and 2 for further discussion. The main technical contributions in this article are summarized as follows.

1) General Smooth Nonconvex Problems: For this problem class, we show the asymptotic convergence of **GT-SAGA** to a first-order stationary point in the almost sure and mean-squared sense. In a big-data regime, where the local batch size m is very large, **GT-SAGA** achieves a network topology-independent convergence rate, leading to a nonasymptotic linear speedup compared with the centralized SAGA [46] at a single node. In large-scale network regimes, i.e., when the number of the nodes and the network spectral gap inverse are relatively large compared to the local batch size m, we show that **GT-SAGA** outperforms the existing best known convergence rate [45]. We also introduce a measure of function heterogeneity across the nodes. Based on this measure, we show that the effect of function heterogeneity on the convergence rate of **GT-SAGA** appears in a fashion that is separable from the effects of local batch size and the network spectral gap. As a consequence, the effect of function heterogeneity often diminishes when the local batch size is large and/or the connectivity of the network is weak, demonstrating the robustness of **GT-SAGA** to function heterogeneity. In contrast, the state-of-the-art decentralized nonconvex variance-reduced method [45] does not achieve such separation and, hence, has worse convergence rate than **GT-SAGA** when the function heterogeneity is large and the network is weakly connected. These improvements are achieved by leveraging the conditional unbiasedness of SAGA estimators to obtain tighter bounds in the stochastic gradient tracking analysis. See Remarks 3–5 for details.

2) Smooth Nonconvex Problems Under the Global Polyak–Łojasiewicz (PL) Condition: For this problem class, we show that **GT-SAGA** achieves linear convergence to an optimal solution in expectation. To the best of our knowledge, this is the first linear rate result for decentralized variance-reduced methods under the PL condition, while the existing ones require strong convexity [39]–[43]. This generalization is nontrivial since the existing analysis essentially uses the unique optimal solution under strong convexity as a reference point to bound-related error terms, while the PL condition allows for the existence of multiple optimal solutions. In comparison with the existing linearly convergent decentralized deterministic batch gradient methods under the PL condition [21], [50], [51], **GT-SAGA** provably achieves faster linear rate, in terms of the component gradient computation complexity at each node, when the local batch size m is large, demonstrating the advantage of the employed variance reduction technique. In a big-data regime where m is large enough, we show that the linear rate of **GT-SAGA** becomes network topology independent. See Remarks 6 and 7 for details.

3) Convergence Analysis: We note that our analysis of SAGA-type variance reduction procedures is different from the existing ones [46], [52], which require careful constructions of Lyapunov functions. We avoid such delicate constructions by adopting a direct analysis approach, based on linear time-invariant (LTI) dynamics, which may be of independent interest and perhaps more readily extendable to other nonconvex problems. We note that the LTI dynamics-based analysis has mainly been used in convex problems in the existing literature of gradient tracking methods, e.g., [18], [20]. Somewhat surprisingly, a special case of our analysis, i.e., when the network is complete, provides the first linear rate result of the *original* centralized SAGA algorithm [34] under the PL condition. Indeed, the existing analysis [46], [52] is only applicable to a *modified* SAGA, which periodically restarts and samples its iterates; see Remark 8 for details. Finally, our analysis is also substantially different from that of the existing decentralized nonconvex variance-reduced methods [44], [45], where the variances of the stochastic gradients are bounded recursively, due to their hybrid nature. In contrast, we introduce a proper auxiliary sequence to bound the variance of **GT-SAGA**; see Sections IV-B and IV-D for details.

C. Outline of This Article and Notation

1) Outline of This Article: Section II presents the **GT-SAGA** algorithm and the main convergence results. Section III

illustrates the main theoretical results with the help of numerical simulations. Section IV presents the convergence analysis. Finally, Section V concludes this article.

2) Notation: The set of positive real numbers is denoted by $\mathbb{R}^+$. We use lowercase bold letters to denote vectors and uppercase bold letters to denote matrices. The matrix, $\mathbf{I}_d$ (respectively, $\mathbf{O}_d$), represents the $d \times d$ identity (respectively, zero matrix). The vector, $\mathbf{1}_d$ (respectively, $\mathbf{0}_d$), is the d-dimensional ones (respectively, zeros). The Kronecker product of two matrices is denoted by $\otimes$. We use $\|\cdot\|$ to denote the 2-norm of a vector or a matrix. For a matrix $\mathbf{X}$, we use $\rho(\mathbf{X})$ to denote its spectral radius and $\mathrm{diag}(\mathbf{X})$ as the diagonal matrix with the diagonal entries of $\mathbf{X}$. Matrix/vector inequalities are stated in the entrywise sense. We fix a proper probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for all random variables in question and $\mathbb{E}[\cdot]$ denotes the expectation; for an event $A \in \mathcal{F}$, its indicator is denoted as $\mathbb{1}_A$. We use $\sigma(\cdot)$ to denote the σ-algebra generated by the random variables and/or events. For two quantities $A, B \in \mathbb{R}^+$, we denote $A \lesssim B$ if there exists a universal c such that $A \leq cB$.

II. GT-SAGA Algorithm and Main Results

GT-SAGA [41], built upon local SAGA estimators [34] and global gradient tracking [16], [17], is formally presented in Algorithm 1. We refer the readers to [10] and [41] for detailed discussion on the development of **GT-SAGA**. In this article, we require for conciseness that all nodes start at the same point, i.e., $\mathbf{x}_i^0 = \overline{\mathbf{x}}^0, \forall i \in \mathcal{V}$. We emphasize that the complexity results of **GT-SAGA** established in this article hold, up to factors of universal constants, for the case where the nodes are initialized differently. We comment on the practical implementation aspects of **GT-SAGA** in comparison with the existing approaches in the following remarks.

Remark 1 (Single-timescale implementation): The existing decentralized variance-reduced methods for nonconvex optimization [44], [45] are based on a two-timescale double-loop implementation. Specifically, these methods, within each inner loop, run a fixed number of stochastic-gradient-type iterations, while, at each outer-loop iteration, a local batch gradient is computed at each node. This double-loop nature imposes challenges on the practical implementation of the two methods in [44] and [45]. First, periodic batch gradient computation incurs a synchronization overhead on the communication network and jeopardizes the actual wall-clock time when the networked nodes have largely heterogeneous computational capabilities. Second, these two methods have an additional parameter to tune, i.e., the length of each inner loop, other than the step size. Although this parameter may be chosen as m [45], this particular choice may not lead to the best performance in practice. In sharp contrast, **GT-SAGA** admits a simple single-timescale implementation since it only evaluates *one* randomly selected component gradient at each iteration. Furthermore, it only has *one* parameter to tune, i.e., the step size α. Therefore, **GT-SAGA** leads to significantly simpler implementation and tuning compared with the existing decentralized nonconvex variance-reduced methods [44], [45], especially over large-scale ad hoc networks. Finally, we note that **GT-SAGA** takes two successive communication rounds per iteration to transmit the state and gradient tracker, respectively, as in other gradient tracking-based methods, e.g., [20], [21], [44], [45].

Algorithm 1: **GT-SAGA** at Each Node i.

Require: $\mathbf{x}_i^0 = \overline{\mathbf{x}}^0 \in \mathbb{R}^p$; $\alpha \in \mathbb{R}^+$; $\{\underline{w}_{ir}\}_{r=1}^n$; $\mathbf{z}_{i,j}^0 = \mathbf{x}_i^0, \forall j \in \{1, \ldots, m\}$; $\mathbf{y}_i^0 = \mathbf{0}_p$; $\mathbf{g}_i^{-1} = \mathbf{0}_p$.

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

Select τ_i^k uniformly at random from $\{1, \ldots, m\}$;

Update the local stochastic gradient estimator:

$$\mathbf{g}_i^k = \nabla f_{i,\tau_i^k}\left(\mathbf{x}_i^k\right) - \nabla f_{i,\tau_i^k}\left(\mathbf{z}_{i,\tau_i^k}^k\right) + \frac{1}{m}\sum_{j=1}^m \nabla f_{i,j}\left(\mathbf{z}_{i,j}^k\right);$$

Update the local gradient tracker:

$$\mathbf{y}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\left(\mathbf{y}_r^k + \mathbf{g}_r^k - \mathbf{g}_r^{k-1}\right);$$

Update the local estimate of the solution:

$$\mathbf{x}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\left(\mathbf{x}_r^k - \alpha \mathbf{y}_r^{k+1}\right);$$

Select s_i^k uniformly at random from $\{1, \ldots, m\}$;

Set $\mathbf{z}_{i,j}^{k+1} = \mathbf{x}_i^k$ for $j = s_i^k$; $\mathbf{z}_{i,j}^{k+1} = \mathbf{z}_{i,j}^k$ for $j \neq s_i^k$;

end for

Remark 2 (Storage requirement): To practically implement **GT-SAGA**, each node i needs to retain a gradient table $\{\nabla f_{i,j}(\mathbf{z}_{i,j}^k)\}_{j=1}^m$ of size $m \times p$ in general, which may be expensive. However, for certain structured problems, the size of the gradient table can be largely reduced [34]. For instance, in nonconvex generalized linear models [53], each component function takes the form $f_{i,j}(\mathbf{x}) = \ell(\mathbf{x}^\top \boldsymbol{\theta}_{i,j})$, where $\ell : \mathbb{R} \to \mathbb{R}$ is a nonconvex loss and $\boldsymbol{\theta}_{i,j}$ is the jth data at the ith node. Clearly, $\nabla f_{i,j}(\mathbf{x}) = \ell'(\mathbf{x}^\top \boldsymbol{\theta}_{i,j})\boldsymbol{\theta}_{i,j}$, and thus, each node i only needs to retain $\{\ell'(\mathbf{z}_{i,j}^\top \boldsymbol{\theta}_{i,j})\}_{j=1}^m$, a gradient table of size $m \times 1$, since the data samples $\{\boldsymbol{\theta}_{i,j}\}_{j=1}^m$ are already stored locally. See Section III-A for numerical experiments based on one such example.

We now enlist the assumptions of interest in this article.

Assumption 1: The family $\{\tau_i^k, s_i^k : i \in \mathcal{V}, k \geq 0\}$ of random variables in Algorithm 1 is independent.

Assumption 1 is standard in stochastic gradient methods. Specifically, the index s_i^k used for updating the gradient table $\{\nabla f_{i,j}(\mathbf{z}_{i,j}^k)\}_{j=1}^m$ is sampled independently from the index τ_i^k used for updating the local SAGA estimator $\mathbf{g}_i^k$ per node per iteration. This independence requirement is straightforward to implement and is often posed to simplify the analysis of SAGA-type estimators for nonconvex problems [46], [52]; see Section IV-D for analysis based on this assumption.

Assumption 2: Each component function $f_{i,j} : \mathbb{R}^p \to \mathbb{R}$ is differentiable and L-smooth, i.e., there exists $L > 0$, such that $\|\nabla f_{i,j}(\mathbf{x}) - \nabla f_{i,j}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p \forall i \in \mathcal{V}$ $\forall j \in \{1, \ldots, m\}$. Moreover, the global function F is bounded below, i.e., $F^* := \inf_{\mathbf{x} \in \mathbb{R}^p} F(\mathbf{x}) > -\infty$.

Under Assumption 2, the local batch functions $\{f_i\}_{i=1}^n$ and the global function F are L-smooth. We note that L stated in Assumption 2 is essentially the maximum of the smoothness parameters of all component functions. We further consider the case when *the global* F additionally satisfies the PL condition described as follows.

Assumption 3: The global function $F:\mathbb{R}^p\to\mathbb{R}$ satisfies $2\mu(F(\mathbf{x})-F^*)\leq\|\nabla F(\mathbf{x})\|^2\ \forall\mathbf{x}\in\mathbb{R}^p$, for some $\mu>0$.

The PL condition, originally introduced in [54], generalizes the notion of strong convexity to nonconvex functions; see [55] for more discussion. When Assumption 3 holds, we denote $\kappa:=\frac{L}{\mu}\geq 1$, which may be interpreted as the condition number of F. Note that the PL condition implies that every stationary point $\mathbf{x}^*$ of F, such that $\nabla F(\mathbf{x}^*)=\mathbf{0}_p$, is a global minimizer of F, while F is not necessarily convex.

Assumption 4: The weight matrix $\underline{\mathbf{W}}=\{\underline{w}_{ir}\}\in\mathbb{R}^{n\times n}$ of the network is primitive and doubly stochastic, i.e., $\underline{\mathbf{W}}\mathbf{1}_n=\mathbf{1}_n$, $\mathbf{1}_n^\top\underline{\mathbf{W}}=\mathbf{1}_n^\top$, and $\lambda:=\lambda_2(\underline{\mathbf{W}})\in[0,1)$, where $\lambda_2(\mathbf{W})$ is the second largest singular value of $\underline{\mathbf{W}}$.

Weight matrices that satisfy Assumption 4 may be designed for strongly connected, weight-balanced, directed networks or for connected undirected networks. We next discuss the performance metrics of **GT-SAGA** for different problem classes. For general smooth nonconvex problems, we define the iteration complexity of **GT-SAGA** as the minimum number of iterations required to achieve an ϵ-accurate stationary point of the global function F, i.e.,

$$\inf\left\{K:\frac{1}{n}\sum_{i=1}^n\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2\right]\leq\epsilon\right\}.$$

When the global function F further satisfies the PL condition, we define the iteration complexity of **GT-SAGA** as

$$\inf\left\{k:\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(F(\mathbf{x}_i^k)-F^*\right)\right]\leq\epsilon\right\}.$$

These are standard metrics for decentralized stochastic nonconvex optimization methods [2], [21], [44], [45]. We refer the iteration complexity as the convergence rate metric of **GT-SAGA**, since it is the same as the communication and component gradient computation complexity at each node. We are now ready to state the main results of **GT-SAGA** in the next subsections and discuss their implications.

A. General Smooth Nonconvex Functions

In this subsection, we present the main convergence results of **GT-SAGA** for general smooth nonconvex functions.

Theorem 1: Let Assumptions 1, 2, and 4 hold. If the step size α of **GT-SAGA** satisfies $0<\alpha\leq\overline{\alpha}_1$, where

$$\overline{\alpha}_1:=\min\left\{\frac{(1-\lambda^2)^2}{48\lambda},\frac{2n^{1/3}}{13m^{2/3}},\frac{1}{2},\frac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}\right\}\frac{1}{L}$$

then all nodes asymptotically agree on a stationary point in both mean-squared and almost sure sense, i.e., $\forall i,r\in\mathcal{V}$

$$\mathbb{P}\left(\lim_{k\to\infty}\|\mathbf{x}_i^k-\mathbf{x}_r^k\|=0\right)=1,\quad\lim_{k\to\infty}\mathbb{E}\left[\|\mathbf{x}_i^k-\mathbf{x}_r^k\|^2\right]=0$$

$$\mathbb{P}\left(\lim_{k\to\infty}\|\nabla F(\mathbf{x}_i^k)\|=0\right)=1,\quad\lim_{k\to\infty}\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2\right]=0.$$

Moreover, if $\alpha=\overline{\alpha}_1$, **GT-SAGA** achieves an ϵ-accurate stationary point in

$$\mathcal{O}\left(\frac{EL(F(\overline{\mathbf{x}}^0)-F^*)}{\epsilon}+\frac{\lambda^2(1-\lambda^2)\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n\epsilon}\right)\quad(2)$$

iterations, where E is given by

$$E:=\max\left\{\frac{m^{2/3}}{n^{1/3}},1,\frac{\lambda}{(1-\lambda)^2},\frac{\lambda^{1/2}m^{1/2}}{(1-\lambda)^{3/4}}\right\}$$

and $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2=\sum_{i=1}^n\|\nabla f_i(\overline{\mathbf{x}}^0)\|^2$.

Theorem 1 is formally proved in Section IV-F. We discuss its implications in the following remarks.

Remark 3 (Effect of the function heterogeneity): We note that $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n$ in the second term of (2) can be viewed as a measure of heterogeneity among the local functions. In particular, when all local functions are identical such that $f_i=f_r=F,\forall i,r\in\mathcal{V}$, this term diminishes, i.e., it can be shown that $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n=\|\nabla F(\overline{\mathbf{x}}^0)\|^2\leq 2L(F(\overline{\mathbf{x}}^0)-F^*)$. On the other hand, when the local functions are significantly different, $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n$ can be fairly large compared with $L(F(\overline{\mathbf{x}}^0)-F^*)$. Based on Theorem 1, it is important to note that the effect of the function heterogeneity $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n$ on the convergence rate of **GT-SAGA** is decoupled from E, the effect of the local batch size m and the network spectral gap $1-\lambda$. It is further interesting to observe that the heterogeneity effect diminishes when the network is sufficiently either well connected or weakly connected. In other words, the function heterogeneity effect is dominated by the network effect in these two extreme cases of interest.

We next view Theorem 1 in two different regimes.

Remark 4 (Big-data regime): We first consider a big-data regime that is often applicable in data centers, where the local batch size m is relatively large compared with the network spectral gap inverse $(1-\lambda)^{-1}$ and the number of the nodes n. In particular, if m large enough such that

$$\max\left\{1,\frac{\lambda}{(1-\lambda)^2},\frac{\lambda^{1/2}m^{1/2}}{(1-\lambda)^{3/4}}\right\}\lesssim\frac{m^{2/3}}{n^{1/3}}\quad(3)$$

Theorem 1 results into an iteration complexity of

$$\mathcal{O}\left(\frac{m^{2/3}L(F(\overline{\mathbf{x}}^0)-F^*)}{n^{1/3}\epsilon}+\frac{\lambda^2(1-\lambda^2)\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n\epsilon}\right).\quad(4)$$

We emphasize that the first term in (4) matches the iteration complexity of the centralized SAGA with a minibatch size n [46], as **GT-SAGA** computes n component gradients across the nodes in parallel at each iteration. We note that under the big-data condition (3), it typically holds that $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n\lesssim m^{2/3}L(F(\overline{\mathbf{x}}^0)-F^*)/n^{1/3}$, i.e., the first term dominates the second term in (4). Therefore, **GT-SAGA** in this regime achieves a nonasymptotic linear speedup, i.e., the total number of component gradient computations required at each node to achieve an ϵ-accurate stationary point is reduced by a factor of $1/n$, compared with the centralized minibatch SAGA that operates on a single machine.

Remark 5 (Large-scale network regime): We now consider the case where a large number of nodes are weakly connected, a scenario that commonly appears in sensor networks, robotic swarms, and ad hoc Internet of Things networks. In this case, the number of the nodes n and the network spectral gap inverse $(1-\lambda)^{-1}$ are relatively large in comparison with the local batch size m. In particular, if

$$\max\left\{1, \frac{m^{2/3}}{n^{1/3}}, \frac{\lambda^{1/2}m^{1/2}}{(1-\lambda)^{3/4}}\right\} \lesssim \frac{\lambda}{(1-\lambda)^2} \tag{5}$$

then the component gradient computation complexity at each node of **GT-SAGA**, according to Theorem 1, becomes

$$\mathcal{O}\left(\frac{\lambda L(F(\overline{\mathbf{x}}^0)-F^*)}{(1-\lambda)^2\epsilon} + \frac{\lambda^2(1-\lambda^2)\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n\epsilon}\right). \tag{6}$$

We note that the component gradient complexity at each node of GT-SARAH [45], the state-of-the-art decentralized nonconvex variance-reduced method, in this regime is

$$\mathcal{O}\left(\frac{\lambda}{(1-\lambda)^2\epsilon}\left(L(F(\overline{\mathbf{x}}^0)-F^*) + \frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}\right)\right). \tag{7}$$

Comparing (7) to (6), we observe that GT-SARAH, unlike **GT-SAGA**, does not achieve a separation between the dependence of the network spectral gap $1-\lambda$ and the function heterogeneity measure $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n$ on the convergence rate. We, hence, conclude that **GT-SAGA** outperforms GT-SARAH if the network is weakly connected and the local functions are largely heterogeneous, i.e., when $1-\lambda$ is small and $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n$ is large. Moreover, we recall from Remark 1 that **GT-SAGA** is single timescale and, thus, is much easier to implement than the two-timescale GT-SARAH over large-scale networks. We also emphasize that the storage requirement of **GT-SAGA** in this regime is significantly relaxed since the data samples are distributed across a large network, leading to a small local batch size m at each node.

B. Global PL Condition

Theorem 2: Let Assumptions 1–4 hold. If the step size α of **GT-SAGA** satisfies $0<\alpha\leq\overline{\alpha}_2$, where

$$\overline{\alpha}_2 := \min\left\{\frac{(1-\lambda^2)^2}{55\lambda L}, \frac{1-\lambda^2}{13\lambda\kappa^{1/4}L}, \frac{(1-\lambda^2)^3}{388\lambda^2 nL}, \frac{n^{1/3}}{10.5m^{2/3}\kappa^{1/3}L}, \frac{1}{36L}, \frac{1-\lambda^2}{2\mu}, \frac{1}{4m\mu}\right\}$$

then all nodes converge linearly at the rate $\mathcal{O}((1-\mu\alpha)^k)$ to a global minimizer of F. In particular, if $\alpha=\overline{\alpha}_2$, then all nodes agree on an ϵ-accurate global minimizer in

$$\mathcal{O}\left(\max\{Q_{\text{opt}}, Q_{\text{net}}\}\log\frac{1}{\epsilon}\right)$$

iterations, where Q_{opt} and Q_{net} are given, respectively, by

$$Q_{\text{opt}} := \max\left\{\frac{m^{2/3}\kappa^{4/3}}{n^{1/3}}, \kappa, m\right\}$$

$$Q_{\text{net}} := \max\left\{\frac{\lambda\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{\lambda^2 n\kappa}{(1-\lambda)^3}, \frac{1}{1-\lambda}\right\}.$$

Theorem 2 is formally proved in Section IV-G. The following remarks discuss a few key aspects of it.

Remark 6 (Linear rate under the global PL condition): Theorem 2 shows that **GT-SAGA** linearly converges to an optimal solution when the global F additionally satisfies the PL condition. This is the first linear rate result for decentralized variance-reduced methods under the PL condition, while the existing ones require strong convexity, e.g., [39]–[43]. A notable feature of the linear rate in Theorem 2 is that the effects of the local batch size m and the network spectral gap $1-\lambda$ are decoupled. Hence, in a big-data regime where the local batch size m is sufficiently large such that $Q_{\text{net}}\lesssim Q_{\text{opt}}$, **GT-SAGA** achieves a network topology-independent rate of $\mathcal{O}(Q_{\text{opt}}\log\frac{1}{\epsilon})$. In addition, we note that Theorem 2 implies the linear rate of **GT-SAGA** in the almost sure sense under the PL condition, by Chebyshev's inequality and the Borel–Cantelli lemma; see [41, Lemma 7] for details.

Remark 7 (Comparison with other decentralized gradient methods): When the local batch size m is relatively large, the linear rate of **GT-SAGA** improves that of the existing decentralized batch gradient methods [21], [50], [51] under the PL condition in terms of the component gradient computation complexity. Moreover, decentralized online stochastic gradient methods, e.g., [21], [56], only exhibit sublinear rate under the PL condition due to the persistent variances of the stochastic gradients. Therefore, **GT-SAGA** achieves faster convergence under the PL condition compared with the existing decentralized methods, demonstrating the advantage of the employed SAGA variance reduction scheme that is able to exploit the finite-sum structure of local functions.

Remark 8 (Improved convergence results for the centralized minibatch SAGA): When $\lambda=0$, i.e., when the underlying network is a complete graph whose weight matrix can be easily chosen as $\underline{\mathbf{W}}=\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, **GT-SAGA** reduces to the centralized minibatch SAGA and achieves the linear rate of $\mathcal{O}(Q_{\text{opt}}\log\frac{1}{\epsilon})$. Hence, a special case of Theorem 2, i.e, $\lambda=0$, provides the first linear rate result under the PL condition for the centralized SAGA. Indeed, the existing linear rate results [46], [52] under the PL condition are only applicable to a modified SAGA that *periodically restarts* $\mathcal{O}(\log\frac{1}{\epsilon})$ times with the output of each cycle being selected randomly from the past iterates in this cycle. This procedure is not feasible particularly in decentralized settings. In contrast, the linear rate shown Theorem 2 is on the *last iterate* of the original SAGA without periodic restarting and sampling.

III. Numerical Experiments

In this section, we present numerical simulations to illustrate our main theoretical results. The network topologies of interest are undirected ring, undirected 2D-grid, directed exponential, undirected geometric, and complete graphs; see [3], [41], and [57] for details of these graphs. The doubly stochastic weights are set to be equal for the ring and exponential graphs and are generated by the lazy Metropolis rule for the grid and geometric graphs. We manually optimize the parameters of all algorithms in all experiments for their best performance.

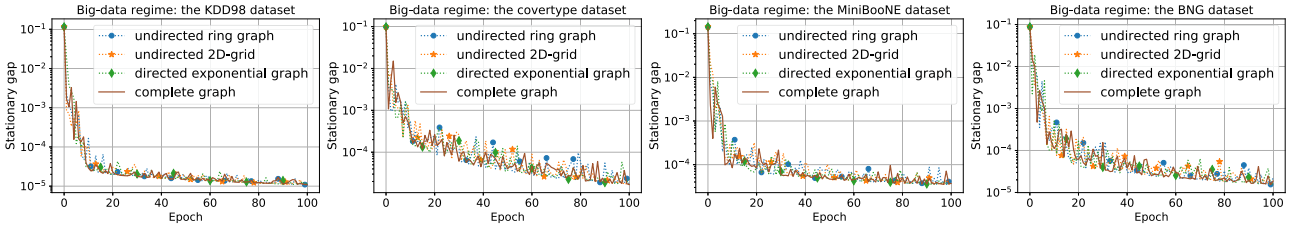

Fig. 1. Big-data regime: the network topology-independent convergence rate of **GT-SAGA** on the KDD98, covertype, MiniBooNE, and BNG (sonar) datasets.

TABLE I
DATASETS USED IN NUMERICAL EXPERIMENTS

Dataset	train ($N = nm$)	dimension (p)
nomao	30,000	119
a9a	48,800	124
w8a	60,000	300
KDD98	80,000	478
covertype	100,000	55
MiniBooNE	100,000	51
BNG(sonar)	100,000	61

[Online]. Available: https://www.openml.org/

A. Nonconvex Binary Classification

In this subsection, we consider a decentralized nonconvex generalized linear model for binary classification. In view of Problem (1), each component cost $f_{i,j}$ is defined as [53]

$$f_{i,j}(\mathbf{x}) := \ell\left(\xi_{i,j}\mathbf{x}^\top\boldsymbol{\theta}_{i,j}\right), \ell(u) := \left(1 - \frac{1}{1+\exp(-u)}\right)^2$$

where $\boldsymbol{\theta}_{i,j} \in \mathbb{R}^p$ is the jth data vector at the ith node, $\xi_{i,j} \in \{-1, +1\}$ is the label of $\boldsymbol{\theta}_{i,j}$, and $\ell : \mathbb{R} \to \mathbb{R}$ is a $\frac{4}{3}$-smooth nonconvex loss. We normalize each data to be $\|\boldsymbol{\theta}_{i,j}\| = 1, \forall i, j$. Since $\nabla^2 f_{i,j}(\mathbf{x}) = \ell''(\xi_{i,j}\mathbf{x}^\top\boldsymbol{\theta}_{i,j})\boldsymbol{\theta}_{i,j}\boldsymbol{\theta}_{i,j}^\top$, it can be verified that $\|\nabla^2 f_{i,j}(\mathbf{x})\| = |\ell''(\xi_{i,j}\mathbf{x}^\top\boldsymbol{\theta}_{i,j})| \leq \frac{4}{3}$. Hence, each component cost $f_{i,j}$ is nonconvex and $\frac{4}{3}$-smooth. We measure the performance of the algorithms in question in terms of the decrease of the stationary gap $\|\nabla F(\overline{\mathbf{x}})\|$ versus epochs, where $\overline{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ for $\mathbf{x}_i$ being the model at node i and each epoch represents m component gradient evaluations at each node. All nodes start from a vector randomly generated from the standard Gaussian distribution. The statistics of the datasets used in the experiments are provided in Table I.

1) Big-Data Regime: We first test the convergence behavior of **GT-SAGA** in the big-data regime by uniformly distributing the KDD98, covertype, MiniBooNE, and BNG (sonar) datasets over a network of $n = 20$ nodes. We consider four different network topologies with decreasing sparsity, i.e., the undirected ring, undirected 2D-grid, directed exponential, and complete graph; their corresponding second largest singular values of the weight matrices are $\lambda = 0.98, 0.97, 0.6, 0$, respectively. It can be verified that the big-data condition (3) holds. The experimental results are shown in Fig. 1, where we observe that the convergence rate of **GT-SAGA** is independent of the network topology in this big-data regime; see Remark 4.

2) Large-Scale Network Regime: We next compare the performance of **GT-SAGA** with DSGD [2] and GT-SARAH [45] in the large-scale network regime. To this aim, we generate a sparse geometric graph of $n = 200$ nodes with $\lambda \approx 0.99$ and uniformly distribute the nomao, a9a, w8a, and BNG (sonar) datasets over the nodes. It can be verified that the large-scale network condition (5) holds. The numerical results are presented in Fig. 2: the first three plots show that **GT-SAGA** achieves the best performance among the algorithms in comparison, while the last plot shows that the convergence rate of **GT-SAGA** is dependent on the network topology in this large-scale network regime; see Remark 5.

3) Robustness to Heterogeneous Data: We now make the data distributions across the nodes significantly heterogeneous by letting each node only have data samples of one label, so that no node can train a valid classification model only from its local data. We compare the performance of **GT-SAGA** under heterogeneous and homogeneous distribution of the nomao dataset. We consider a well-connected graph, i.e., the 20-node exponential graph, and a weakly connected graph, i.e., the 200-node geometric graph. The numerical results are shown in Fig. 3, where we observe that the convergence rate of **GT-SAGA** is not affected by the data heterogeneity over both graphs; see Remark 3.

B. Synthetic Functions That Satisfy the PL Condition

Finally, we verify the linear rate of **GT-SAGA** when the global function F satisfies the PL condition. Specifically, we choose each component function $f_{i,j} : \mathbb{R} \to \mathbb{R}$ as

$$f_{i,j}(x) = x^2 + 3\sin^2(x) + a_{i,j}\cos(x) + b_{i,j}x$$

where $\sum_{i=1}^n\sum_{j=1}^m a_{i,j} = 0$ and $\sum_{i=1}^n\sum_{j=1}^m b_{i,j} = 0$ such that $a_{i,j} \neq 0$, $b_{i,j} \neq 0$, $\forall i, j$. This formulation, hence, leads to the global function $F(x) = x^2 + 3\sin^2(x)$. It can be verified that F is nonconvex and satisfies the PL condition [55]. Note that each $f_{i,j}$ is nonlinear and highly deviated from F; see the last three plots in Fig. 4 for a comparison of local and global geometries. We use the 20-node exponential graph and set $m = 5$. It can be observed from the first plot in Fig. 4 that **GT-SAGA** achieves linear rate to the optimal solution, while DSGD converges to an inexact solution; see Remark 7.

IV. CONVERGENCE ANALYSIS

In this section, we present the convergence analysis of **GT-SAGA**, i.e., the sublinear convergence for general smooth nonconvex functions and the linear convergence when the global

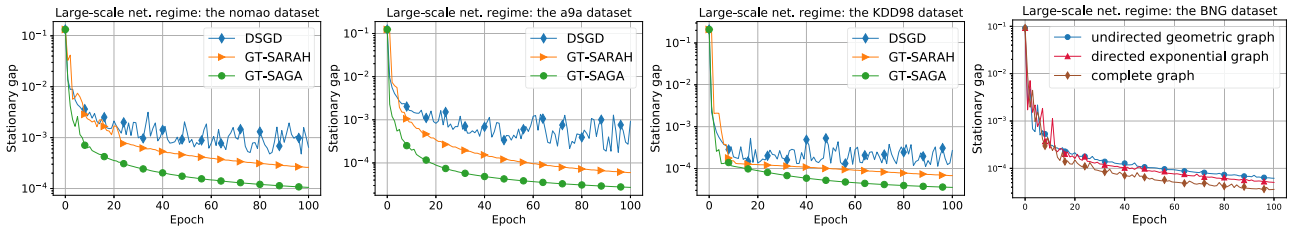

Fig. 2. Large-scale network regime: the first three plots present the performance comparison between **GT-SAGA**, DSGD, and GT-SARAH on the nomao, a9a, and KDD98 datasets; the last plot presents the performance of **GT-SAGA** over different graph topologies in this regime on the BNG (sonar) dataset.

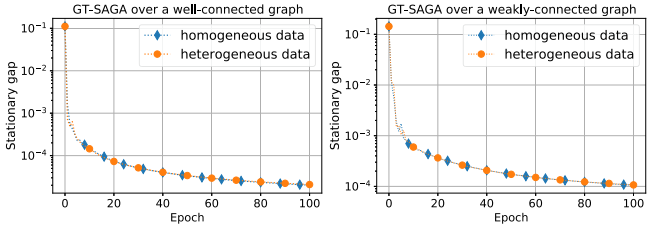

Fig. 3. Robustness of **GT-SAGA** to heterogeneous data over well connected and weakly connected graphs on the nomao dataset.

function F additionally satisfies the PL condition. Throughout this section, we assume that Assumptions 1, 2, and 4 hold without explicitly stating them; we only assume that Assumption 3 hold in Section IV-G. In Sections IV-B–IV-E, we establish key relationships between several important quantities, based on which the proofs of Theorems 1 and 2 are derived in Sections IV-F and IV-G, respectively. We start by presenting some preliminary facts.

A. Preliminaries

GT-SAGA can be written in the following form: $\forall k \geq 0$

$$\mathbf{y}^{k+1} = \mathbf{W}\left(\mathbf{y}^k + \mathbf{g}^k - \mathbf{g}^{k-1}\right) \tag{8a}$$

$$\mathbf{x}^{k+1} = \mathbf{W}\left(\mathbf{x}^k - \alpha \mathbf{y}^{k+1}\right) \tag{8b}$$

where $\mathbf{x}^k, \mathbf{y}^k$, and $\mathbf{g}^k$ are random vectors in $\mathbb{R}^{np}$ that concatenate all local states $\{\mathbf{x}_i^k\}_{i=1}^n$, gradient trackers $\{\mathbf{y}_i^k\}_{i=1}^n$, and local SAGA estimators $\{\mathbf{g}_i^k\}_{i=1}^n$, respectively, and $\mathbf{W} = \underline{\mathbf{W}} \otimes \mathbf{I}_p$. We denote $\mathcal{F}^k$ as the filtration of **GT-SAGA**, i.e., $\forall k \geq 1$

$$\mathcal{F}^k := \sigma\left(\left\{\tau_i^t, s_i^t : i \in \mathcal{V}, t \leq k-1\right\}\right), \qquad \mathcal{F}^0 := \{\phi, \Omega\}$$

where ϕ is the empty set. It can be verified that $\mathbf{x}^k$, $\mathbf{y}^k$, and $\mathbf{z}_{i,j}^k, \forall i, j$, are $\mathcal{F}^k$-measurable and $\mathbf{g}^k$ is $\mathcal{F}^{k+1}$-measurable for all $k \geq 0$. We use $\mathbb{E}[\cdot|\mathcal{F}^k]$ to denote the conditional expectation with respect to $\mathcal{F}^k$. For the ease of exposition, we introduce the following quantities:

$$\mathbf{J} := \left(\mathbf{1}_n \mathbf{1}_n^\top / n\right) \otimes \mathbf{I}_p$$

$$\nabla \mathbf{f}(\mathbf{x}^k) = \left[\nabla f_1(\mathbf{x}_1^k)^\top, \ldots, \nabla f_n(\mathbf{x}_n^k)^\top\right]^\top$$

$$\overline{\nabla \mathbf{f}}(\mathbf{x}^k) = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \nabla \mathbf{f}(\mathbf{x}^k), \qquad \overline{\mathbf{x}}^k = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \mathbf{x}^k$$

$$\overline{\mathbf{y}}^k = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \mathbf{y}^k, \qquad \overline{\mathbf{g}}^k = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \mathbf{g}^k.$$

We assume that $\overline{\mathbf{x}}^0 \in \mathbb{R}^p$ is constant, and hence, all random variables generated by **GT-SAGA** have bounded second moment. The following lemma lists several well-known facts in the context of gradient tracking and SAGA estimators, which may be found in [17], [18], [34], [41], and [54].

Lemma 1: The following relationships hold.

a) $\forall \mathbf{x} \in \mathbb{R}^{np}$, $\|\mathbf{W}\mathbf{x} - \mathbf{J}\mathbf{x}\| \leq \lambda \|\mathbf{x} - \mathbf{J}\mathbf{x}\|$.

b) $\overline{\mathbf{y}}^{k+1} = \overline{\mathbf{g}}^k, \forall k \geq 0$.

c) $\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k) - \nabla F(\overline{\mathbf{x}}^k)\|^2 \leq \frac{L^2}{n} \|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2, \forall k \geq 0$.

d) $\mathbb{E}[\mathbf{g}_i^k | \mathcal{F}^k] = \nabla f_i(\mathbf{x}_i^k), \forall i \in \mathcal{V}, \forall k \geq 0$.

e) $\|\nabla F(\mathbf{x})\|^2 \leq 2L(F(\mathbf{x}) - F^*)$.

Note that Lemma 1(e) is a consequence of the L-smoothness of the global function F and is only used in Section IV-G, while other statements in Lemma 1 are frequently utilized throughout the analysis. The next lemma states some standard inequalities on the network consensus error [21], [41].

Lemma 2: The following inequality holds: $k \geq 0$

$$\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\|^2 \leq \tfrac{1+\lambda^2}{2} \|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \tfrac{2\alpha^2\lambda^2}{1-\lambda^2} \|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 \tag{9}$$

$$\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\|^2 \leq 2\lambda^2 \|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 2\alpha^2\lambda^2 \|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 \tag{10}$$

$$\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\| \leq \lambda \|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\| + \alpha\lambda \|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|. \tag{11}$$

B. Bounds on the Variance of Local SAGA Estimators

In this subsection, we bound the variance of the local SAGA gradient estimators $\mathbf{g}_i^k$s. For analysis purposes, we construct two auxiliary $\mathcal{F}^k$-adapted sequences: $\forall i \in \mathcal{V}, \forall k \geq 0$

$$t_i^k := \frac{1}{m} \sum_{j=1}^m \|\overline{\mathbf{x}}^k - \mathbf{z}_{i,j}^k\|^2, \qquad t^k := \frac{1}{n} \sum_{i=1}^n t_i^k.$$

These two sequences are essential in the convergence analysis. We note that t^k measures the average distance between the mean state $\overline{\mathbf{x}}^k$ of the networked nodes and the latest iterates $\mathbf{z}_{i,j}^k$s, where the component gradients were computed at iteration k in the gradient tables. Intuitively, t^k goes to 0 as all nodes in **GT-SAGA** reach consensus on a stationary point. We will establish a contraction argument in t^k in Section IV-D. In the following lemma, we show that the variance of $\mathbf{g}_i^k$ may be bounded by the network consensus error and t^k.

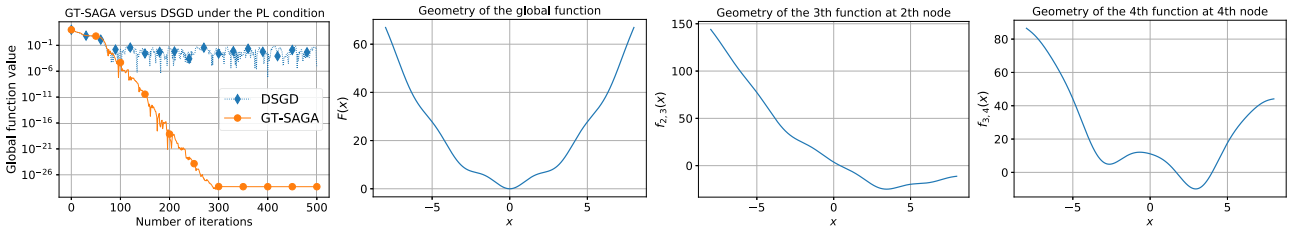

Fig. 4. PL condition: the first plot presents the performance comparison between **GT-SAGA** and DSGD when the global function satisfies the PL condition; the last three plots present the geometry comparison of the global and local component functions.

Lemma 3: The following inequality holds: $\forall k \geq 0$

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2|\mathcal{F}^k\right] \leq 2L^2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 2nL^2t^k \tag{12}$$

$$\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right] \leq \frac{2L^2}{n^2}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \frac{2L^2}{n}t^k + \|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2. \tag{13}$$

Proof: We denote $\widehat{\nabla}_i^k := \nabla f_{i,\tau_i^k}(\mathbf{x}_i^k) - \nabla f_{i,\tau_i^k}(\mathbf{z}_{i,\tau_i^k}^k), \forall i \in \mathcal{V}$, $\forall k \geq 0$, for the ease of exposition. We first observe from Algorithm 1 that $\forall k \geq 0, \forall i \in \mathcal{V}$

$$\mathbb{E}\left[\widehat{\nabla}_i^k|\mathcal{F}^k\right] = \nabla f_i(\mathbf{x}_i^k) - \frac{1}{m}\sum_{j=1}^m \nabla f_{i,j}\left(\mathbf{z}_{i,j}^k\right). \tag{14}$$

In light of (14), we bound the variance of $\mathbf{g}_i^k$ in the following: $\forall k \geq 0, \forall i \in \mathcal{V}$

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k)\|^2|\mathcal{F}^k\right] \\
&= \mathbb{E}\left[\|\widehat{\nabla}_i^k - \mathbb{E}\left[\widehat{\nabla}_i^k|\mathcal{F}^k\right]\|^2|\mathcal{F}^k\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\|\widehat{\nabla}_i^k\|^2|\mathcal{F}^k\right] \\
&= \mathbb{E}\left[\sum_{j=1}^m \mathbb{1}_{\{\tau_i^k=j\}}\left\|\nabla f_{i,j}\left(\mathbf{x}_i^k\right) - \nabla f_{i,j}\left(\mathbf{z}_{i,j}^k\right)\right\|^2|\mathcal{F}^k\right] \\
&\overset{(ii)}{=} \frac{1}{m}\sum_{j=1}^m\left\|\nabla f_{i,j}\left(\mathbf{x}_i^k\right) - \nabla f_{i,j}\left(\mathbf{z}_{i,j}^k\right)\right\|^2 \\
&\overset{(iii)}{\leq} \frac{L^2}{m}\sum_{j=1}^m\left\|\mathbf{x}_i^k - \mathbf{z}_{i,j}^k\right\|^2 \\
&\leq 2L^2\left\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\right\|^2 + 2L^2t_i^k.
\end{aligned} \tag{15}$$

where (i) is the conditional variance decomposition, (ii) uses that $\|\nabla f_{i,j}(\mathbf{x}_i^k) - \nabla f_{i,j}(\mathbf{z}_{i,j}^k)\|^2$ is $\mathcal{F}^k$-measurable and that τ_i^k is independent of $\mathcal{F}^k$, and (iii) uses the L-smoothness of each $f_{i,j}$. Summing up (15) over i from 1 to n gives (12). Toward (13), we have: $\forall k \geq 0$

$$\begin{aligned}
\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right] &\overset{(i)}{=} \mathbb{E}\left[\|\overline{\mathbf{g}}^k - \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k\right] + \|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 \\
&\overset{(ii)}{=} \frac{1}{n^2}\mathbb{E}\left[\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2|\mathcal{F}^k\right] + \|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2
\end{aligned} \tag{16}$$

where (i) uses that $\mathbb{E}[\overline{\mathbf{g}}^k|\mathcal{F}^k] = \overline{\nabla\mathbf{f}}(\mathbf{x}^k)$ and that $\overline{\nabla\mathbf{f}}(\mathbf{x}^k)$ is $\mathcal{F}^k$-measurable, while (ii) uses that, whenever $i \neq j$, $\mathbb{E}[\langle\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k), \mathbf{g}_j^k - \nabla f_j(\mathbf{x}_j^k)\rangle|\mathcal{F}^k] = 0$, since τ_i^k is independent of $\sigma(\sigma(\tau_j^k), \mathcal{F}^k)$ and $\mathbb{E}[\mathbf{g}^k|\mathcal{F}^k] = \nabla\mathbf{f}(\mathbf{x}^k)$. The proof follows by applying (12) to (16). ■

C. Descent Inequality

In this subsection, we provide a key descent inequality that characterizes the expected decrease of the global function value at each iteration of **GT-SAGA**.

Lemma 4: If $0 < \alpha \leq \frac{1}{2L}$, then $\forall k \geq 0$

$$\begin{aligned}
\mathbb{E}\left[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k\right] \leq F(\overline{\mathbf{x}}^k) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k)\|^2 - \frac{\alpha}{4}\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 \\
+ \frac{\alpha L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \frac{\alpha^2L^3}{n}t^k.
\end{aligned}$$

Proof: Since F is L-smooth, we have [4]: $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle\nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{17}$$

We multiply (8b) by $\frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)$ and use Lemma 1(b) to obtain $\overline{\mathbf{x}}^{k+1} = \overline{\mathbf{x}}^k - \alpha\overline{\mathbf{y}}^{k+1} = \overline{\mathbf{x}}^k - \alpha\overline{\mathbf{g}}^k, \forall k \geq 0$. Setting $\mathbf{y} = \overline{\mathbf{x}}^{k+1}$ and $\mathbf{x} = \overline{\mathbf{x}}^k$ in (17) obtains: $\forall k \geq 0$

$$F(\overline{\mathbf{x}}^{k+1}) \leq F(\overline{\mathbf{x}}^k) - \alpha\langle\nabla F(\overline{\mathbf{x}}^k), \overline{\mathbf{g}}^k\rangle + \frac{\alpha^2L}{2}\|\overline{\mathbf{g}}^k\|^2. \tag{18}$$

Conditioning (18) with respect to $\mathcal{F}^k$, since $\nabla F(\overline{\mathbf{x}}^k)$ is $\mathcal{F}^k$-measurable, we have

$$\begin{aligned}
\mathbb{E}\left[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k\right] \leq F(\overline{\mathbf{x}}^k) - \alpha\left\langle\nabla F(\overline{\mathbf{x}}^k), \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\rangle \\
+ \frac{\alpha^2L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right].
\end{aligned} \tag{19}$$

Using $2\langle\mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2, \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, in (19), we obtain: $\forall k \geq 0$

$$\begin{aligned}
\mathbb{E}\left[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k\right] \leq F(\overline{\mathbf{x}}^k) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k)\|^2 - \frac{\alpha}{2}\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 \\
+ \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k) - \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{\alpha^2L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right].
\end{aligned} \tag{20}$$

Applying Lemma 1(c) and (13) to (20), we have: $\forall k \geq 0$

$$\begin{aligned}
&\mathbb{E}[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k] \\
&\leq F(\overline{\mathbf{x}}^k) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k)\|^2 - \frac{\alpha(1-\alpha L)}{2}\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2
\end{aligned}$$

$$+\left(\frac{\alpha L^2}{2n}+\frac{\alpha^2L^3}{n^2}\right)\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2+\frac{\alpha^2L^3}{n}t^k. \quad (21)$$

The proof follows by the fact that if $0<\alpha\leq\frac{1}{2L}$, we have $-\frac{\alpha(1-\alpha L)}{2}\leq-\frac{\alpha}{4}$ and $\frac{\alpha L^2}{2n}+\frac{\alpha^2L^3}{n^2}\leq\frac{\alpha L^2}{n}$. ■

Compared with the corresponding descent inequality for centralized batch gradient descent [4], Lemma 4 exhibits two additional bias terms, i.e., $\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|$ and t^k, which are due to the decentralized nature of the problem and sampling. To establish the convergence of **GT-SAGA**, we, therefore, bound these bias terms by $\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|$ and show that they are dominated by the descent effect $-\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|$.

D. Bounds on the Auxiliary Sequence t^k

In this subsection, we analyze the evolution of the auxiliary sequence t^k and establish useful bounds.

Lemma 5: The following inequality holds: $\forall k\geq 0$

$$\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right]\leq\theta t^k+\left(2\alpha^2+\frac{\alpha}{\beta}\right)\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2+\left(\frac{2\alpha^2L^2}{n}+\frac{2}{m}\right)\frac{1}{n}\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2$$

where the parameter $\theta\in\mathbb{R}$ is given by

$$\theta:=1-\frac{1}{m}+\alpha\beta+\frac{2\alpha^2L^2}{n} \quad (22)$$

and $\beta>0$ is an arbitrary positive constant.

Proof: We define $\mathcal{A}^k:=\sigma(\cup_{i=1}^n\sigma(\tau_i^k),\mathcal{F}^k)$ and clearly $\mathcal{F}^k\subseteq\mathcal{A}^k$. By the tower property of the conditional expectation, we have: $\forall i\in\mathcal{V},\forall k\geq 0$

$$\mathbb{E}\left[t_i^{k+1}|\mathcal{F}^k\right]=\frac{1}{m}\sum_{j=1}^m\mathbb{E}\left[\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}-\mathbf{z}_{i,j}^{k+1}\|^2|\mathcal{A}^k\right]|\mathcal{F}^k\right]. \quad (23)$$

Since s_i^k is independent of $\mathcal{A}^k$ under Assumption 1, we have: $\forall i\in\mathcal{V},\forall j\in\{1,\ldots,m\},k\geq 0$

$$\mathbb{E}\left[\mathbb{1}_{\{s_i^k=j\}}|\mathcal{A}^k\right]=\frac{1}{m}\text{and}\mathbb{E}\left[\mathbb{1}_{\{s_i^k\neq j\}}|\mathcal{A}^k\right]=1-\frac{1}{m}. \quad (24)$$

In light of (24), we have: $\forall i\in\mathcal{V},\forall j\in\{1,\ldots,m\},k\geq 0$

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}-\mathbf{z}_{i,j}^{k+1}\|^2|\mathcal{A}^k\right]\\
&=\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1}-\left(\mathbb{1}_{\{s_i^k=j\}}\mathbf{x}_i^k+\mathbb{1}_{\{s_i^k\neq j\}}\mathbf{z}_{i,j}^k\right)\right\|^2\Big|\mathcal{A}^k\right]\\
&=\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}\|^2|\mathcal{A}^k\right]+\mathbb{E}\left[\left\|\mathbb{1}_{\{s_i^k=j\}}\mathbf{x}_i^k+\mathbb{1}_{\{s_i^k\neq j\}}\mathbf{z}_{i,j}^k\right\|^2\Big|\mathcal{A}^k\right]\\
&\quad-2\mathbb{E}\left[\left\langle\overline{\mathbf{x}}^{k+1},\mathbb{1}_{\{s_i^k=j\}}\mathbf{x}_i^k+\mathbb{1}_{\{s_i^k\neq j\}}\mathbf{z}_{i,j}^k\right\rangle\Big|\mathcal{A}^k\right]\\
&\overset{(i)}{=}\|\overline{\mathbf{x}}^{k+1}\|^2-2\left\langle\overline{\mathbf{x}}^{k+1},\frac{1}{m}\mathbf{x}_i^k+\left(1-\frac{1}{m}\right)\mathbf{z}_{i,j}^k\right\rangle\\
&\quad+\frac{1}{m}\|\mathbf{x}_i^k\|^2+\left(1-\frac{1}{m}\right)\|\mathbf{z}_{i,j}^k\|^2\\
&=\frac{1}{m}\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}_i^k\|^2+\left(1-\frac{1}{m}\right)\|\overline{\mathbf{x}}^{k+1}-\mathbf{z}_{i,j}^k\|^2 \quad (25)
\end{aligned}$$

where (i) uses (24) and that $\overline{\mathbf{x}}^{k+1},\mathbf{x}_i^k$, and $\mathbf{z}_{i,j}^k$ are $\mathcal{A}^k$-measurable. Using (25) in (23), we obtain: $\forall i\in\mathcal{V},\forall k\geq 0$

$$\mathbb{E}\left[t_i^{k+1}|\mathcal{F}^k\right]=\frac{1}{m}\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}_i^k\right\|^2\Big|\mathcal{F}^k\right]+\left(1-\frac{1}{m}\right)\frac{1}{m}\sum_{j=1}^m\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1}-\mathbf{z}_{i,j}^k\right\|^2\Big|\mathcal{F}^k\right]. \quad (26)$$

We next bound the two terms on the right-hand side (RHS) of (26) separately. For the first term, we have: $\forall i\in\mathcal{V},k\geq 0$

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}_i^k\|^2|\mathcal{F}^k\right]\\
&=\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}-\overline{\mathbf{x}}^k+\overline{\mathbf{x}}^k-\mathbf{x}_i^k\|^2|\mathcal{F}^k\right]\\
&=\alpha^2\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right]-2\left\langle\alpha\overline{\nabla\mathbf{f}}(\mathbf{x}^k),\overline{\mathbf{x}}^k-\mathbf{x}_i^k\right\rangle+\|\overline{\mathbf{x}}^k-\mathbf{x}_i^k\|^2\\
&\leq\alpha^2\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right]+\alpha^2\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2+2\|\mathbf{x}_i^k-\overline{\mathbf{x}}^k\|^2 \quad (27)
\end{aligned}$$

where the last line uses the Cauchy–Schwarz inequality. Toward the second term on the RHS of (26), we have: $\forall i\in\mathcal{V},\ j\in\{1,\ldots,m\},\forall k\geq 0,\forall\beta>0$

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}-\mathbf{z}_{i,j}^k\|^2|\mathcal{F}^k\right]\\
&=\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}-\overline{\mathbf{x}}^k+\overline{\mathbf{x}}^k-\mathbf{z}_{i,j}^k\|^2\big|\mathcal{F}^k\right]\\
&=\alpha^2\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right]-2\alpha\left\langle\overline{\nabla\mathbf{f}}(\mathbf{x}^k),\overline{\mathbf{x}}^k-\mathbf{z}_{i,j}^k\right\rangle+\|\overline{\mathbf{x}}^k-\mathbf{z}_{i,j}^k\|^2\\
&\leq\alpha^2\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right]+(1+\alpha\beta)\|\overline{\mathbf{x}}^k-\mathbf{z}_{i,j}^k\|^2+\frac{\alpha}{\beta}\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 \quad (28)
\end{aligned}$$

where the last line uses Young's inequality. Now, we apply (27) and (28) to (26) to obtain: $\forall i\in\mathcal{V},\forall k\geq 0$

$$\mathbb{E}\left[t_i^{k+1}|\mathcal{F}^k\right]\leq\left(1-\frac{1}{m}\right)(1+\alpha\beta)t_i^k+\alpha^2\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right]+\frac{2}{m}\|\mathbf{x}_i^k-\overline{\mathbf{x}}^k\|^2+\left(\frac{\alpha^2}{m}+\left(1-\frac{1}{m}\right)\frac{\alpha}{\beta}\right)\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2. \quad (29)$$

We average (29) over i from 1 to n and use (13) in the resulting inequality to obtain: $\forall k\geq 0$

$$\begin{aligned}
\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right]\leq&\left(\frac{2\alpha^2L^2}{n}+\frac{2}{m}\right)\frac{1}{n}\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2\\
&+\left(\alpha^2+\frac{\alpha^2}{m}+\left(1-\frac{1}{m}\right)\frac{\alpha}{\beta}\right)\left\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|^2\\
&+\left(\frac{2\alpha^2L^2}{n}+\left(1-\frac{1}{m}\right)(1+\alpha\beta)\right)t^k. \quad (30)
\end{aligned}$$

We conclude by using $\frac{1}{m}+1\leq 2$ and $1-\frac{1}{m}\leq 1$ in (30). ■

Next, we specify some particular choices of β and the range of α in Lemma 5 to obtain useful bounds on the auxiliary sequence t^k. The following corollary shows that t^k has an intrinsic contraction property.

Corollary 1: If $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, then $\forall k \geq 0$

$$\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \leq \left(1-\frac{1}{4m}\right)t^k + 4m\alpha^2\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{9}{4mn}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2.$$

Proof: We choose $\beta = \frac{1}{2m\alpha}$ in Lemma 5 to obtain: if $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, i.e., $\frac{2\alpha^2L^2}{n} \leq \frac{1}{4m}$, then

$$\theta = 1 - \frac{1}{m} + \alpha\beta + \frac{2\alpha^2L^2}{n} \leq 1 - \frac{1}{4m} \tag{31}$$

$$2\alpha^2 + \frac{\alpha}{\beta} = 2\alpha^2 + 2m\alpha^2 \leq 4m\alpha^2 \tag{32}$$

$$\frac{2\alpha^2L^2}{n} + \frac{2}{m} \leq \frac{1}{4m} + \frac{2}{m} = \frac{9}{4m}. \tag{33}$$

We conclude by applying (31)–(33) to Lemma 5. ■

The following corollary of Lemma 5 will be only used to bound $\mathbb{E}[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k]$.

Corollary 2: If $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, then $\forall k \geq 0$

$$\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \leq 2t^k + 3\alpha^2\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{9}{4mn}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2.$$

Proof: Setting $\beta = 1/\alpha$ in Lemma 5, we have: if $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, i.e., $\frac{2\alpha^2L^2}{n} \leq \frac{1}{4m}$, then

$$\theta = 1 - \frac{1}{m} + \alpha\beta + \frac{2L^2\alpha^2}{n} \leq 2 \tag{34}$$

$$2\alpha^2 + \frac{\alpha}{\beta} = 3\alpha^2 \tag{35}$$

$$\frac{2\alpha^2L^2}{n} + \frac{2}{m} \leq \frac{1}{4m} + \frac{2}{m} = \frac{9}{4m}. \tag{36}$$

We conclude by applying (34)–(36) to Lemma 5. ■

With the help of (10), (12), and Corollary 2, we provide an upper bound on $\mathbb{E}[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k]$.

Lemma 6: If $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, then $\forall k \geq 0$

$$\mathbb{E}[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k] \leq 8.5L^2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 4nL^2t^k + 6n\alpha^2L^2\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + 4\alpha^2L^2\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2|\mathcal{F}^k\right].$$

Proof: By the tower property of the conditional expectation, we have: $\forall k \geq 0$

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^{k+1}\right]|\mathcal{F}^k\right] \\
&\leq 2L^2\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\|^2|\mathcal{F}^k\right] + 2nL^2\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \\
&\leq 2L^2\left(2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 2\alpha^2\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2|\mathcal{F}^k\right]\right) \\
&\quad + 2nL^2\left(2t^k + 3\alpha^2\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{9}{4mn}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right)
\end{aligned}$$

where the second line uses (12) and the third line uses (10) and Corollary 2. The desired inequality then follows. ■

E. Bounds on Stochastic Gradient Tracking Process

In this subsection, we analyze the variance-reduced stochastic gradient tracking process (8a).

Lemma 7: The following inequality holds: $\forall k \geq 0$

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\right] \\
&\leq \lambda^2\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] + \lambda^2\mathbb{E}\left[\|\mathbf{g}^{k+1} - \mathbf{g}^k\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\langle(\mathbf{W}-\mathbf{J})\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\right)\right\rangle\right] \\
&\quad + 2\mathbb{E}\left[\left\langle(\mathbf{W}-\mathbf{J})\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^k) - \mathbf{g}^k\right)\right\rangle\right].
\end{aligned}$$

Proof: Using (8a) and $\mathbf{J}\mathbf{W} = \mathbf{J}$, we have: $\forall k \geq 0$

$$\begin{aligned}
&\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2 \\
&= \|\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1} + (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\|^2 \\
&= \|\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 + \|(\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\|^2 \\
&\quad + 2\left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\right\rangle \\
&\leq \lambda^2\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 + \lambda^2\|\mathbf{g}^{k+1} - \mathbf{g}^k\|^2 \\
&\quad + 2\left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\right\rangle
\end{aligned} \tag{37}$$

where the last line uses Lemma 1(a) and $\|\mathbf{W} - \mathbf{J}\| = \lambda$. To proceed, we observe that $\forall k \geq 0$

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\right\rangle|\mathcal{F}^{k+1}\right] \\
&= \left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \mathbf{g}^k\right)\right\rangle \\
&= \left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\right)\right\rangle \\
&\quad + \left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^k) - \mathbf{g}^k\right)\right\rangle
\end{aligned} \tag{38}$$

where the first line uses that $\mathbb{E}[\mathbf{g}^{k+1}|\mathcal{F}^{k+1}] = \nabla \mathbf{f}(\mathbf{x}^{k+1})$ and that $\mathbf{y}^{k+1}$ and $\mathbf{g}^k$ are $\mathcal{F}^{k+1}$-measurable for all $k \geq 0$. We conclude by using (38) in (37) and taking the expectation. ■

We next bound the third term in Lemma 7.

Lemma 8: The following inequality holds: $\forall k \geq 0$

$$\begin{aligned}
&\left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\right)\right\rangle \\
&\leq \left(\lambda\alpha L + 0.5\eta_1 + \eta_2\right)\lambda^2\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 \\
&\quad + 0.5\eta_1^{-1}\lambda^2\alpha^2L^2n\|\overline{\mathbf{g}}^k\|^2 + \eta_2^{-1}\lambda^2L^2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2
\end{aligned}$$

where $\eta_1 > 0$ and $\eta_2 > 0$ are arbitrary.

Proof: Using Lemma 1(a) and $\|\mathbf{W} - \mathbf{J}\| = \lambda$, we have

$$\begin{aligned}
&\left\langle\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\right)\right\rangle \\
&\leq \lambda^2L\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \qquad \forall k \geq 0.
\end{aligned} \tag{39}$$

Observe that $\forall k \geq 0$

$$\begin{aligned}
&\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \\
&= \|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1} + \mathbf{J}\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^k + \mathbf{J}\mathbf{x}^k - \mathbf{x}^k\| \\
&\leq \|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\| + \sqrt{n}\alpha\|\overline{\mathbf{g}}^k\| + \|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\| \\
&\leq 2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\| + \sqrt{n}\alpha\|\overline{\mathbf{g}}^k\| + \alpha\lambda\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|
\end{aligned} \tag{40}$$

where the last line is due to (11). We use (40) in (39) to obtain

$$\begin{aligned}&\left\langle \mathbf{W}\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}(\mathbf{x}^{k+1})-\nabla\mathbf{f}(\mathbf{x}^{k})\right)\right\rangle\\ &\le \lambda^3\alpha L\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2+\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|\sqrt{n}\alpha L\|\overline{\mathbf{g}}^k\|\\ &\quad+2\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|L\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\| \qquad \forall k\ge 0. \end{aligned}\tag{41}$$

By Young's inequality, we have: $\forall k\ge 0$, for some $\eta_1>0$

$$\begin{aligned}&\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|\sqrt{n}\alpha L\|\overline{\mathbf{g}}^k\|\\ &\le 0.5\lambda^2\left(\eta_1\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2+\eta_1^{-1}n\alpha^2L^2\|\overline{\mathbf{g}}^k\|^2\right)\end{aligned}\tag{42}$$

and, $\forall k\ge 0$, for some $\eta_2>0$

$$\begin{aligned}&2\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|L\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|\\ &\le \lambda^2\eta_2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2+\lambda^2\eta_2^{-1}L^2\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2.\end{aligned}\tag{43}$$

The proof follows by applying (42) and (43) to (41). ■

We next bound the fourth term in Lemma 7.

Lemma 9: The following inequality holds: $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{W}\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}(\mathbf{x}^k)-\mathbf{g}^k\right)\right\rangle\right]\\ &\le \mathbb{E}\left[\|\mathbf{g}^k-\nabla\mathbf{f}(\mathbf{x}^k)\|^2\right]/n.\end{aligned}$$

Proof: In the following, we denote $\nabla\mathbf{f}^k:=\nabla\mathbf{f}(\mathbf{x}^k)$ to simplify the notation. Observe that $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{W}\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}^k-\mathbf{g}^k\right)\right\rangle|\mathcal{F}^k\right]\\ &\overset{(i)}{=}\mathbb{E}\left[\left\langle \mathbf{W}^2(\mathbf{y}^k+\mathbf{g}^k-\mathbf{g}^{k-1}),(\mathbf{W}-\mathbf{J})(\nabla\mathbf{f}^k-\mathbf{g}^k)\right\rangle|\mathcal{F}^k\right]\\ &\overset{(ii)}{=}\mathbb{E}\left[\left\langle \mathbf{W}^2\mathbf{g}^k,(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}^k-\mathbf{g}^k\right)\right\rangle|\mathcal{F}^k\right]\\ &\overset{(iii)}{=}\mathbb{E}\left[\left\langle \mathbf{W}^2\left(\mathbf{g}^k-\nabla\mathbf{f}^k\right),(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}^k-\mathbf{g}^k\right)\right\rangle|\mathcal{F}^k\right]\\ &\overset{(iv)}{=}\mathbb{E}\left[(\mathbf{g}^k-\nabla\mathbf{f}^k)^\top(\mathbf{J}-\mathbf{W}^\top\mathbf{W}^2)(\mathbf{g}^k-\nabla\mathbf{f}^k)|\mathcal{F}^k\right]\end{aligned}\tag{44}$$

where (i) uses (8a) and $\mathbf{J}\mathbf{W}=\mathbf{J}$, (ii) and (iii) use that $\mathbf{y}^k$, $\mathbf{g}^{k-1}$, and $\nabla\mathbf{f}^k$ are $\mathcal{F}^k$-measurable and that $\mathbb{E}[\mathbf{g}^k|\mathcal{F}^k]=\nabla\mathbf{f}^k$ for all $k\ge 0$, and (iv) uses $\mathbf{J}\mathbf{W}=\mathbf{J}$. Since, whenever $i\ne j\in\mathcal{V}$, $\mathbb{E}[\langle\mathbf{g}_i^k-\nabla f_i(\mathbf{x}_i^k),\mathbf{g}_j^k-\nabla f_j(\mathbf{x}_j^k)\rangle|\mathcal{F}^k]=0$, and $\mathbf{W}^\top\mathbf{W}^2$ is nonnegative, we have: $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[(\mathbf{g}^k-\nabla\mathbf{f}^k)^\top\left(\mathbf{J}-\mathbf{W}^\top\mathbf{W}^2\right)\left(\mathbf{g}^k-\nabla\mathbf{f}^k\right)|\mathcal{F}^k\right]\\ &=\mathbb{E}\left[(\mathbf{g}^k-\nabla\mathbf{f}^k)^\top\mathrm{diag}\left(\mathbf{J}-\mathbf{W}^\top\mathbf{W}^2\right)\left(\mathbf{g}^k-\nabla\mathbf{f}^k\right)|\mathcal{F}^k\right]\\ &\le\mathbb{E}\left[(\mathbf{g}^k-\nabla\mathbf{f}^k)^\top\mathrm{diag}(\mathbf{J})\left(\mathbf{g}^k-\nabla\mathbf{f}^k\right)|\mathcal{F}^k\right].\end{aligned}\tag{45}$$

The proof follows by taking the expectation of (45). ■

We finally bound the second term in Lemma 7.

Lemma 10: The following inequality holds: $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{g}^{k+1}-\mathbf{g}^k\|^2\right]\le 12\lambda^2\alpha^2L^2\mathbb{E}\left[\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2\right]\\ &+2\mathbb{E}\left[\|\mathbf{g}^k-\nabla\mathbf{f}(\mathbf{x}^k)\|^2\right]+\mathbb{E}\left[\|\mathbf{g}^{k+1}-\nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2\right]\\ &+18L^2\mathbb{E}\left[\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2\right]+6n\alpha^2L^2\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2\right].\end{aligned}$$

Proof: Since $\mathbf{g}^k$ and $\nabla\mathbf{f}(\mathbf{x}^{k+1})$ are $\mathcal{F}^{k+1}$-measurable, and $\mathbb{E}[\mathbf{g}^{k+1}|\mathcal{F}^{k+1}]=\nabla\mathbf{f}(\mathbf{x}^{k+1})$, we have: $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{g}^{k+1}-\mathbf{g}^k\|^2|\mathcal{F}^{k+1}\right]\\ &=\mathbb{E}\left[\|\mathbf{g}^{k+1}-\nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^{k+1}\right]+\|\nabla\mathbf{f}(\mathbf{x}^{k+1})-\mathbf{g}^k\|^2\\ &\le\mathbb{E}\left[\|\mathbf{g}^{k+1}-\nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^{k+1}\right]\\ &\quad+2\|\nabla\mathbf{f}(\mathbf{x}^{k+1})-\nabla\mathbf{f}(\mathbf{x}^k)\|^2+2\|\nabla\mathbf{f}(\mathbf{x}^k)-\mathbf{g}^k\|^2\\ &\le\mathbb{E}\left[\|\mathbf{g}^{k+1}-\nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^{k+1}\right]\\ &\quad+2L^2\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2+2\|\nabla\mathbf{f}(\mathbf{x}^k)-\mathbf{g}^k\|^2.\end{aligned}\tag{46}$$

Similar to the derivation of (40), we have: $\forall k\ge 0$

$$\begin{aligned}&\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\\ &\le 3\|\mathbf{x}^{k+1}-\mathbf{J}\mathbf{x}^{k+1}\|^2+3n\alpha^2\|\overline{\mathbf{g}}^k\|^2+3\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2\\ &\le 9\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2+3n\alpha^2\|\overline{\mathbf{g}}^k\|^2+6\alpha^2\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2\end{aligned}$$

where the last line is due to (10). We conclude by applying the last line above to (46) and taking the expectation. ■

Now, we apply Lemmas 8–10 to Lemma 7.

Lemma 11: The following inequality holds: $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{y}^{k+2}-\mathbf{J}\mathbf{y}^{k+2}\|^2\right]\\ &\le\left(1+2\lambda\alpha L+\eta_1+2\eta_2+12\lambda^2\alpha^2L^2\right)\lambda^2\\ &\quad\times\mathbb{E}\left[\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2\right]\\ &\quad+\left(2\eta_2^{-1}+18\right)\lambda^2L^2\mathbb{E}\left[\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2\right]\\ &\quad+\left(\eta_1^{-1}+6\right)\lambda^2\alpha^2L^2n\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2\right]\\ &\quad+\left(2\lambda^2+2/n\right)\mathbb{E}\left[\|\mathbf{g}^k-\nabla\mathbf{f}(\mathbf{x}^k)\|^2\right]\\ &\quad+\lambda^2\mathbb{E}\left[\|\mathbf{g}^{k+1}-\nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2\right].\end{aligned}$$

Proof: Apply Lemmas 8–10 to Lemma 7. ■

Finally, we use Lemmas 3 and 6 to refine Lemma 11 and establish a contraction in the gradient tracking process.

Lemma 12: If $0<\alpha\le\min\{\frac{1-\lambda^2}{16\lambda},\frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then we have: $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{y}^{k+2}-\mathbf{J}\mathbf{y}^{k+2}\|^2\right]\\ &\le\frac{1+\lambda^2}{2}\mathbb{E}\left[\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2\right]+\frac{30.5L^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2\right]\\ &\quad+\frac{97L^2n}{8}\mathbb{E}\left[t^k\right]+\frac{16\lambda^2\alpha^2L^2n}{1-\lambda^2}\mathbb{E}\left[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\right].\end{aligned}$$

Proof: We apply Lemmas 3 and 6 to Lemma 11 to obtain: if $0<\alpha\le\frac{\sqrt{n}}{\sqrt{8m}L}$, then $\forall k\ge 0$

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{y}^{k+2}-\mathbf{J}\mathbf{y}^{k+2}\|^2\right]\\ &\le\left(1+2\lambda\alpha L+\eta_1+2\eta_2+\left(12\lambda^2+4\right)\alpha^2L^2\right)\lambda^2\\ &\quad\times\mathbb{E}\left[\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2\right]\\ &+\left((2\eta_2^{-1}+18)\lambda^2+(\eta_1^{-1}+6)\frac{2\lambda^2\alpha^2L^2}{n}+\frac{4}{n}+12.5\lambda^2\right)\\ &\quad\times L^2\mathbb{E}\left[\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2\right]\\ &+\left(2(\eta_1^{-1}+6)\lambda^2\alpha^2L^2+\left(2\lambda^2+1/n\right)4n\right)L^2\mathbb{E}\left[t^k\right]\\ &+(\eta_1^{-1}+12)\lambda^2\alpha^2L^2n\mathbb{E}\left[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\right].\end{aligned}\tag{47}$$

We fix $\eta_1 = \frac{1-\lambda^2}{16\lambda^2}$ and $\eta_2 = \frac{1-\lambda^2}{8\lambda^2}$. It can then be verified that $1 + 2\lambda\alpha L + \eta_1 + 2\eta_2 + (12\lambda^2+4)\alpha^2L^2 \leq \frac{1+\lambda^2}{2\lambda^2}$, if $0 < \alpha \leq \frac{1-\lambda^2}{16\lambda L}$. The proof then follows by applying this inequality and the values of η_1 and η_2 to (47). ■

F. Proof of Theorem 1

In this subsection, we prove the convergence of **GT-SAGA** for general smooth nonconvex functions. To this aim, we write the contraction inequalities in (9), Corollary 1, and Lemma 12 as an LTI dynamics that jointly characterizes the evolution of the consensus, gradient tracking, and the auxiliary sequence t^k.

Proposition 1: If $0 < \alpha \leq \min\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then

$$\mathbf{u}^{k+1} \leq \mathbf{G}_\alpha \mathbf{u}^k + \mathbf{b}^k \qquad \forall k \geq 0$$

where $\mathbf{u}^k \in \mathbb{R}^3$, $\mathbf{G}_\alpha \in \mathbb{R}^{3\times 3}$, and $\mathbf{b}^k \in \mathbb{R}^3$ are given by

$$\mathbf{u}^k := \begin{bmatrix} \mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \\ \mathbb{E}\left[t^k\right] \\ \mathbb{E}\left[\frac{1}{nL^2}\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] \end{bmatrix}, \quad \mathbf{b} := \begin{bmatrix} 0 \\ 4m\alpha^2 \\ \frac{16\lambda^2\alpha^2}{1-\lambda^2} \end{bmatrix}$$

$$\mathbf{G}_\alpha := \begin{bmatrix} \frac{1+\lambda^2}{2} & 0 & \frac{2\lambda^2\alpha^2L^2}{1-\lambda^2} \\ \frac{9}{4m} & 1-\frac{1}{4m} & 0 \\ \frac{30.5}{1-\lambda^2} & \frac{97}{8} & \frac{1+\lambda^2}{2} \end{bmatrix}$$

and $\mathbf{b}^k := \mathbf{b}\mathbb{E}[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2]$.

We first derive the range of the step size α under which the spectral radius of $\mathbf{G}_\alpha$ defined in Proposition 1 is less than 1, with the help of the following Lemma from [58].

Lemma 13: Let $\mathbf{X} \in \mathbb{R}^{d\times d}$ be a nonnegative matrix and $\mathbf{x} \in \mathbb{R}^d$ be a positive vector. If $\mathbf{X}\mathbf{x} < \mathbf{x}$, then $\rho(\mathbf{X}) < 1$. Moreover, if $\mathbf{X}\mathbf{x} \leq \beta\mathbf{x}$, for some $\beta \in \mathbb{R}$, then $\rho(\mathbf{X}) \leq \beta$.

Lemma 14: If $0 < \alpha \leq \min\{\frac{(1-\lambda^2)^2}{35\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then we have $\rho(\mathbf{G}_\alpha) < 1$, and thus, $\sum_{k=0}^\infty \mathbf{G}_\alpha^k = (\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}$.

Proof: In light of Lemma 13, we find a positive vector $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \epsilon_3]^\top$ and the range of α s.t. $\mathbf{G}_\alpha\boldsymbol{\epsilon} < \boldsymbol{\epsilon}$, i.e.,

$$\alpha^2 < \frac{(1-\lambda^2)^2}{4\lambda^2L^2}\frac{\epsilon_1}{\epsilon_3} \tag{48}$$

$$9\epsilon_1 < \epsilon_2 \tag{49}$$

$$\frac{61}{(1-\lambda^2)^2}\epsilon_1 + \frac{97}{4(1-\lambda^2)}\epsilon_2 < \epsilon_3. \tag{50}$$

Based on (49), we set $\epsilon_1 = 1$ and $\epsilon_2 = 10$. Then, based on (50), we set $\epsilon_3 = \frac{303.5}{(1-\lambda^2)^2}$. The proof follows by using the values of ϵ_1 and ϵ_3 in (48). ■

Based on the LTI dynamics in Proposition 1, we derive the following lemma that is the key to establish the convergence of **GT-SAGA** for general smooth nonconvex functions.

Lemma 15: If $0 < \alpha \leq \min\{\frac{(1-\lambda^2)^2}{35\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then we have: $\forall K \geq 1$

$$\sum_{k=0}^K \mathbf{u}^k \leq (\mathbf{I} - \mathbf{G}_\alpha)^{-1}\left(\mathbf{u}^0 + \mathbf{b}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\right]\right).$$

Proof: We recursively apply the dynamics in Proposition 1 to obtain: $\mathbf{u}^k \leq \mathbf{G}_\alpha^k\mathbf{u}^0 + \sum_{r=0}^{k-1}\mathbf{G}_\alpha^r\mathbf{b}^{k-1-r}, \forall k \geq 1$. We sum this inequality over k to obtain: $\forall K \geq 1$

$$\sum_{k=0}^K \mathbf{u}^k \leq \sum_{k=0}^K \mathbf{G}_\alpha^k\mathbf{u}^0 + \sum_{k=1}^K\sum_{r=0}^{k-1}\mathbf{G}_\alpha^r\mathbf{b}^{k-1-r}$$
$$\leq \left(\sum_{k=0}^\infty \mathbf{G}_\alpha^k\right)\mathbf{u}^0 + \sum_{k=0}^{K-1}\left(\sum_{k=0}^\infty \mathbf{G}_\alpha^k\right)\mathbf{b}^k.$$

The proof follows by $\sum_{k=0}^\infty \mathbf{G}_\alpha^k = (\mathbf{I} - \mathbf{G}_\alpha)^{-1}$ and the definition of $\mathbf{b}^k$ in Proposition 1. ■

Lemma 16: If $0 < \alpha \leq \min\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then

$$(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1} \leq \begin{bmatrix} \star & \frac{776\lambda^2m\alpha^2L^2}{(1-\lambda^2)^3} & \frac{16\lambda^2\alpha^2L^2}{(1-\lambda^2)^3} \\ \star & 8m & \frac{114\lambda^2\alpha^2L^2}{(1-\lambda^2)^3} \\ \star & \star & \star \end{bmatrix}$$

$$(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}\mathbf{b} \leq \begin{bmatrix} \left(3104m^2 + \frac{256\lambda^2}{1-\lambda^2}\right)\frac{\lambda^2\alpha^4L^2}{(1-\lambda^2)^3} \\ 33m^2\alpha^2 \\ \star \end{bmatrix}$$

where the $\star$ entries are not needed for further derivations.

Proof: In the following, for a matrix $\mathbf{X}$, we denote $\mathbf{X}^*$ as its adjugate and $[\mathbf{X}]_{i,j}$ as its (i,j)th entry. We first note that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{48\lambda L}$, $\det(\mathbf{I}_3 - \mathbf{G}_\alpha) \geq \frac{(1-\lambda^2)^2}{32m}$. We next derive upper bounds for entries of $(\mathbf{I}_3 - \mathbf{G}_\alpha)^*$

$$[(\mathbf{I} - \mathbf{G}_\alpha)^*]_{1,2} = \frac{97\lambda^2\alpha^2L^2}{4(1-\lambda^2)}, [(\mathbf{I} - \mathbf{G}_\alpha)^*]_{1,3} = \frac{\lambda^2\alpha^2L^2}{2m(1-\lambda^2)}$$

$$[(\mathbf{I} - \mathbf{G}_\alpha)^*]_{2,2} \leq \frac{(1-\lambda^2)^2}{4}, [(\mathbf{I} - \mathbf{G}_\alpha)^*]_{2,3} = \frac{9\lambda^2\alpha^2L^2}{2m(1-\lambda^2)}.$$

The upper bound on $(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}$ then follows by using the above relations. Finally, we have

$$(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}\mathbf{b} \leq \begin{bmatrix} \frac{3104\lambda^2m^2\alpha^4L^2}{(1-\lambda^2)^3} + \frac{256\lambda^4\alpha^4L^2}{(1-\lambda^2)^4} \\ 32m^2\alpha^2 + \frac{2304\lambda^4\alpha^4L^2}{(1-\lambda^2)^4} \\ \star \end{bmatrix}.$$

If $0 < \alpha \leq \frac{(1-\lambda^2)^2}{48\lambda L}$, then $32m^2\alpha^2 + \frac{2304\lambda^4\alpha^4L^2}{(1-\lambda^2)^4} \leq 33m^2\alpha^2$ and the bound on $(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}\mathbf{b}$ follows. ■

We now bound two important quantities as follows.

Lemma 17: If $0 < \alpha \leq \min\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then we have: $\forall K \geq 1$

$$\sum_{k=0}^K \mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \leq \frac{16\lambda^4\alpha^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}$$
$$+ \left(97m^2 + \frac{8\lambda^2}{1-\lambda^2}\right)\frac{32\lambda^2\alpha^4L^2}{(1-\lambda^2)^3}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\right] \tag{51}$$

and $\forall K \geq 1$

$$\sum_{k=0}^K \mathbb{E}\left[t^k\right] \leq \frac{114\lambda^4\alpha^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}$$

$$+ 33m^2\alpha^2 \sum_{k=0}^{K-1} \mathbb{E}\left[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\right]. \tag{52}$$

Proof: By (8a), we have $\|\mathbf{y}^1 - \mathbf{J}\mathbf{y}^1\|^2 = \|(\mathbf{W}-\mathbf{J})(\mathbf{y}^0 + \mathbf{g}^0 - \mathbf{g}^{-1})\|^2 \le \lambda^2\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2$. The proof then follows by applying this inequality and Lemma 16 to Lemma 15. ■

Now, we are ready to prove Theorem 1.

Proof of Theorem 1: We sum up the inequality in Lemma 4 over k to obtain: if $0 < \alpha \le \frac{1}{2L}$, then $\forall K \ge 1$

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{x}}^K)\right] \le F(\overline{\mathbf{x}}^0) &- \frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\overline{\mathbf{x}}^k)\|^2\right] \\ &- \frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\right] + \frac{\alpha^2 L^3}{n}\sum_{k=0}^{K-1}\mathbb{E}\left[t^k\right] \\ &+ \alpha L^2 \sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right]. \end{aligned} \tag{53}$$

By the L-smoothness of F, we have $\frac{1}{2n}\sum_{i=1}^n \|\nabla F(\mathbf{x}_i^k)\|^2 \le \|\nabla F(\overline{\mathbf{x}}^k)\|^2 + \frac{L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2$, $\forall k \ge 0$. Using this inequality in (53), we obtain: if $0 < \alpha \le \frac{1}{2L}$, then $\forall K \ge 1$

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{x}}^K)\right] \le F(\overline{\mathbf{x}}^0) &- \frac{\alpha}{4n}\sum_{i=1}^n\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2\right] \\ &- \frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\right] + \frac{\alpha^2 L^3}{n}\sum_{k=0}^{K-1}\mathbb{E}\left[t^k\right] \\ &+ \frac{3\alpha L^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right]. \end{aligned} \tag{54}$$

Applying (52) to (54), we obtain the following inequality: if $0 < \alpha \le \min\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}, \frac{1}{2}\}\frac{1}{L}$, then $\forall K \ge 1$

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{x}}^K)\right] \le F(\overline{\mathbf{x}}^0) &- \frac{\alpha}{4n}\sum_{i=1}^n\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2\right] \\ &- \frac{\alpha}{8}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\right] + \frac{114\lambda^4\alpha^4L^3}{n(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n} \\ &+ \frac{3\alpha L^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \\ &- \frac{\alpha}{8}\left(1 - \frac{264m^2\alpha^3L^3}{n}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\right]. \end{aligned} \tag{55}$$

If $0 < \alpha \le \frac{2n^{1/3}}{13m^{2/3}L}, 1 - \frac{264m^2\alpha^3L^3}{n} \ge 0$, and thus, the last term in (55) may be dropped. We then use (51) in (55) to obtain: if $0 < \alpha \le \min\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{2n^{1/3}}{13m^{2/3}}, \frac{1}{2}\}\frac{1}{L}$, then $\forall K \ge 1$

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{x}}^K)\right] \le F(\overline{\mathbf{x}}^0) &- \frac{\alpha}{4n}\sum_{i=1}^n\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2\right] \\ &- \frac{\alpha L^2}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \\ &+ \left(\frac{114\alpha L}{28n} + 1\right)\frac{28\lambda^4\alpha^3L^2}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n} \\ &- \frac{\alpha}{8}\left(1 - \max\left\{97m^2, \frac{8\lambda^2}{1-\lambda^2}\right\}\frac{896\lambda^2\alpha^4L^4}{(1-\lambda^2)^3}\right) \\ &\times \sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\right]. \end{aligned} \tag{56}$$

We note that if $0 < \alpha \le \min\{\frac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}, \frac{1-\lambda^2}{12\lambda}\}\frac{1}{L}$, then $\max\{97m^2, \frac{8\lambda^2}{1-\lambda^2}\}\frac{896\lambda^2\alpha^4L^4}{(1-\lambda^2)^3} \le 1$ and the last term in (56) may be dropped. Therefore, if $0 < \alpha \le \overline{\alpha}_1$ for $\overline{\alpha}_1$ defined in Theorem 1, we obtain from (56) that $\forall K \ge 1$

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{x}}^K)\right] \le F(\overline{\mathbf{x}}^0) &- \frac{\alpha}{4n}\sum_{i=1}^n\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2\right] \\ &- \frac{\alpha L^2}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] + \frac{112\lambda^4\alpha^3L^2}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}. \end{aligned} \tag{57}$$

Since F is bounded below by F^*, (57) leads to $\forall K \ge 1$

$$\begin{aligned}&\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2 + L^2\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2\right] \\ &\le \frac{4(F(\overline{\mathbf{x}}^0) - F^*)}{\alpha} + \frac{448\lambda^4\alpha^2L^2}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}. \end{aligned} \tag{58}$$

Since the RHS of (58) is finite and independent of K, we let $K \to \infty$ in (58) to obtain

$$\sum_{k=0}^{\infty}\sum_{i=1}^n\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2 + \|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2\right] < \infty \tag{59}$$

which shows that all nodes in **GT-SAGA** asymptotically agree on a stationary point of F in the mean-squared sense. Moreover, since the series on the left-hand side (LHS) of (59) is nonnegative, we may exchange the order of the series and expectation to obtain [59]: $\mathbb{E}[\sum_{k=0}^{\infty}\sum_{i=1}^n(\|\nabla F(\mathbf{x}_i^k)\|^2 + \|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2)] < \infty$, which implies that

$$\mathbb{P}\left(\sum_{k=0}^{\infty}\sum_{i=1}^n\left(\|\nabla F(\mathbf{x}_i^k)\|^2 + \|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2\right) < \infty\right) = 1 \tag{60}$$

i.e., all nodes in **GT-SAGA** asymptotically agree on a stationary point of F in the almost sure sense. Finally, toward the iteration complexity of **GT-SAGA**, we set $\alpha = \overline{\alpha}_1$ in (58) and divide the resulting inequality by K to obtain: $\forall K \ge 1$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\mathbf{x}_i^k)\|^2\right] \\ &\le \frac{4(F(\overline{\mathbf{x}}^0) - F^*)}{\overline{\alpha}_1 K} + \frac{448\lambda^4\overline{\alpha}_1^2L^2}{(1-\lambda^2)^3K}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}. \end{aligned} \tag{61}$$

Based on (61), the iteration complexity of **GT-SAGA** then follows by recalling the definition of $\overline{\alpha}_1$ in Theorem 1 and that $\frac{448\lambda^4\overline{\alpha}_1^2L^2}{(1-\lambda^2)^3} \leq \frac{\lambda^2(1-\lambda^2)}{4}$ since $0 < \overline{\alpha}_1 \leq \frac{(1-\lambda^2)^2}{48\lambda L}$. ■

G. Proof of Theorem 2

In this subsection, we prove the linear rate of **GT-SAGA** when the global function F additionally satisfies the PL condition. In particular, we use the PL condition and Lemma 1(e) to refine the descent inequality in Lemma 4 and the previously obtained LTI system in Proposition 1.

Lemma 18: If $0 < \alpha \leq \frac{1}{2L}$, then $\forall k \geq 0$

$$\mathbb{E}\left[F(\overline{\mathbf{x}}^{k+1}) - F^*|\mathcal{F}^k\right]$$

$$\leq (1-\mu\alpha)(F(\overline{\mathbf{x}}^k) - F^*) + \frac{\alpha L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \frac{\alpha^2 L^3}{n}t^k.$$

Proof: Apply the PL condition to Lemma 4 and then subtract F^* from the resulting inequality. ■

Next, we refine Corollary 1 as follows.

Lemma 19: If $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, then $\forall k \geq 0$

$$\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \leq \left(1 - \frac{1}{4m}\right)t^k + 16m\alpha^2 L\left(F(\overline{\mathbf{x}}^k) - F^*\right)$$
$$+ \left(8m\alpha^2L^2 + \frac{9}{4m}\right)\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2.$$

Proof: By Lemmas 1(c) and 1(e), we have: $\forall k \geq 0$

$$\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 \leq 2\|\nabla F(\overline{\mathbf{x}}^k)\|^2 + 2\|\nabla F(\overline{\mathbf{x}}^k) - \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2$$
$$\leq 4L\left(F(\overline{\mathbf{x}}^k) - F^*\right) + \frac{2L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2. \quad (62)$$

The proof follows by applying (62) to Corollary 1. ■

We finally refine Lemma 12 as follows.

Lemma 20: If $0 < \alpha \leq \min\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then $\forall k \geq 0$

$$\mathbb{E}\left[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\right]$$
$$\leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] + \frac{31L^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right]$$
$$+ \frac{97L^2n}{8}\mathbb{E}\left[t^k\right] + \frac{64\lambda^2\alpha^2L^3n}{1-\lambda^2}\mathbb{E}\left[F(\overline{\mathbf{x}}^k) - F^*\right].$$

Proof: Applying (62) to Lemma 12, we have: if $0 < \alpha \leq \min\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\}\frac{1}{L}$, then $\forall k \geq 0$

$$\mathbb{E}\left[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right]$$
$$+ \left(30.5 + 32\lambda^2\alpha^2L^2\right)\frac{L^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right]$$
$$+ \frac{97L^2n}{8}\mathbb{E}\left[t^k\right] + \frac{64\lambda^2\alpha^2L^3n}{1-\lambda^2}\mathbb{E}\left[F(\overline{\mathbf{x}}^k) - F^*\right].$$

We conclude by $30.5 + 32\lambda^2\alpha^2L^2 \leq 31$ if $0 < \alpha \leq \frac{1-\lambda^2}{16\lambda L}$. ■

Now, we write (9) and Lemmas 18–20 in an LTI system.

Proposition 2: If $0 < \alpha \leq \min\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}, \frac{1}{2}\}\frac{1}{L}$, then

$$\mathbf{v}^{k+1} \leq \mathbf{H}_\alpha\mathbf{v}^k \qquad \forall k \geq 0$$

where $\mathbf{v}^k \in \mathbb{R}^4$ and $\mathbf{H}_\alpha \in \mathbb{R}^{4\times4}$ are given by

$$\mathbf{v}^k := \begin{bmatrix} \mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \\ \frac{1}{L}\mathbb{E}\left[F(\overline{\mathbf{x}}^k) - F^*\right] \\ \mathbb{E}\left[t^k\right] \\ \mathbb{E}\left[\frac{1}{nL^2}\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] \end{bmatrix}$$

$$\mathbf{H}_\alpha := \begin{bmatrix} \frac{1+\lambda^2}{2} & 0 & 0 & \frac{2\lambda^2\alpha^2L^2}{1-\lambda^2} \\ \alpha L & 1-\mu\alpha & \frac{\alpha^2L^2}{n} & 0 \\ 8m\alpha^2L^2 + \frac{9}{4m} & 16m\alpha^2L^2 & 1 - \frac{1}{4m} & 0 \\ \frac{31}{1-\lambda^2} & \frac{64\lambda^2\alpha^2L^2}{1-\lambda^2} & \frac{97}{8} & \frac{1+\lambda^2}{2} \end{bmatrix}.$$

We are ready to prove Theorem 2, i.e., to establish an upper bound on $\rho(\mathbf{H}_\alpha)$ that characterizes the explicit linear rate of **GT-SAGA** under the PL condition.

Proof of Theorem 2: In light of Lemma 13, we solve for the range of α under which there exists a positive vector $\mathbf{s}_\alpha = [s_1, s_2, s_3, s_4]^\top$ s.t. $\mathbf{H}_\alpha\mathbf{s}_\alpha \leq (1 - \frac{\mu\alpha}{2})\mathbf{s}_\alpha$, i.e.,

$$\frac{2\lambda^2\alpha^2L^2}{1-\lambda^2}s_4 \leq \left(\frac{1-\lambda^2}{2} - \frac{\mu\alpha}{2}\right)s_1 \quad (63)$$

$$\alpha L s_1 + \frac{\alpha^2L^2}{n}s_3 \leq \frac{\mu\alpha}{2}s_2 \quad (64)$$

$$\left(8m\alpha^2L^2 + \frac{9}{4m}\right)s_1 + 16m\alpha^2L^2s_2 \leq \frac{1-2m\mu\alpha}{4m}s_3 \quad (65)$$

$$\frac{31}{1-\lambda^2}s_1 + \frac{64\lambda^2\alpha^2L^2}{1-\lambda^2}s_2 + \frac{97}{8}s_3 \leq \frac{1-\lambda^2-\mu\alpha}{2}s_4. \quad (66)$$

We first note that (64) is equivalent to $\frac{\alpha L^2}{n}s_3 \leq \frac{\mu}{2}s_2 - Ls_1$, based on which we set the values of s_1, s_2, and s_3 as

$$s_1 = 1/(4\kappa), \qquad s_2 = 1, \qquad s_3 = n/(4\alpha\kappa L) \quad (67)$$

where $\kappa = L/\mu$. Next, we write (65) equivalently as

$$8m\alpha^2L^2\left(s_1 + 2s_2\right) \leq \frac{1-2m\mu\alpha}{4m}s_3 - \frac{9}{4m}s_1. \quad (68)$$

According to (68), we enforce $0 < \alpha \leq \frac{1}{4m\mu}$, i.e., $\frac{1-2m\mu\alpha}{4m} \geq \frac{1}{8m}$; therefore, to make (68) hold, with the help of the values of s_1, s_2, and s_3 in (67), it suffices to further choose α such that

$$18m\alpha^2L^2 \leq \frac{1}{16m\kappa}\left(\frac{n}{2\alpha L} - 9\right). \quad (69)$$

According to (69), we enforce $0 < \alpha \leq \frac{n}{36L}$, i.e., $\frac{n}{2\alpha L} - 9 \geq \frac{n}{4\alpha L}$, and therefore, to make (69) hold, it suffices to further choose α such that $0 < \alpha \leq \frac{n^{1/3}}{10.5m^{2/3}\kappa^{1/3}L}$. Next, according to (66), we further enforce $0 < \alpha \leq \frac{1-\lambda^2}{2\mu}$, i.e., $\frac{1-\lambda^2-\mu\alpha}{2} \geq \frac{1-\lambda^2}{4}$, and therefore, to make (66) hold, we set s_4 as

$$s_4 = \frac{124}{(1-\lambda^2)^2}s_1 + \frac{256\lambda^2\alpha^2L^2}{(1-\lambda^2)^2}s_2 + \frac{97}{2(1-\lambda^2)}s_3. \quad (70)$$

Finally, since $0 < \alpha \leq \frac{1-\lambda^2}{2\mu}$, to make (63) hold, it suffices to further choose α such that $\frac{8\lambda^2\alpha^2L^2}{(1-\lambda^2)^2}\frac{s_4}{s_1} \leq 1$, which, using the values of s_1 and s_4, becomes

$$\frac{992\lambda^2\alpha^2L^2}{(1-\lambda^2)^4} + \frac{8192\kappa\lambda^4\alpha^4L^4}{(1-\lambda^2)^4} + \frac{388\lambda^2n\alpha L}{(1-\lambda^2)^3} \leq 1. \quad (71)$$

If $0 < \alpha \leq \min\{\frac{(1-\lambda^2)^2}{55\lambda}, \frac{1-\lambda^2}{13\lambda\kappa^{1/4}}, \frac{(1-\lambda^2)^3}{388\lambda^2 n}\}\frac{1}{L}$, then the terms on the LHS of (71) are, respectively, less than $\frac{1}{3}$, and thus, (71) holds. Based on the above derivations and Lemma 13, we have: if $0 < \alpha \leq \overline{\alpha}_2$ for $\overline{\alpha}_2$ defined in Theorem 2, then $\rho(\mathbf{H}_\alpha) \leq 1 - \frac{\mu\alpha}{2}$, which concludes the proof. ■

V. Conclusion

In this article, we analyze **GT-SAGA**, a decentralized randomized incremental gradient method that combines node-level variance reduction and network-level gradient tracking. For both general smooth nonconvex problems and problems, where the global function additionally satisfies the PL condition, we prove that **GT-SAGA** achieves fast convergence rate. We further identify practical regimes, where **GT-SAGA** outperforms the existing approaches. We also present numerical simulations to verify the theoretical results in this article. Future research includes generalization of **GT-SAGA** to the setting of time-varying directed networks [60] and of zeroth-order gradient computation [50], [61], [62]. It is also of interest to incorporate weighted sampling techniques [63] in **GT-SAGA** to improve the dependence of the smoothness parameters of the component functions on the convergence rate.

References

[1] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. AC-31, no. 9, pp. 803–812, Sep. 1986.

[2] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5330–5340.

[3] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 344–353.

[4] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA, USA: SIAM, 2017.

[5] R. Xin, S. Pu, A. Nedić, and U. A. Khan, "A general framework for decentralized optimization with first-order methods," *Proc. IEEE*, vol. 108, no. 11, pp. 1869–1889, Nov. 2020.

[6] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[7] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3575–3605, Jun. 2012.

[8] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

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

[10] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 102–113, May 2020.

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

[12] K. Yuan, S. A. Alghunaim, B. Ying, and A. H. Sayed, "On the influence of bias-correction on distributed stochastic optimization," *IEEE Trans. Signal Process.*, vol. 68, pp. 4352–4367, 2020.

[13] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4494–4506, Sep. 2019.

[14] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2: Decentralized training over decentralized data," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4848–4856.

[15] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "DLM: Decentralized linearized alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 4051–4064, Aug. 2015.

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

[17] P. D. Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Netw. Process.*, vol. 2, no. 2, pp. 120–136, Jun. 2016.

[18] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, Sep. 2018.

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

[20] S. Pu and A. Nedich, "Distributed stochastic gradient tracking methods," *Math. Prog.*, vol. 187, pp. 409–457, 2021.

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

[22] M. Maros and J. Jaldén, "A geometrically converging dual method for distributed optimization over time-varying graphs," *IEEE Trans. Autom. Control*, vol. 66, no. 6, pp. 2465–2479, Jun. 2021.

[23] H. Wai, J. Lafond, A. Scaglione, and E. Moulines, "Decentralized Frank-Wolfe algorithm for convex and nonconvex problems," *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 5522–5537, Nov. 2017.

[24] D. Jakovetić, "A unification and generalization of exact distributed first-order methods," *IEEE Trans. Signal Inf. Process. Netw. Process.*, vol. 5, no. 1, pp. 31–46, Mar. 2019.

[25] S. A. Alghunaim, E. Ryu, K. Yuan, and A. H. Sayed, "Decentralized proximal gradient algorithms with linear convergence rates," *IEEE Trans. Autom. Control*, vol. 66, no. 6, pp. 2787–2794, Jun. 2021.

[26] J. Xu, Y. Tian, Y. Sun, and G. Scutari, "Distributed algorithms for composite optimization: Unified framework and convergence analysis," *IEEE Trans. Signal Process.*, 2021.

[27] X. Wu and J. Lu, "A unifying approximate method of multipliers for distributed composite optimization," 2020, *arXiv:2009.12732*.

[28] S. Chen, A. Garcia, and S. Shahrampour, "On distributed non-convex optimization: Projected subgradient method for weakly convex problems in networks," *IEEE Trans. Autom. Control*, to be published, doi: 10.1109/TAC.2021.3056535.

[29] S. A. Alghunaim and A. H. Sayed, "Distributed coupled multiagent stochastic optimization," *IEEE Trans. Autom. Control*, vol. 65, no. 1, pp. 175–190, Jan. 2020.

[30] I. Notarnicola, Y. Sun, G. Scutari, and G. Notarstefano, "Distributed big-data optimization via block-wise gradient tracking," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2045–2060, May 2021.

[31] S. Pu, A. Olshevsky, and I. C. Paschalidis, "A sharp estimate on the transient time of distributed stochastic gradient descent," 2019, *arXiv:1906.02702*.

[32] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—Part I: Agreement at a linear rate," *IEEE Trans. Signal Process.*, vol. 69, pp. 1242–1256, 2021.

[33] B. Swenson, R. Murray, S. Kar, and H. V. Poor, "Distributed stochastic gradient descent and convergence to local minima," 2020, *arXiv:2003.02818*.

[34] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1646–1654.

[35] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 315–323.

[36] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takac, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2613–2621.

[37] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 689–699.

[38] J. Konevcnỳ and P. Richtárik, "Semi-stochastic gradient descent methods," *Front. Appl. Math. Statist.*, vol. 3, p. 9, 2017.

[39] A. Mokhtari and A. Ribeiro, "DSA: Decentralized double stochastic averaging gradient algorithm," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2165–2199, 2016.

[40] K. Yuan, B. Ying, J. Liu, and A. H. Sayed, "Variance-reduced stochastic learning by networked agents under random reshuffling," *IEEE Trans. Signal Process.*, no. 2, pp. 351–366, Jan. 2019.
[41] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *IEEE Trans. Signal Process.*, vol. 68, pp. 6255–6271, 2020.
[42] B. Li, S. Cen, Y. Chen, and Y. Chi, "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction," *J. Mach. Learn. Res.*, vol. 21, no. 180, pp. 1–51, 2020.
[43] H. Li, Z. Lin, and Y. Fang, "Optimal accelerated variance reduced EXTRA and DIGing for strongly convex and smooth decentralized optimization," 2020, *arXiv:2009.04373*.
[44] H. Sun, S. Lu, and M. Hong, "Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9217–9228.
[45] R. Xin, U. A. Khan, and S. Kar, "Fast decentralized non-convex finite-sum optimization with recursive variance reduction," *SIAM J. Optim.*, 2021.
[46] S. J. Reddi, S. Sra, B. Póczos, and A. Smola, "Fast incremental method for smooth nonconvex optimization," in *Proc. IEEE Conf. Decis. Control*, 2016, pp. 1971–1977.
[47] M. Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo, "On the convergence rate of incremental aggregated gradient algorithms," *SIAM J. Optim.*, vol. 27, no. 2, pp. 1035–1048, 2017.
[48] H. Wai, W. Shi, C. A. Uribe, A. Nedić, and A. Scaglione, "Accelerating incremental gradient optimization with curvature information," *Comput. Optim. Appl.*, vol. 76, no. 2, pp. 347–380, 2020.
[49] A. Mokhtari, M. Gurbuzbalaban, and A. Ribeiro, "Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate," *SIAM J. Optim.*, vol. 28, no. 2, pp. 1420–1447, 2018.
[50] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 1, pp. 269–281, Mar. 2021.
[51] X. Yi, S. Zhang, T. Yang, K. H. Johansson, and T. Chai, "Linear convergence for distributed optimization under the Polyak-Lojasiewicz condition," 2019, *arXiv:1912.12110*.
[52] S. J. Reddi, S. Sra, B. Poczos, and A. J. Smola, "Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1145–1153.
[53] L. Zhao, M. Mammadov, and J. Yearwood, "From convex to nonconvex: A loss function analysis for binary classification," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2010, pp. 1281–1288.
[54] B. T. Polyak, *Introduction to Optimization*, vol. 1. New York, NY, USA: Optimization Software, Inc., Publications Division, 1987.
[55] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. New York, NY, USA: Springer, 2016, pp. 795–811.
[56] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "A primal-dual SGD algorithm for distributed nonconvex optimization," 2020, *arXiv:2006.03474*.
[57] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proc. IEEE*, vol. 106, no. 5, pp. 953–976, May 2018.
[58] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[59] D. Williams, *Probability With Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.
[60] F. Saadatniaki, R. Xin, and U. A. Khan, "Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices," *IEEE Trans. Autom. Control*, vol. 65, no. 11, pp. 4769–4780, Nov. 2020.
[61] A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, "Distributed zeroth order optimization over random networks: A Kiefer-Wolfowitz stochastic approximation approach," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 4951–4958.
[62] Y. Pang and G. Hu, "Randomized gradient-free distributed optimization methods for a multiagent system with unknown cost function," *IEEE Trans. Autom. Control*, vol. 65, no. 1, pp. 333–340, Jan. 2020.
[63] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM J. Optim.*, vol. 24, no. 4, pp. 2057–2075, 2014.

Ran Xin received the B.S. degree in mathematics and applied mathematics from Xiamen University, Xiamen, China, in 2016, and the M.S. degree in electrical and computer engineering from Tufts University, Medford, MA, USA, in 2018. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA.

His research interests include convex and nonconvex optimization, stochastic approximation, and machine learning.

Usman A. Khan received the B.S. degree from the University of Engineering and Technology, Lahore, Pakistan, in 2002, the M.S. degree from the University of Wisconsin–Madison, Madison, WI, USA, in 2004, and the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, USA, in 2009, all in electrical and computer engineering.

He is currently an Associate Professor of Electrical and Computer Engineering with Tufts University, Medford, MA, USA. His research interests include statistical signal processing, network science, and decentralized optimization over multiagent systems.

Dr. Khan is an Associate Editor for IEEE CONTROL SYSTEM LETTERS, IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, and IEEE OPEN JOURNAL OF SIGNAL PROCESSING.

Soummya Kar received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 2005, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2010.

From June 2010 to May 2011, he was a Postdoctoral Research Associate with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. He is currently a Professor of Electrical and Computer Engineering with Carnegie Mellon University. His research interests include several aspects of decision making in large-scale networked dynamical systems with applications to problems in network science, cyber-physical systems, and energy systems.