Variance-Reduced Decentralized Stochastic Optimization with Accelerated Convergence

Ran Xin†, Usman A. Khan‡, and Soumya Kar†
†Carnegie Mellon University, Pittsburgh, PA ‡Tufts University, Medford, MA

Abstract—This paper studies decentralized optimization to minimize a finite-sum of functions available over a network of \( n \) nodes. We propose two efficient algorithms, namely \( \text{GT-SAGA} \) and \( \text{GT-SVRG} \), that leverage decentralized stochastic gradient-tracking methods and variance-reduction techniques. We show that both algorithms achieve accelerated linear convergence for smooth and strongly-convex functions. We further describe the regimes in which \( \text{GT-SAGA} \) and \( \text{GT-SVRG} \) achieve non-asymptotic network-independent convergence rates that are \( n \) times faster than that of the centralized SAGA and SVRG and exhibit superior performance (in terms of the number of parallel local component gradient computations required) with respect to the existing decentralized schemes. We also highlight different trade-offs between the proposed algorithms in various settings.

Index Terms—Decentralized optimization, stochastic first-order gradient methods, variance reduction, multi-agent systems.

I. INTRODUCTION

We consider \( n \) nodes connected over a static and directed communication graph \( \mathcal{G} \) such that each node \( i \) has access to a local (possibly private) cost function \( f_i : \mathbb{R}^p \rightarrow \mathbb{R} \). The goal of the networked nodes is to cooperatively solve \( \min_{\mathbf{x} \in \mathbb{R}^p} \sum_{i=1}^n f_i(\mathbf{x}) \), where each node \( i \) is only allowed to process its own local function and to send/receive information with its neighboring nodes. This formulation is well-known as decentralized optimization [3]–[6] that has been studied extensively by the signal processing and control communities over the past decade. Related literature on decentralized optimization methods include the early work on Decentralized Gradient Descent (DGD) [4]–[6], dual averaging [7], [8], and ADMM [9]–[12].

More recently, significant effort has been made to design first-order gradient methods that achieve exact linear convergence for smooth and strongly-convex functions. Examples of such approaches include: primal methods, i.e., EXTRA [13], [14], Exact Diffusion [15], [16], and DLM [17]; methods based on gradient-tracking [18]–[25] and AB/Push-Pull [30], [31]; and methods based on dual decomposition, [32]–[34].

In this paper, we focus on the following refined formulation of decentralized optimization problems:

\[
P_1: \min_{\mathbf{x} \in \mathbb{R}^p} \sum_{i=1}^n f_i(\mathbf{x}), \quad f_i(\mathbf{x}) = \frac{1}{m_i} \sum_{j=1}^{m_i} f_{i,j}(\mathbf{x}),
\]

where each local cost \( f_i \) is the average of \( m_i \) component functions \( \{f_{i,j}\}_{j=1}^{m_i} \). This formulation is motivated by recent data-science and machine learning applications, where large amounts of data is collected by and/or distributed over a network of nodes that aim to cooperatively train a model \( \mathbf{x} \in \mathbb{R}^p \) utilizing data across all nodes. However, when the local data batch at each node is very large, i.e., \( m_i \gg 1 \), full gradient computation becomes prohibitively expensive. In order to avoid extensive computation, stochastic optimization methods use a random subset of the local data for gradient computation. Several stochastic variants of DGD, EXTRA, Exact Diffusion and gradient tracking methods have been recently studied [35]–[45]. These methods converge sub-linearly and outperform their deterministic counterparts when local data batches are large and low-precision solutions suffice [46].

Finite-sum optimization of the form in Problem \( P_1 \) has also been extensively studied in the centralized settings, where various variance-reduction (VR) techniques have been developed to accelerate the standard Stochastic Gradient Descent (SGD) [47]. Well-known approaches include SAG [48], SVRG [49], [50], SAGA [51], S2GD [52], and SARAH [53]. These methods achieve accelerated linear convergence to the minimizer of smooth and strongly-convex functions, while maintaining comparable low per-iteration computation cost as the plain-vanilla SGD. It is therefore natural to introduce VR to the decentralized scenarios of interest in this paper in order to improve the convergence and complexity aspects.

Main contributions: We develop a novel class of decentralized, first-order methods that systematically integrate the aforementioned gradient tracking and variance reduction techniques. In particular, we describe \( \text{GT-SAGA} \) and \( \text{GT-SVRG} \), based on SAGA [51] and SVRG [49], and show that both algorithms achieve accelerated linear convergence for smooth and strongly-convex objective functions. We further discuss the scenarios where \( \text{GT-SAGA} \) and \( \text{GT-SVRG} \) achieve non-asymptotic network-independent convergence rates and exhibit linear speedup (in terms of the total number of nodes) compared with their centralized counterparts. To the best of our knowledge, this is the first work to show network-independent convergence and linear speedup for decentralized VR approaches without requiring computationally-expensive dual gradients or proximal mappings of the objective functions. A detailed comparison of the proposed algorithms with existing decentralized VR methods is provided in Section III-A.

1See [28], [56], [38], [45], [54] where network independence and linear speedup have been shown but without variance-reduction.

RX and SK are with the Electrical and Computer Engineering (ECE) Dept. at Carnegie Mellon University, \{ranx, soumyyak\}@andrew.cmu.edu. UAK is with the ECE Dept. at Tufts University, khan@ece.tufts.edu. The work of SK and RX has been partially supported by NSF under award #1513936. The work of UAK has been partially supported by NSF under awards #1350264, #1903972, and #1935555. Preliminary arXiv versions of this paper appeared in [1], [2].
Basic notation: We use lowercase bold letters to denote vectors and \( ||\cdot|| \) to denote the Euclidean norm of a vector. The matrix \( I_d \) is the \( d \times d \) identity, and \( L_d \) is the \( d \)-dimensional column vector of all ones. For two matrices \( X, Y \in \mathbb{R}^{d \times d} \), \( X \otimes Y \) denotes their Kronecker product. The spectral radius of a matrix \( X \) is denoted by \( \rho(X) \) while its spectral norm is denoted by \( \|X\|_2 \). For a positive vector \( w = [w_1, \ldots, w_d]^\top \) and an arbitrary vector \( x = [x_1, \ldots, x_d]^\top \), the weighted infinity norm of \( x \) is defined as \( \|x\|_w = \max_i |x_i|/w_i \) and \( \|\cdot\|_w \) is the weighted matrix norm induced by the vector norm \( \|\cdot\|_w \). A distributed stochastic iterative algorithm is said to achieve \( \epsilon \)-optimal solution with \( T_{\epsilon} \) iterations, if \( \mathbb{E}[\|x^k - x^*\|^2] \leq \epsilon, \forall k \geq T_{\epsilon} \), where \( x^k \) is the estimate of the optimal solution \( x^* \) at node \( i \) and time \( k \).

Structure of the paper: Section II develops the class of decentralized VR algorithms proposed in this paper while Section III presents the main convergence results and a comparison with the state-of-the-art both in theory and numerical simulations. Section IV presents a unified approach to cast and analyze the proposed algorithms. Sections V and VI contain the convergence analysis for GT-SAGA and GT-SVRG, respectively, and Section VII concludes the paper.

II. ALGORITHM DEVELOPMENT

We first review the standard decentralized gradient descent with gradient tracking (GT-DGD) \[19\]–[24], [30], [51] to solve Problem P1. GT-DGD iteratively updates two vectors at each node \( i \), i.e., \( x^k_i \in \mathbb{R}^p \), the local estimate of the global optimal solution, and \( y^k_i \in \mathbb{R}^p \), the local tracker of the global gradient, initialized with an arbitrary \( x^0_i \) and \( y^0_i = \nabla f_i(x^0_i) \):

\[
\begin{align*}
x^{k+1}_i &= \sum_{r=1}^n w_{ir} x^k_r - \alpha y^k_i, \quad (1a) \\
y^{k+1}_i &= \sum_{r=1}^n w_{ir} y^k_r + \nabla f_i(x^{k+1}_r) - \nabla f_i(x^k_i), \quad (1b)
\end{align*}
\]

where \( \bar{W} = \{w_{ir}\} \) is a doubly-stochastic weight matrix. Note that (1b) tracks the global gradient, i.e., \( y^k_i \rightarrow \nabla f_i(x^k_i) \), as \( k \rightarrow \infty \), and enables linear convergence to the optimal for smooth and strongly-convex functions with a constant step-size \[53\].

When each node has a large amount of data samples, GT-DGD is not practically feasible due to the expensive computation of the local full gradient \( \nabla f_i \) at each iteration. GT-DSGD, the stochastic variant of GT-DGD, replaces each local full gradient with a subset of randomly sampled component gradients \[39\], [40], [42\]. However, GT-DSGD converges sublinearly and, in practice, requires a carefully tuned sequence of decaying step-sizes to ensure a satisfactory performance, potentially due to the large variance of the stochastic gradients. Inspired by centralized VR techniques \[48\]–[53], a class of decentralized VR methods can be obtained by replacing each local full gradient \( \nabla f_i(x^k_i) \) with its estimator whose variance progressively reduces to 0. Clearly all existing VR schemes are applicable here, including \[48\]–[53]. In this paper, we explore two popular VR methods, i.e., SAGA \[51\] and SVRG \[49\]. The resulting algorithms are formally described next.

GT-SAGA, formally described in Algorithm 1, implements the gradient tracking \( y^k_i \) on the estimate \( g^k_i \) of the local batch gradient at each node. The auxiliary variable \( z^k_{i,j} \) is the most recent iterate at which the component gradient \( \nabla f_{i,j} \) was evaluated before time \( k \). At each time \( k \), node \( i \) draws an index \( x_i^k \) to randomly select one component function from its local data batch. To implement the SAGA gradient estimator, each node must maintain a table of all local component gradients. After \( g^k_i \) is updated, the \( \nabla f_{i,s} \) entry in the gradient table is replaced by \( \nabla f_{i,k}(x^k_i) \), while the others remain unchanged. This implementation procedure results in a storage cost of \( O(m_i) \) at each node \( i \) that can be improved to \( O(m_i) \) for certain problems with favorable structures \[48\].

Algorithm 1 GT-SAGA at each node \( i \)

Require: \( x^0_i; z^0_{i,j} = x^0_i, \forall j \in \{1, \ldots, m_i\}; \alpha; \{\bar{w}_{ir}\}_{r=1}^n; \)

1: for \( k = 0, 1, 2, \ldots \) do

2: Update the local estimate of the solution:

\[
x^{k+1}_i = \sum_{r=1}^n \bar{w}_{ir} x^k_r - \alpha y^{k}_i;
\]

3: Select \( s^{k+1}_i \) uniformly at random from \( \{1, \ldots, m_i\} \);

4: Update the local stochastic gradient estimator:

\[
g^{k+1}_i = \nabla f_{i,s^{k+1}_i}(x^{k+1}_i) - \nabla f_{i,s^{k+1}_i}(z^{k+1}_{i,s^{k+1}_i}) + \frac{1}{m_i} \sum_{j=1}^{m_i} \nabla f_{i,j}(z^{k+1}_{i,j});
\]

5: if \( j = s^{k+1}_i \), then \( z^{k+2}_{i,j} = x^{k+1}_i \); else \( z^{k+2}_{i,j} = z^{k+1}_{i,j} \).

6: Update the local gradient tracker:

\[
y^{k+1}_i = \sum_{r=1}^n \bar{w}_{ir} y^{k}_r + g^{k+1}_i - g^k_i;
\]

7: end for

GT-SVRG is formally described in Algorithm 2. In contrast to GT-SAGA that computes an estimate \( g^k_i \) of \( \nabla f_i \) from a random sample of the local data batch at every iteration, GT-SVRG realizes VR by periodically computing full \( \nabla f_i \) from the entire local data batch. GT-SVRG can thus be considered as a “double-loop” method, where each node \( i \), at every outer-loop update \( \{x^T_i\} \rightarrow T \geq 0 \), calculates a local full gradient \( \nabla f_i(x^T_i) \) that is retained in the subsequent inner-loop iterations to update the local gradient estimator \( y^T_i \): for \( k \in \{T, (T + 1)T - 1\} \)

\[
y^{k+1}_i = \nabla f_{i,s^{k+1}_i}(x^{k+1}_i) - \nabla f_{i,s^{k+1}_i}(x^{T}_i) + \nabla f_{i,s^{k+1}_i}(x^{T}_i);
\]

Clearly, GT-SVRG eliminates the need of storing all local component gradients at each node and thus has a favorable storage cost compared to GT-SAGA. However, this advantage comes at the expense of evaluating two component gradients \( \nabla f_{i,s^{k+1}_i}(x^{k+1}_i) \) and \( \nabla f_{i,s^{k+1}_i}(x^{T}_i) \) at every iteration, in addition to calculating the full local gradient \( \nabla f_i \) periodically. See Remarks 1 and 2 for additional discussion.

III. MAIN RESULTS

The convergence results for GT-SAGA and GT-SVRG are based on the following assumptions.

Assumption 1. The global objective function \( F \) is \( \mu \)-strongly-convex, i.e., \( \forall x, y \in \mathbb{R}^p \) and for some \( \mu > 0 \), we have

\[
F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2.
\]

We note that under Assumption 1, the global objective function \( F \) has a unique minimizer, denoted as \( x^* \).
Algorithm 2 GT–SVRG at each node $i$

Require: $x_i^0, \tau_i^0 = x_i^0; \alpha; \{ w_{ir} \}_{r=1}^n; T; y_i^0 = \nabla f_i(x_i^0)$.
1: for $k = 0, 1, 2, \ldots$ do
2: Update the local estimate of the solution:
   $$x_i^{k+1} = \sum_{r=1}^n w_{ir} x_r^k - \alpha y_i^k;$$
3: Select $k_i^{k+1}$ uniformly at random from $\{ 1, \ldots, m_i \}$;
4: If $\text{mod}(k+1, T) = 0$ then $\tau_i^{k+1} = x_i^{k+1}$; else $\tau_i^{k+1} = \tau_i^k$.
5: Update the local stochastic gradient estimator:
   $$v_i^{k+1} = \nabla f_{i,k_i^{k+1}}(x_i^{k+1}) - \nabla f_{i,k_i^{k+1}}(\tau_i^{k+1})$$
   $$+ \nabla f_i(\tau_i^{k+1});$$
6: Update the local gradient tracker:
   $$y_i^{k+1} = \sum_{r=1}^n w_{ir} y_r^k + v_i^{k+1} - v_i^k;$$
7: end for

Assumption 2. Each local objective function $f_{i,j}$ is $L$-smooth, i.e., $\forall x, y \in \mathbb{R}^p$ and for some $L > 0$, we have
$$\| \nabla f_{i,j}(x) - \nabla f_{i,j}(y) \| \leq L \| x - y \|.$$ Clearly, under Assumption 2 the global objective $F$ is also $L$-smooth and $L \geq \mu$. We use $Q = L/\mu$ to denote the condition number of the global objective function $F$.

Assumption 3. The weight matrix $\tilde{W} = \{ w_{ir} \}$ associated with the network $\tilde{G}$ is primitive and doubly-stochastic.

Assumption 3 is satisfied by strongly-connected and weight-balanced directed graphs that admit doubly-stochastic weights. This assumption implies that the second largest singular value of $\tilde{W}$ is less than 1, i.e., $\sigma = \| \tilde{W} - \frac{1}{\tilde{n}} \mathbf{1}_n \mathbf{1}_n^T \| < 1$.

We denote $M := \max_i m_i$ and $m := \min_i m_i$, where $m_i$ is the number of local component functions at node $i$. The main convergence results of GT–SAGA and GT–SVRG are summarized respectively in the following theorems.

Theorem 1 (GT–SAGA). Let Assumptions 7, 2 and 3 hold. If the step-size $\alpha$ in GT–SAGA is such that
$$0 < \alpha \leq \frac{1}{\kappa} := \min \left\{ \mathcal{O} \left( \frac{1}{m^3} \right), \mathcal{O} \left( \frac{m}{M} \frac{(1-\sigma^2)^2}{1-L^2} \right) \right\},$$
then GT–SAGA linearly converges to the optimal solution $x^*$. If $\alpha = \frac{1}{\kappa}$, then GT–SAGA achieves $\epsilon$-optimal solution in
$$\mathcal{O} \left( \max \left\{ M, \frac{m}{M} \frac{Q^2}{1-\sigma^2} \right\} \log \frac{1}{\epsilon} \right)$$
iterations (parallel local component gradient computations).

Theorem 2 (GT–SVRG). Let Assumptions 7, 2 and 3 hold. If the step-size $\alpha$ and the length $T$ of the inner-loop are such that
$$\alpha = \mathcal{O} \left( \frac{(1-\sigma^2)^2}{1-L^2} \right), \quad T = \mathcal{O} \left( \frac{Q^2 \log Q}{1-\sigma^2} \right),$$
then GT–SVRG achieves $\epsilon$-optimal solution in
$$\mathcal{O} \left( \left( M + \frac{Q^2 \log Q}{1-\sigma^2} \right) \log \frac{1}{\epsilon} \right)$$
iterations (parallel local component gradient computations).

The formal proofs of these theorems are developed in the next sections. In particular, Section 4 presents a dynamical system approach that unifies the proposed algorithms and develops the results that are common to both. Next, Section 5 discusses GT–SAGA and builds the proof of Theorem 1 while the analysis of GT–SVRG and the proof of Theorem 2 is completed in Section 6. We discuss some salient features of the proposed algorithms next and compare them with the state-of-the-art both in theory and numerical simulations.

Remark 1. (Network-independent convergence rates). In the “big-data” regimes when each node has a large dataset, and when the nodes are well-connected, i.e., $M \approx m \gg \frac{Q^2}{1-\sigma^2}$, both GT–SAGA and GT–SVRG achieve a network-independent convergence rate of $\mathcal{O}(M \frac{1}{\log 1/\epsilon})$, which, in addition, is $n$ times faster than their centralized counterparts, SAGA and SVRG. GT–SAGA and GT–SVRG, therefore, achieve a non-asymptotic linear speedup in terms of the total number of nodes.

Remark 2. (GT–SAGA versus GT–SVRG). It can be observed from Theorem 1 and 2 that when data samples are unevenly distributed across the nodes, i.e., $\frac{M}{m} \gg 1$, GT–SVRG achieves a lower iteration complexity than GT–SAGA. However, an uneven data distribution may adversely impact the practical implementation of GT–SVRG as well. This is because GT–SVRG requires a highly synchronized communication network as all nodes need to evaluate their local full gradients every $T$ iterations and cannot proceed to the next inner-loop until all nodes complete their local computation. Therefore, the nodes with small data samples have a relatively long idle time at the end of each inner-loop that leads to an increase in overall wall-clock time. Indeed, the inherent trade-off between GT–SAGA and GT–SVRG is the network synchrony versus the gradient storage. For structured problems, where the component gradients can be stored efficiently, GT–SAGA may be preferred due to its flexibility of implementation and less dependence on the network synchrony. Conversely, if the problem of interest is large-scale, i.e., $m$ is very large, and storing all component gradients is not feasible, GT–SVRG may become a more appropriate choice.

A. Comparison with Related Work

Existing variance-reduced decentralized stochastic optimization methods include: DSA [57] that integrates EXTRA [13] with SAGA [51] and was the first decentralized VR method; Diffusion-AVRG that combines Exact Diffusion [15] and AVRG [59]; DSBA [60] that uses proximal mapping [61] to accelerate DSA; Ref. [62] that applies edge-based method [63] to DSA; ADFS [64] that is the decentralized version of the accelerated randomized proximal coordinate

| Algorithm       | Convergence Rate                        |
|-----------------|-----------------------------------------|
| Acc-DNGD [26]   | $\mathcal{O} \left( \frac{m \sigma^2}{\sqrt{t}} \log \frac{1}{\epsilon} \right)$ |
| DSA [57]        | $\mathcal{O} \left( \max \left\{ \frac{mQ^2}{1-\sigma^2}, \frac{1}{1-\sigma^2} \right\} \log \frac{1}{\epsilon} \right)$ |
| Diffusion-AVRG  | linear (no explicit rate)               |
| GT–SAGA (this work) | $\mathcal{O} \left( \max \left\{ \frac{mQ^2}{1-\sigma^2}, \frac{1}{1-\sigma^2} \right\} \log \frac{1}{\epsilon} \right)$ |
| GT–SVRG (this work) | $\mathcal{O} \left( \left( M + \frac{Q^2 \log Q}{1-\sigma^2} \right) \log \frac{1}{\epsilon} \right)$ |
gradient method [65] based on the dual of Problem P1. We compare the iteration complexity of \textit{GT-SAGA} and \textit{GT-SVRG} with several state-of-the-art first-order primal methods that solve Problem P1 in Table 1, where, for the simplicity of presentation, we assume that all nodes have the same number \( \tilde{m} \) of local functions, i.e., \( M = m = \tilde{m} \). Clearly, \textit{GT-SAGA} and \textit{GT-SVRG} improve upon the convergence rates in terms of the joint dependence on \( Q \) and \( \tilde{m} \), especially in the “big-data” scenarios where \( \tilde{m} \) is very large, with the exception of DSBA [60] and ADFS [64], both of which achieve better iteration complexity albeit at the expense of computing the proximal mapping of a component function at each iteration.

\[ f(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \log \left( 1 + e^{-\langle x^T \theta_{ij}, \xi_{ij} \rangle} \right) + \frac{\lambda}{2} \|x\|^2, \]

where \( \theta_{ij} \in \mathbb{R}^{785} \) is the feature vector and \( \xi_{ij} \in \{-1, +1\} \) is the corresponding binary label. We compare the performance of the proposed algorithms, \textit{GT-SAGA} and \textit{GT-SVRG}, with state-of-the-art primal first-order methods, including the accelerated decentralized Nesterov gradient descent (Acc-DNGD) [26], GT-DSGD [39] (with decaying step-sizes to ensure exact convergence), DSA [57] and Diffusion-AVRG [58]. For comparison, we consider the performance metric of the average residual \( \frac{1}{n} \sum_{i=1}^{n} \|x^k_i - x^*\|^2 \) versus number of local epochs (effective passes of local data set). The experimental results are shown in Fig. 1.

We observe that GT-DSGD, which converges sublinearly, progresses very fast at the beginning and then drastically slows down, similar to the centralized SGD. On the other hand, \textit{GT-SAGA} and \textit{GT-SVRG} achieve accelerated linear convergence that is faster than Acc-DNGD that adds Nesterov momentum [63] to the GT-DGD framework. This is due to the fact that in contrast to Acc-DNGD, \textit{GT-SAGA} and \textit{GT-SVRG} incorporate variance-reduction techniques that exploit the finite-sum structure of each local cost function. This is consistent with the finite-sum optimization in the centralized settings, where VR methods [48]–[53] often outperform Nesterov gradient descent [68] when the number of data samples is large. Although DSA and Diffusion-AVRG are observed to achieve similar practical performance as \textit{GT-SAGA} and \textit{GT-SVRG}, their theoretical guarantee is relatively weak, see Table 1. It is also worth mentioning that SAGA-based methods, i.e., \textit{GT-SAGA} and DSA, are typically faster than \textit{GT-SVRG} and Diffusion-AVRG that execute in cycles, however, at the expense of storing all component gradients.

B. Numerical Experiments

In this subsection, we numerically verify the accelerated linear convergence of \textit{GT-SAGA} and \textit{GT-SVRG} and compare them with the state-of-the-art. We consider the problem of decentralized training of a regularized logistic regression model to classify handwritten digits \( \{0, 7\} \), represented by feature vectors in \( \mathbb{R}^{784} \), from the MNIST dataset. We generate a random, connected and undirected geometric graph of 20 nodes using the nearest neighbor rule and the associated doubly-stochastic weight matrix is generated by the Laplacian method [67]. Each node in the network has \( m_i = 50 \) data samples and we normalize the feature vectors such that they have a mean of 0 and a standard deviation of 1. The networked nodes cooperatively minimize the following decentralized smooth and strongly-convex cost function:

\[ f(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \log \left( 1 + e^{-\langle x^T \theta_{ij}, \xi_{ij} \rangle} \right) + \frac{\lambda}{2} \|x\|^2, \]

where \( \theta_{ij} \in \mathbb{R}^{785} \) is the feature vector and \( \xi_{ij} \in \{-1, +1\} \) is the corresponding binary label.
We recall that [55] is a stochastic gradient tracking method [39], [40], [42] as an application of dynamic consensus [55]. It is straightforward to verify by induction that [55]:

\[ r^k = x^k, \quad \forall k \geq 0. \]

Clearly, the randomness of both GT–SAGA and GT–SVRG lies in the set of independent random variables \( \{s_i^k\}_{i=1}^{l_{\text{max}}} \). We denote \( \mathcal{F}^k \) as the history of the dynamical system generated by \( \{s_i^k\}_{i=1}^{l_{\text{max}}} \). For both GT–SAGA and GT–SVRG, \( r^k \) is an unbiased estimator of \( \nabla f(x^k) \) given \( \mathcal{F}^k \) [49], [51], i.e.,

\[ \mathbb{E}[r^k | \mathcal{F}^k] = \nabla f(x^k), \quad \mathbb{E}[\mathbf{y}^k | \mathcal{F}^k] = \mathbb{E}[r^k | \mathcal{F}^k] = \overline{\nabla f}(x^k). \]

In the following, we first present a few well-known results related to decentralized gradient tracking methods whose proofs can be found in, e.g., [21], [22], [30], [31].

**Lemma 1.** Let Assumption 7 and 2 hold. If \( 0 < \alpha \leq \frac{1}{2} \), we have \( \|x - \alpha \nabla F(x) - x^*\| \leq (1 - \mu \alpha) \|x - x^*\|, \forall x \in \mathbb{R}^p. \)

**Lemma 2.** Let Assumption 2 hold. Consider the iterates \( \{x^k\} \) generated by the dynamical system (2). We have that \( \|\nabla F(x^k) - \nabla F(x^*)\| \leq \frac{\alpha}{\sqrt{n}} \|W - W_\infty x^*\|, \forall k \geq 0. \)

**Lemma 3.** Let Assumption 2 hold. We have that \( \forall x \in \mathbb{R}^{np}, \|Wx - W_\infty x\| \leq \sigma \|W - W_\infty x\|, \) \( \)where \( W_\infty = \frac{1}{n} I_p \).

**B. Auxiliary Results**

In this subsection, we analyze the general dynamical system (2) by establishing the interrelationships between the mean-squared consensus error \( \mathbb{E}[\|x^k - W_\infty x^\star\|^2] \), network optimality gap \( \mathbb{E}[\|\mathbf{x}^k - x^\star\|^2] \) and gradient tracking error \( \mathbb{E}[\|y^k - W_\infty y^\star\|^2] \).

**Lemma 4.** Let Assumption 2 hold. Consider the iterates \( \{x^k\} \) generated by (2). We have the following hold: \( \forall k \geq 0, \)

\[ \mathbb{E}\left[\|x^{k+1} - W_\infty x^{k+1}\|^2\right] \leq \left(\begin{array}{c} 1 + \frac{\sigma^2}{2} \\
 + 2 \alpha^2 - \frac{1}{\sigma} \end{array}\right) \mathbb{E}\left[\|x^k - W_\infty x^k\|^2\right]. \]

**Proof.** Using (2a) and the fact that \( W_\infty W = W_\infty \), we have:

\[ \|x^{k+1} - W_\infty x^{k+1}\|^2 = \|Wx^{k} - W_\infty x^k - \alpha (y^k - W_\infty y^k)\|^2 \]

Next, we use Young’s inequality that \( ||a+b||^2 \leq (1+\eta)||a||^2 + (1+\frac{1}{\eta})||b||^2, \forall a, b \in \mathbb{R}^{np}, \forall \eta > 0, \) and Lemma 3 in (5) to obtain: \( \forall k \geq 0, \)

\[ \|x^{k+1} - W_\infty x^{k+1}\|^2 \leq (1 + \eta \sigma^2) \|x^k - W_\infty x^k\|^2 + (1 + \frac{1}{\eta} \alpha^2) \|y^k - W_\infty y^k\|^2 \]

Setting \( \eta = \frac{1 - \sigma^2}{2} \) and 1 in the above inequality respectively leads to (3) and (4).

Next, we establish an inequality for \( \mathbb{E}[\|x^{k+1} - x^\star\|^2]. \)

**Lemma 5.** Let Assumptions 2 and 2 hold. Consider the iterates \( \{x^k\} \) generated by (2). If \( 0 < \alpha \leq \frac{1}{L} \), we have the following inequalities: \( \forall k \geq 0, \)

\[ \mathbb{E}\left[n \|x^{k+1} - x^\star\|^2\right] \leq \frac{L^2 \alpha}{L_\infty} \mathbb{E}\left[\|x^k - W_\infty x^k\|^2\right] \]

\[ + (1 - \mu \alpha) \mathbb{E}\left[n \|x^k - x^\star\|^2\right] \]

\[ + \left(\begin{array}{c} \frac{\alpha^2}{n} \\mathbb{E}\left[\|r^k - \nabla f(x^k)\|^2\right] \end{array}\right). \]

**Proof.** Multiplying \( \frac{1}{n} I_p \) to (2a), we have that \( \forall k \geq 0, \)

\[ x^{k+1} = x^k - \alpha y^k = x^k - \alpha r^k. \]

We expand \( \mathbb{E}[\|x^{k+1} - x^\star\|^2]|\mathcal{F}^k| \) as follows.

\[ \mathbb{E}\left[n \|x^{k+1} - x^\star\|^2|\mathcal{F}^k\right] = \mathbb{E}\left[\|x^k - \alpha r^k - x^\star\|^2|\mathcal{F}^k\right] \]

\[ = \mathbb{E}\left[\|x^k - \alpha \nabla F(x^k) - x^\star + \alpha (\nabla F(x^k) - r^k)\|^2|\mathcal{F}^k\right] \]

\[ \leq \mathbb{E}\left[\|x^k - \alpha \nabla F(x^k) - x^\star\|^2 + \alpha^2 \mathbb{E}\left[\|
abla \nabla F(x^k) - r^k\|^2|\mathcal{F}^k\right]\right] \]

\[ + 2 \alpha \mathbb{E}\left[\left(\nabla F(x^k) - x^\star, \nabla F(x^k) - \overline{\nabla f}(x^k)\right)\right], \]

where in the last equality we used that \( \mathbb{E}[\nabla \nabla F(x^k)|\mathcal{F}^k] = \overline{\nabla f}(x^k). \)

Next, we expand and simplify \( \mathbb{E}[\|
abla F(x^k) - r^k\|^2|\mathcal{F}^k]\): \( \mathbb{E}\left[\|
abla F(x^k) - r^k\|^2|\mathcal{F}^k\right] \]

\[ = \mathbb{E}\left[\|
abla F(x^k) - \overline{\nabla f}(x^k)\|^2 + \mathbb{E}\left[\|
abla \nabla f(x^k) - r^k\|^2|\mathcal{F}^k\right]\right] \]

where we used the fact that \( \left(\nabla F(x^k) - \overline{\nabla f}(x^k), \mathbb{E}[\nabla \nabla f(x^k)|\mathcal{F}^k] - r^k\right) = 0. \)

For the last term in (9), we have that:

\[ \mathbb{E}\left[\|
abla \nabla F(x^k) - r^k\|^2|\mathcal{F}^k\right] \]

\[ = \frac{1}{n^2} \mathbb{E}\left[\left(\sum_{i=1}^{n} (r_i^k - \nabla \nabla f_i(x_i^k))\right)^2|\mathcal{F}^k\right] \]

\[ = \frac{1}{n^2} \mathbb{E}\left[\|r^k - \nabla \nabla f(x^k)\|^2|\mathcal{F}^k\right], \]

where in the equality above we used the fact that \( \{r_i^k\}_{i=1}^{n} \) are independent from each other given \( \mathcal{F}^k \) and therefore \( \mathbb{E}[\sum_{i \neq j} (r_i^k - \nabla \nabla f_i(x_i^k), r_j^k - \nabla \nabla f_j(x_j^k))|\mathcal{F}^k] = 0. \) Now, we use (9), (10) and Lemma 1 in (8) to obtain:

\[ \mathbb{E}\left[n \|x^{k+1} - x^*\|^2|\mathcal{F}^k\right] \]

\[ \leq (1 - \mu \alpha^2) \|x^k - x^*\|^2 + \alpha^2 \mathbb{E}\left[\|
abla F(x^k) - \overline{\nabla f}(x^k)\|^2\right] \]

\[ + 2 \alpha (1 - \mu \alpha) \mathbb{E}[\|x^k - x^*\|^2|\mathcal{F}^k] \]

\[ + \alpha^2 \mathbb{E}[\|r^k - \nabla \nabla f(x^k)\|^2|\mathcal{F}^k]. \]
Finally, we apply Young’s inequality such that

\[ 2\alpha \| x^k - x^* \| \| \nabla F(x^k) - \nabla T(x^k) \| \leq \mu \alpha \| x^k - x^* \|^2 + \mu^{-1} \alpha \| \nabla F(x^k) - \nabla T(x^k) \|^2 \]

and \( \| \nabla T(x^k) - \nabla F(x^k) \| \leq \frac{1}{\sqrt{\sigma}} \| x^k - W_k x^k \|, \forall k \geq 0 \), from Lemma [2] to (11) and take the total expectation; the resulting inequality is exactly (6). Similarly, using

\[ 2\alpha \| x^k - x^* \| \| \nabla F(x^k) - \nabla T(x^k) \| \leq \| x^k - x^* \|^2 + \alpha^2 \| \nabla F(x^k) - \nabla T(x^k) \|^2 \]

and Lemma [2] in (11) leads to (7).

Next, we derive an inequality for \( \| y^{k+1} - W_k y^{k+1} \|^2 \).

**Lemma 6.** Let Assumption [2] and Assumption [3] hold. Consider the iterates \( \{ y^k \} \) generated by (2). If \( 0 < \alpha \leq \frac{1}{4 \sqrt{2} L} \), we have the following inequality holds: \( \forall k \geq 0 \),

\[
\mathbb{E} \left[ \| y^{k+1} - W_k y^{k+1} \|^2 \right] 
\leq \frac{33L^2}{1 - \sigma^2} \mathbb{E} \left[ \| x^k - W_k x^k \|^2 \right] + \frac{L^2}{1 - \sigma^2} \mathbb{E} \left[ \| x^k - x^* \|^2 \right]
\]

\[
+ \left( \frac{1 + \sigma^2}{2} + \frac{32L^2 \alpha^2}{1 - \sigma^2} \right) \mathbb{E} \left[ \| y^k - W_k y^k \|^2 \right]
\]

\[
+ \frac{5}{1 - \sigma^2} \mathbb{E} \left[ \| r^k - \nabla f(x^k) \|^2 \right]
\]

\[
+ \frac{4}{1 - \sigma^2} \mathbb{E} \left[ \| r^{k+1} - \nabla f(x^{k+1}) \|^2 \right].
\]

**Proof.** Using (2b) and the fact that \( W_k W = W \), we have:

\[
\| y^{k+1} - W_k y^{k+1} \|^2 
= \| W_k y^k + r^{k+1} - W_k \| W_k x^k \| + r^{k+1} - r^k \| \|^2
\]

\[
= \| W_k y^k - W_k W_k x^k \| + (I_{n_p} - W_k) \| r^{k+1} - r^k \| \|^2.
\]  (12)

To proceed from (12), we use Young’s inequality that \( \| a + b \|^2 \leq (1 + \eta) \| a \|^2 + (1 + \frac{\eta}{\eta^2}) \| b \|^2 \), \( \forall a, b \in \mathbb{R}^n \) with \( \eta = \frac{\sigma^2}{1 - \sigma^2} \) and that \( I_{n_p} - W_k = 1 \) together with Lemma [3] to obtain:

\[
\| y^{k+1} - W_k y^{k+1} \|^2
\leq \left( 1 + \frac{1 - \sigma^2}{2 \sigma^2} \right) \| W_k y^k - W_k W_k x^k \|^2
\]

\[
+ \left( 1 + \frac{2 \sigma^2}{1 - \sigma^2} \right) \| (I_{n_p} - W_k) (r^{k+1} - r^k) \|^2
\]

\[
\leq \frac{1 + \sigma^2}{2} \| y^k - W_k W_k x^k \|^2
+ \frac{2}{1 - \sigma^2} \| r^{k+1} - r^k \|^2. \]  (13)

We then take the total expectation to obtain:

\[
\mathbb{E} \left[ \| y^{k+1} - W_k y^{k+1} \|^2 \right]
\leq \frac{1 + \sigma^2}{2} \mathbb{E} \left[ \| y^k - W_k W_k x^k \|^2 \right]
+ \frac{2}{1 - \sigma^2} \mathbb{E} \left[ \| r^{k+1} - r^k \|^2 \right]. \]  (14)

Now, we derive an upper bound for \( \mathbb{E} [\| r^{k+1} - r^k \|^2] \). Firstly,

\[
\mathbb{E} \left[ \| r^{k+1} - r^k \|^2 \right]
\leq 2 \mathbb{E} \left[ \| r^{k+1} - r^k - (\nabla f(x^{k+1}) - \nabla f(x^k)) \|^2 \right]
+ 2 \mathbb{E} \left[ \| \nabla f(x^{k+1}) - \nabla f(x^k) \|^2 \right]
\leq 2 \mathbb{E} \left[ \| r^k - \nabla f(x^k) \|^2 \right]
+ 2 \mathbb{E} \left[ \| r^{k+1} - \nabla f(x^{k+1}) \|^2 \right]
+ 2L^2 \mathbb{E} \left[ \| r^{k+1} - r^k \|^2 \right]. \]  (15)

where in the last inequality above we used that

\[
\mathbb{E} \left[ (r^{k+1} - \nabla f(x^{k+1}), r^k - \nabla f(x^k)) \right]
= \mathbb{E} \left[ (r^{k+1} - \nabla f(x^{k+1}), r^k - \nabla f(x^k)) | x^{k+1} \right] = 0.
\]

We next bound \( \mathbb{E} [\| x^{k+1} - x^k \|^2] \). Using (2a) leads to:

\[
\| x^{k+1} - x^k \|^2 = 2 \mathbb{E} \left[ \| x^k - \nabla f(x^k) \| \right]
\]

\[
+ 2 \mathbb{E} \left[ \| x^k - x^* \|^2 \right]
\]

\[
+ 4 \mathbb{E} \left[ \| r^{k+1} - \nabla f(x^{k+1}) \|^2 \right]. \]  (16)

where in (16), we used the fact that \( \| W - I_{n_p} \| \leq 2 \). We then denote \( \nabla f(x^*) = (\nabla f_1(x^*)^T, \ldots, \nabla f_n(x^*)^T) \) and note that \( (1_n \otimes I_p) \nabla f(x^*) = 0_p \). We bound \( \| y^k \| \) as follows.

\[
\| y^k \| = \| y^k - W_k y^k + W_k r^k - W_k \nabla f(x^k) \|
\]

\[
+ W_k \nabla f(x^k) - W_k \nabla f(x^*) \|
\leq \| y^k - W_k y^k \| + \| r^k - \nabla f(x^k) \|
+ L \| x^k - x^* \|
\leq \| y^k - W_k y^k \| + \| r^k - \nabla f(x^k) \|
+ L \| x^k - W_k x^k \|
+ \sqrt{nL} \| x^k - x^* \|. \]  (17)

Using (17) in (16) with the requirement that \( 0 < \alpha \leq \frac{1}{4 \sqrt{2} L} \) and taking the total expectation, we have:

\[
\mathbb{E} \left[ \| x^{k+1} - x^k \|^2 \right]
\leq 8.25 \mathbb{E} \left[ \| x^k - W_k x^k \|^2 \right]
+ 0.25 \mathbb{E} \left[ \| x^k - x^* \|^2 \right]
\]

\[
+ 8 \alpha^2 \mathbb{E} \left[ \| y^k - W_k y^k \|^2 \right]
+ 8 \alpha^2 \mathbb{E} \left[ \| r^k - \nabla f(x^k) \|^2 \right]. \]  (18)

Finally, we apply (18) in (15) with \( 0 < \alpha \leq \frac{1}{4 \sqrt{2} L} \) to obtain:

\[
\mathbb{E} \left[ \| r^{k+1} - r^k \|^2 \right]
\leq 16.5 \mathbb{E} \left[ \| x^k - W_k x^k \|^2 \right]
+ 0.5 \mathbb{E} \left[ \| x^k - x^* \|^2 \right]
+ 16 \alpha^2 \mathbb{E} \left[ \| y^k - W_k y^k \|^2 \right]
+ 2.5 \mathbb{E} \left[ \| r^k - \nabla f(x^k) \|^2 \right]
+ 2 \mathbb{E} \left[ \| r^{k+1} - \nabla f(x^{k+1}) \|^2 \right].
\]

Using the above inequality in (14) completes the proof.
Remark 3. We note that in contrast to SGD-based methods \cite{49-52}, the convergence results of VR methods \cite{48-53} and the ones derived in this paper are independent of the variance of the stochastic gradient. Because of this reason, an explicit bound on the variance of the stochastic gradient is not required as an assumption in the analysis of VR methods.

With the help of the auxiliary results on the general dynamical system \cite{2} established in this section, we now derive explicit convergence rates for the proposed algorithms, GT–SAGA and GT–SVRG, in the next sections.

V. CONVERGENCE ANALYSIS OF GT–SAGA

In this section, we establish the linear convergence of GT–SAGA described in Algorithm \cite{1}. Following the unified representation in \cite{2}, we note that the local gradient estimator \(g_i^k\) in GT–SAGA: \(\forall i \in \{1, \cdots, n\}, \forall k \geq 1,\)

\[
g_i^k = \nabla f_i(s^k_i(x^k_i)) - \nabla f_i(s^k_i(z^k_{i,j})) + \frac{1}{m_i} \sum_{j=1}^{m_i} \nabla f_i(z^k_{i,j}),
\]

where \(s^k_i\) is selected uniformly at random from \(\{1, \cdots, m_i\}\) and the auxiliary variable \(z^k_{i,j}\) is the most recent iterate where the component gradient \(\nabla f_i(z^k_{i,j})\) was evaluated before time \(k\).

A. Bounding the variance of the gradient estimator

We first derive an upper bound for \(\mathbb{E}[\|g_i^k - \nabla f(x^k)\|^2]\) that is the variance of the gradient estimator \(g_i^k\). To do this, we define \(t_i^k\) as the averaged optimality gap of the auxiliary variables of \(\{z^k_{i,j}\}_{j=1}^{m_i}\) at node \(i\) as follows:

\[
t_i^k := \frac{1}{m_i} \sum_{j=1}^{m_i} \|z^k_{i,j} - x^*\|^2, \quad t^k := \sum_{i=1}^{n} t_i^k.
\]

(19)

The following lemma shows that \(t^k\) has an intrinsic contraction property. Recall that \(M = \max_i m_i\) and \(m = \min_i m_i\).

Lemma 7. Consider the iterates \(\{x^k\}\) generated by GT–SAGA. We have the following holds: \(\forall k \geq 1,\)

\[
\mathbb{E}[t^{k+1}] \leq \left(1 - \frac{1}{M}\right) \mathbb{E}[t^k] + \frac{2}{m} \mathbb{E} \left[\|x^k - W_\infty x^k\|^2\right] + \frac{2}{m} \mathbb{E} \left[n \|x^k - x^*\|^2\right].
\]

Proof. Recall Algorithm \cite{1} and note that \(\forall k \geq 1, z^{k+1}_{i,j} = z^k_{i,j}\) with probability \(1 - \frac{1}{m_i}\) and \(z^{k+1}_{i,j} = x^k_i\) with probability \(\frac{1}{m_i}\) given \(F_k\). Then we have the following holds: \(\forall i, \forall k \geq 1,\)

\[
\mathbb{E}[t^{k+1} | F_k] = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{E}[\|z^{k+1}_{i,j} - x^*\|^2 | F_k] = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{E} \left[\left(1 - \frac{1}{m_i}\right) \|z^k_{i,j} - x^*\|^2 + \frac{1}{m_i} \|x^k_i - x^*\|^2\right] | F_k] = \left(1 - \frac{1}{m_i}\right) t_i^k + \frac{1}{m_i} \|x^k_i - x^*\|^2 \leq \left(1 - \frac{1}{M}\right) t_i^k + \frac{2}{m} \|x^k_i - x^*\|^2 + \frac{2}{m} \|x^k - x^*\|^2.
\]

(20)

The proof follows by summing (20) over \(i\) and taking the total expectation.

In the next lemma, we bound the stochastic gradient variance \(\mathbb{E} \left[\|g_i^k - \nabla f(x^k)\|^2\right]\) by the mean-square consensus error and the optimality gap of \(x^k\) and \(t^k\).

Lemma 8. Let Assumption \cite{2} hold. Consider the iterates \(\{g_i^k\}\) generated by GT–SAGA. Then we have the following inequality holds: \(\forall k \geq 1,\)

\[
\mathbb{E} \left[\|g_i^k - \nabla f(x^k)\|^2\right] \leq 4L^2 \mathbb{E} \left[\|x^k - W_\infty x^k\|^2\right] + 4L^2 \mathbb{E} \left[n \|x^k_i - x^*\|^2\right] + 2L^2 \mathbb{E} \left[t^k\right].
\]

Proof. Recall the local gradient estimator \(g_i^k\) from Algorithm \cite{1} and proceed as follows.

\[
\mathbb{E} \left[\|g_i^k - \nabla f_i(x^k)\|^2 | F_k\right] = \mathbb{E} \left[\left\|\nabla f_i(s^k_i(x^k_i)) - \nabla f_i(s^k_i(z^k_{i,j}))\right\|^2 | F_k\right] \leq \mathbb{E} \left[\left\|\nabla f_i(s^k_i(x^k_i)) - \nabla f_i(z^k_{i,j})\right\|^2 | F_k\right] + \mathbb{E} \left[\left\|\nabla f_i(z^k_{i,j}) - \nabla f_i(x^k_i)\right\|^2 | F_k\right] = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{E} \left[\left\|\nabla f_i(s^k_i(x^k_i)) - \nabla f_i(z^k_{i,j})\right\|^2 | F_k\right] + \mathbb{E} \left[\left\|\nabla f_i(z^k_{i,j}) - \nabla f_i(x^k_i)\right\|^2 | F_k\right]
\]

\[
\leq 2L^2 \|x^k_i - x^*\|^2 + 2L^2 t_i^k + 4L^2 \|x^k - x^*\|^2 + 2L^2 t^k,
\]

(21)

where the second inequality uses the standard conditional variance decomposition

\[
\mathbb{E} \left[\|a_i^k - \mathbb{E}[a_i^k | F_k]\|^2 | F_k\right] = \mathbb{E} \left[\|a_i^k\|^2 | F_k\right] - \mathbb{E} \left[\|a_i^k | F_k\|^2\right] \leq \mathbb{E} \left[\|a_i^k\|^2 | F_k\right],
\]

(22)

with \(a_i^k = \nabla f_i(s^k_i(x^k_i)) - \nabla f_i(s^k_i(z^k_{i,j}))\). The proof follows by summing (21) over \(i\) and taking the total expectation.

Lemma 8 clearly shows that as \(x^k\) and \(z^k_{i,j}\) approach to an agreement on \(x^*\), the variance of the gradient estimator decays to zero. We have the following corollary.

Corollary 1. Let Assumption \cite{2} and \cite{3} hold. Consider the iterates \(\{g_i^k\}\) generated by GT–SAGA. If \(0 < \alpha \leq \frac{1}{4\sqrt{2L}}\), then the following inequality holds \(\forall k \geq 0,\)

\[
\mathbb{E} \left[\|g_i^{k+1} - \nabla f(x^{k+1})\|^2\right] \leq 12.75L^2 \mathbb{E} \left[\|x^k - W_\infty x^k\|^2\right] + 2L^2 \mathbb{E} \left[n \|x^k - x^*\|^2\right] + 8L^2 \alpha^2 \mathbb{E} \left[\|y^k - W_\infty y^k\|^2\right] + 2.25L^2 \mathbb{E} \left[t^k\right].
\]

Proof. Following directly from Lemma 8 we have: \(\forall k \geq 0,\)

\[
\mathbb{E} \left[\|g_i^{k+1} - \nabla f(x^{k+1})\|^2\right] \leq 4L^2 \mathbb{E} \left[\|x^{k+1} - W_\infty x^{k+1}\|^2\right] + 4L^2 \mathbb{E} \left[n \|x^{k+1} - x^*\|^2\right] + 2L^2 \mathbb{E} \left[t^k\right].
\]
Using \([4, 7]\) and Lemma \([7]\) in the previous sections, we now derive a more refined bound on the gradient variance for \(\text{GT-SAGA}\). We have the following:

\[\begin{align*}
\mathbb{E} \left[ \|g^{k+1} - \nabla f(x^{k+1})\|^2 \right] &\leq 4L^2 \left( 2\mathbb{E} \left[ \|x^k - W_\infty x^k\|^2 \right] + 2\alpha \mathbb{E} \left[ \|y^k - W_\infty y^k\|^2 \right] 
+ 4L^2 \alpha E \left[ \|x^k - W_\infty x^k\|^2 \right] + 2\mathbb{E} \left[ n \|\bar{x}^k - x^k\|^2 \right] 
+ \frac{\alpha^2}{n} E \left[ \|g^k - \nabla f(x^k)\|^2 \right] \right) 
+ 2L^2 \left( \frac{1}{m} \right) E \left[ \|k\right] 
+ \frac{2}{m} E \left[ n \|\bar{x}^k - x^k\|^2 \right] \right) 
\leq 12.25L^2 \mathbb{E} \left[ \|x^k - W_\infty x^k\|^2 \right] + 12L^2 \mathbb{E} \left[ n \|\bar{x}^k - x^k\|^2 \right] 
+ 8L^2 \alpha E \left[ \|y^k - W_\infty y^k\|^2 \right] 
+ 2L^2 \mathbb{E} \left[ \|k\right] + 0.125 \mathbb{E} \left[ \|g^k - \nabla f(x^k)\|^2 \right].
\end{align*}\]

The proof follows by applying the bound on \(\mathbb{E}[\|g^k - \nabla f(x^k)\|^2]\) in Lemma \([6]\) to the above inequality.

\(\square\)

**B. Main results for \(\text{GT-SAGA}\)**

With the bounds on the gradient variance for \(\text{GT-SAGA}\) derived in the previous section, we now derive a more refined bound on the gradient variance for \(\text{GT-SAGA}\). We have the following:

\[\begin{align*}
\mathbb{E} \left[ n \|\bar{x}^{k+1} - x^k\|^2 \right] &\leq L^2 \alpha \left( \frac{1}{\mu} + \frac{4\alpha}{n} \right) E \left[ \|x^k - W_\infty x^k\|^2 \right] 
+ \left( 1 - \mu \alpha + \frac{4L^2 \alpha^2}{n} \right) E \left[ n \|x^k - x^k\|^2 \right] 
+ \frac{2L^2 \alpha^2}{n} E \left[ \|k\right] 
\end{align*}\]

If \(0 < \alpha \leq \frac{1}{4\mu}\), then \(\frac{1}{\mu} + \frac{4\alpha}{n} \leq \frac{2}{\mu}\); if \(0 < \alpha \leq \frac{\mu n}{8L^2}\), then we have \(1 - \mu \alpha + \frac{4L^2 \alpha^2}{n} \leq 1 - \frac{\mu \alpha}{2}\). Therefore, if \(0 < \alpha \leq \frac{\mu n}{8L^2}\), we have the following:

\[\begin{align*}
\mathbb{E} \left[ n \|\bar{x}^{k+1} - x^k\|^2 \right] &\leq \frac{2L^2 \alpha}{\mu} E \left[ \|x^k - W_\infty x^k\|^2 \right] + \left( 1 - \frac{\mu \alpha}{2} \right) E \left[ n \|x^k - x^k\|^2 \right] 
+ \frac{2L^2 \alpha^2}{n} E \left[ \|k\right] 
\end{align*}\]

Second, we apply the upper bounds on \(\mathbb{E}[\|g^k - \nabla f(x^k)\|^2]\) and \(\mathbb{E}[\|g^{k+1} - \nabla f(x^{k+1})\|^2]\) in Lemma \([6]\) and Corollary \([1]\) to obtain the following:

\[\begin{align*}
\mathbb{E} \left[ \|y^{k+1} - W_\infty y^{k+1}\|^2 \right] &\leq \frac{104L^2}{1 - \sigma^2} E \left[ \|x^k - W_\infty x^k\|^2 \right] + \frac{71L^2}{1 - \sigma^2} E \left[ n \|x^k - x^k\|^2 \right] 
+ \frac{19L^2}{1 - \sigma^2} E \left[ \|k\right] 
+ \frac{(1 + \sigma^2)}{2} \frac{64L^2 \alpha^2}{1 - \sigma^2} E \left[ \|x^k - W_\infty x^k\|^2 \right].
\end{align*}\]

Therefore, if \(0 < \alpha \leq \frac{1 - \sigma^2}{16L}\), we have that:

\[\begin{align*}
\mathbb{E} \left[ \|y^{k+1} - W_\infty y^{k+1}\|^2 \right] &\leq \frac{104L^2}{1 - \sigma^2} E \left[ \|x^k - W_\infty x^k\|^2 \right] + \frac{71L^2}{1 - \sigma^2} E \left[ n \|x^k - x^k\|^2 \right] 
+ \frac{19L^2}{1 - \sigma^2} E \left[ \|k\right] + \frac{3 + \sigma^2}{4} E \left[ \|y^k - W_\infty y^k\|^2 \right].
\end{align*}\]

To proceed, we write \([6, 23]\), Lemma \([7]\) and \([23]\) jointly as a linear matrix inequality in the following proposition.

**Proposition 1.** Let Assumptions \([6, 2, 3]\) hold and consider the iterates \(\{x^k\}, \{y^k\}, \{k\}\) generated by \(\text{GT-SAGA}\). If the step-size \(\alpha\) follows \(0 < \alpha \leq \frac{\mu(1 - \sigma^2)}{16L}\), we have:

\[\begin{align*}
\|u^{k+1}\| &\leq G_{\alpha} \|u^k\|,
\end{align*}\]

where \(u^k \in \mathbb{R}^d\) and \(G_{\alpha} \in \mathbb{R}^{d \times d}\) are defined as follows:

\[\begin{align*}
G_{\alpha} &= \begin{bmatrix}
\mathbb{E} \left[ \|x^k - W_\infty x^k\|^2 \right] 
E \left[ n \|\bar{x}^k - x^k\|^2 \right] 
E \left[ \|y^k - W_\infty y^k\|^2 \right]

\mathbb{E} \left[ \|k\right]
\end{bmatrix}
\begin{bmatrix}
1 + \frac{\alpha^2}{2}
0
0
\frac{2L^2 \alpha}{1 - \sigma^2}
\frac{1 - \mu \alpha}{2}
\frac{2L^2 \alpha^2}{n}
\frac{m}{2}
\frac{m}{2}
\frac{1}{M}
\frac{3 + \sigma^2}{4}
\end{bmatrix}
\end{align*}\]

Clearly, to show the linear convergence of \(\text{GT-SAGA}\), it suffices to derive the range of \(\alpha\) such that \(\rho(G_{\alpha}) < 1\). To do this, we present a useful lemma from [56].

**Lemma 9.** Let \(A \in \mathbb{R}^{d \times d}\) be non-negative and \(x \in \mathbb{R}^d\) be positive. If \(Ax \leq \beta x\) for \(\beta > 0\), then \(\rho(A) \leq \|x\|_\infty \leq \beta\).

We are ready to prove Theorem \([1]\) based on Proposition \([1]\) and Corollary \([1]\).

**Proof of Theorem \([1]\)** Recall from Proposition \([1]\) that if \(0 < \alpha \leq \frac{\mu(1 - \sigma^2)}{16L}\), we have \(u^{k+1} \leq G_{\alpha} u^k\). In the light of Lemma \([9]\) we solve for the range of the step-size \(\alpha\) and a positive vector \(\epsilon = [\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4]^T\) such that the following (entry-wise) linear matrix inequality holds:

\[G_{\alpha} \epsilon \leq \left( 1 - \frac{\mu \alpha}{4} \right) \epsilon,\]

which can be written equivalently in the following form:

\[\begin{align*}
\frac{\mu \alpha}{4} + \frac{2L^2 \epsilon_4}{4 - \sigma^2} \epsilon_1 &\leq \frac{1 - \sigma^2}{2}
\frac{2L^2 \epsilon_3}{n} \epsilon_\alpha &\leq \frac{2L^2 \epsilon_3}{\mu}
\frac{\mu \alpha}{4} &\leq 1 - \frac{2 \epsilon_1}{m \epsilon_3} - \frac{2 \epsilon_2}{m \epsilon_3}
\frac{\mu \alpha}{4} &\leq 1 - \frac{\epsilon_1}{M} - \frac{2 \epsilon_1}{m \epsilon_3} - \frac{2 \epsilon_2}{M \epsilon_3}
\end{align*}\]
Clearly, that $29$–$31$ hold for some feasible range of $\alpha$ is equivalent to the RHS of $29$–$31$ being positive. Based on this observation, we will next fix the values of $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ that are independent of $\alpha$. First, for the RHS of $29$ to be positive, we set $\epsilon_1 = 1, \epsilon_2 = 8.5Q^2, \epsilon_3 = 17MQ^2$. Second, the RHS of $30$ being positive is equivalent to

$$\epsilon_3 > \frac{2M}{m} + \frac{2M}{m} - \frac{2M}{m} + 17MQ^2.$$  \hfill (32)

We therefore set $\epsilon_3 = \frac{20MQ^2}{m}$. Third, we note that the RHS of $31$ being positive is equivalent to the following:

$$\epsilon_4 > \frac{4}{(1-\sigma^2)^2} \left( \frac{104\epsilon_1 + 71\epsilon_2 + 19\epsilon_3}{m} \right) = \frac{4}{(1-\sigma^2)^2} \left( 104 + 603.5Q^2 + \frac{380MQ^2}{m} \right)$$

Note that $104 + 603.5Q^2 + \frac{380MQ^2}{m} \leq 1087Q^2$. We therefore set $\epsilon_4 = \frac{8700}{1-\sigma^2} \frac{MQ}{m}$.

We now solve for the range of $\alpha$ from $28$–$31$ given the previously fixed $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$. From $29$, we have that

$$\alpha \leq \frac{n}{2L^2\epsilon_3} \left( \frac{\mu}{4} \epsilon_2 - \frac{2L^2}{\mu} \epsilon_1 \right) = \frac{n}{m} \frac{1}{320QL}.$$  \hfill (33)

Moreover, it is straightforward to verify that if $\alpha$ satisfies

$$0 < \alpha \leq \frac{m}{M} \frac{(1-\sigma^2)^2}{320QL}$$  \hfill (34)

then $28$ holds. Next, to make $30$ hold, it suffices to make $\alpha$:

$$\alpha \leq \frac{1}{5\mu M}.$$  \hfill (35)

Finally, to make $31$ hold, it suffices to make

$$\alpha \leq \frac{1-\sigma^2}{2\mu}.$$  \hfill (36)

To summarize, combining $34$–$36$, we conclude that if the step-size $\alpha$ satisfies

$$0 < \alpha \leq \overline{\alpha} := \min \left\{ \frac{1}{5\mu M}, \frac{m}{320M} \frac{(1-\sigma^2)^2}{LQ} \right\},$$  \hfill (37)

then $27$ holds with some $\epsilon > 0$ and thus $\rho(G_\alpha) \leq 1 - \frac{\omega}{4}$ according to Lemma $9$. Further if $\alpha = \overline{\alpha}$, we have

$$\rho(G_\alpha) \leq 1 - \min \left\{ \frac{1}{20M}, \frac{1280M}{1280M} \frac{(1-\sigma^2)^2}{Q^2} \right\},$$

which completes the proof.

**VI. CONVERGENCE ANALYSIS OF GT–SVRG**

In this section, we conduct the complexity analysis of GT–SVRG in Algorithm 2 based on the auxiliary results derived for the general dynamical system $2$ in Section IV. Recall from Algorithm 2 that the gradient estimator $v^k_i$ at each node $i$ in GT–SVRG is given by the following: $\forall k \geq 1$, choose $s_i^k$ uniformly at random in $\{1, \cdots, m_i\}$ and

$$\phi_i = \nabla f_{i,s_i^k}(x^k_i) - \nabla f_{i,s_i^k}(\tau^k_i) + \nabla f_i(\tau^k_i)$$  \hfill (38)

where $\tau^k_i = x^k_i$ if $\text{mod}(k, T) = 0$, where $T$ is the length of each inner-loop iterations of GT–SVRG, otherwise $\tau^k_i = \tau^{k-1}_i$. To facilitate the convergence analysis, we define an auxiliary variable $\tau^k := \frac{1}{n} \sum_{i=1}^n \tau^k_i, \forall k \geq 0$.

**A. Bounding the variance of the gradient estimator**

We first bound the variance of the gradient estimator $\phi^k_i$, following a similar procedure as the proof of Lemma 8.

**Lemma 10.** Let Assumption 2 hold and consider the iterates $\{v^k_i\}$ generated by GT–SVRG in Algorithm 2. The following inequality holds $\forall k \geq 0$:

$$\mathbb{E} \left[ \|v^k_i - \nabla f(x^k_i)\|^2 \right] \leq 4L^2\mathbb{E} \left[ \|x^k - W_\infty x^k\|^2 + 4L^2 \mathbb{E} \left[ n \|x^k - x^*\|^2 \right] + 4L^2 \mathbb{E} \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + 4L^2 \mathbb{E} \left[ n \|\tau^k - x^*\|^2 \right] \right].$$

**Proof.** From Lemma 10 we have: $\forall k \geq 0$,

$$\mathbb{E} \left[ \|v^{k+1}_i - \nabla f(x^{k+1}_i)\|^2 \right] \leq 4L^2\mathbb{E} \left[ \|x^{k+1} - W_\infty x^{k+1}\|^2 \right] + 4L^2 \mathbb{E} \left[ n \|x^{k+1} - x^*\|^2 \right] + 4L^2 \mathbb{E} \left[ \|\tau^{k+1} - W_\infty \tau^{k+1}\|^2 \right] + 4L^2 \mathbb{E} \left[ n \|\tau^{k+1} - x^*\|^2 \right].$$  \hfill (40)
Recall that $\tau^{k+1} = x^{k+1}$ if mod$(k + 1, T) = 0$; otherwise, $\tau^{k+1} = \tau^k$. We first derive upper bounds on the last two terms in (40) for these two cases separately. On the one hand, if $\text{mod}(k + 1, T) \neq 0$, we have that

$$4L^2E \left[ \|\tau^{k+1} - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ \|\tau^k - x^*\|^2 \right] = 4L^2E \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ n \|\tau^k - x^*\|^2 \right]. \quad (41)$$

On the other hand, if $\text{mod}(k + 1, T) = 0$, we have that

$$4L^2E \left[ \|\tau^{k+1} - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ \|\tau^k - x^*\|^2 \right] = 4L^2E \left[ \|x^{k+1} - W_\infty x^k\|^2 \right] + 4L^2E \left[ n \|x^k - x^*\|^2 \right]. \quad (42)$$

Therefore, combining (41) and (42), we have that $\forall k \geq 0$:

$$4L^2E \left[ \|\tau^{k+1} - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ \|\tau^k - x^*\|^2 \right] \leq 4L^2E \left[ \|x^{k+1} - W_\infty x^k\|^2 \right] + 4L^2E \left[ n \|x^k - x^*\|^2 \right] + 4L^2E \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ n \|\tau^k - x^*\|^2 \right]. \quad (43)$$

Next, we apply (43) in (40) to obtain

$$E \left[ \|y^{k+1} - \nabla f(x^{k+1})\|^2 \right] \leq 8L^2E \left[ \|x^{k+1} - W_\infty x^k\|^2 \right] + 8L^2E \left[ n \|x^k - x^*\|^2 \right] + 4L^2E \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ n \|\tau^k - x^*\|^2 \right]. \quad (44)$$

We use (4), (7) in (44) to proceed. If $0 < \alpha \leq \frac{1}{8L}$,

$$E \left[ \|\nu^{k+1} - \nabla f(x^{k+1})\|^2 \right] \leq 8L^2 \left( 2E \left[ \|x^k - W_\infty x^k\|^2 \right] + 2\alpha^2E \left[ \|y^k - W_\infty y^k\|^2 \right] \right) + 8\frac{L^2\alpha^2}{n}E \left[ \|x^k - W_\infty x^k\|^2 \right] + 2E \left[ n \|x^k - x^*\|^2 \right] + \frac{\alpha^2}{n}E \left[ \|\nu^k - \nabla f(x^k)\|^2 \right] \right) + 4L^2E \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ n \|\tau^k - x^*\|^2 \right], \leq 16.25L^2E \left[ \|x^k - W_\infty x^k\|^2 \right] + 16L^2E \left[ \|y^k - W_\infty y^k\|^2 \right] + 16L^2E \left[ n \|x^k - x^*\|^2 \right] + 4L^2E \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + 4L^2E \left[ n \|\tau^k - x^*\|^2 \right] + 0.125E \left[ \|\nu^k - \nabla f(x^k)\|^2 \right], \quad (45)$$

The proof follows by using the bound on $E \left[ \|y^k - \nabla f(x^k)\|^2 \right]$ from Lemma 10 in the above inequality.

\[\square\]

**B. Main results for GT–SVRG**

We now use the upper bounds on the variance of the gradient estimator $v^k$ in GT–SVRG obtained in the previous subsection to refine the inequalities derived for the general dynamical system \(2\) in Section IV and establish the explicit complexity for GT–SVRG. We first apply the upper bound on $E\left[\|v^k - \nabla f(x^k)\|^2\right]$ in Lemma 10 to (7) to obtain $\forall k \geq 0$:

$$E \left[ n \|x^{k+1} - x^*\|^2 \right] \leq L^2\alpha \left( \frac{1}{\mu} + \frac{4\alpha}{n} \right) E \left[ \|x^k - W_\infty x^k\|^2 \right] + \left( 1 - \mu\alpha + \frac{4L^2\alpha^2}{n} \right) E \left[ n \|\tau^k - x^*\|^2 \right] + \frac{4L^2\alpha^2}{n} E \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + \frac{4L^2\alpha^2}{n} E \left[ n \|\tau^k - x^*\|^2 \right]. \quad (46)$$

If $0 < \alpha \leq \frac{1}{8L}$, we have $(\frac{1}{\mu} + \frac{4\alpha}{n}) \leq \frac{3}{2}$; if $0 < \alpha \leq \frac{\mu}{8L}$, we have $1 - \mu\alpha + \frac{4L^2\alpha^2}{n} \leq 1 - \frac{3\alpha}{2}$. Therefore, if $0 < \alpha \leq \frac{\mu}{8L}$, we have:

$$E \left[ n \|x^{k+1} - x^*\|^2 \right] \leq 2L^2\alpha E \left[ \|x^k - W_\infty x^k\|^2 \right] + \left( 1 - \frac{\mu\alpha}{2} \right) E \left[ n \|\tau^k - x^*\|^2 \right] + \frac{4L^2\alpha^2}{n} E \left[ \|\tau^k - W_\infty \tau^k\|^2 \right] + \frac{4L^2\alpha^2}{n} E \left[ n \|\tau^k - x^*\|^2 \right]. \quad (47)$$

Now, we write Lemma 3, (46) and (47) jointly in an entrywise linear matrix inequality that characterizes the evolution of GT–SVRG in the following proposition.

**Proposition 2.** Let Assumptions 1, 2 and 3 hold and Consider the iterates $\{x^k\}, \{y^k\}, \{v^k\}$ generated by GT–SVRG. If the step-size $\alpha$ follows $0 < \alpha \leq \frac{\mu}{144L^2}$, then the following linear matrix inequality hold $\forall k \geq 0$:

$$u^{k+1} \leq J_\alpha u^k + H_\alpha u^k, \quad (48)$$
where \( u^k, \tilde{u}^k \in \mathbb{R}^3 \) and \( J_\alpha, H_\alpha \in \mathbb{R}^{3 \times 3} \) are defined as follows:

\[
\begin{align*}
J_\alpha &= \begin{bmatrix}
\frac{1 + \sigma^2}{2} & 0 & 2\alpha^2 L_z^2 \rho \\
\frac{2 L_z^2 \alpha}{\mu} & 1 - \frac{\mu \alpha}{2} & 0 \\
1 - \sigma^2 & 3 + \sigma^2 & 4
\end{bmatrix}, \\
H_\alpha &= \begin{bmatrix}
0 & 0 & 0 \\
\frac{4L_z^2 \alpha}{n} & \frac{4L_z^2 \alpha^2}{n} & 0 \\
\frac{n}{2} & \frac{n}{2} & 0
\end{bmatrix}.
\end{align*}
\]

Recall that \( T \) is the number of the inner-loop iterations of GT-SVRG. We will show that the subsequence \( \{u^{tT}\}_{t \geq 0} \) of \( \{u^k\}_{k \geq 0} \), which corresponds to the outer-loop updates of GT-SVRG, converges to zero linearly, based on which the total complexity of GT-SVRG will be established, in terms of the number of parallel local component gradient computations required to find the solution \( x^* \). Recall from Algorithm 2 that \( \forall k \geq 0, \tau^{k+1} = x^k + \text{mod}(k + 1, T) = 0; \) else \( \tau^{k+1} = \tau^k \). Therefore, \( \forall t \geq 0 \) and \( tT \leq k \leq (t + 1)T - 1 \), we have \( \tau^k = x^{tT} \). Based on this discussion, (48) can be rewritten as the following dynamical system with delays:

\[
u^{k+1} \leq J_\alpha u^k + H_\alpha u^{tT}, \quad \forall k \in \{tT, (t + 1)T - 1\}, \quad \forall t \geq 0.
\]

We then recursively apply the above inequality over \( k \) to obtain the evolution of the outer-loop iterates \( \{u^{tT}\}_{t \geq 0} \):

\[
u^{(t+1)T} \leq (J_\alpha^T + \sum_{t=0}^{T-1} J_\alpha H_\alpha) u^{tT}, \quad \forall t \geq 0.
\]

(49)

Clearly, to show the linear decay of \( \{u^{tT}\}_{t \geq 0} \), it suffices to find the range of \( \alpha \) such that \( \rho(J_\alpha^T + \sum_{t=0}^{T-1} J_\alpha H_\alpha) < 1 \). To this aim, we first derive the range of \( \alpha \) such that \( \rho(J_\alpha) < 1 \).

**Lemma 11.** Let Assumptions 7, 2, 3 hold and consider the system matrix \( J_\alpha, H_\alpha \) defined in Proposition 2. If the step-size \( \alpha \) follows \( 0 < \alpha \leq \frac{(1-\sigma^2)^2}{187Q^2} \), then

\[
\rho(J_\alpha) \leq \|J_\alpha\|_\infty \leq 1 - \frac{\mu \alpha}{4},
\]

(50)

where \( \delta = \left[1, 8Q^2, \frac{6528Q^2}{(1-\sigma^2)^2}\right]^T \).

**Proof.** In the light of Lemma 9 we solve for the range of \( \alpha \) and a positive vector \( \delta = [\delta_1, \delta_2, \delta_3] \) such that the following entry-wise linear inequality holds:

\[
J_\alpha \delta \leq \left(1 - \frac{\mu \alpha}{4}\right) \delta,
\]

which can be written equivalently as

\[
\begin{align*}
\mu \alpha &\leq \frac{1}{4} - \frac{1}{\sigma^2} \delta_3 \\
8Q^2 \delta_1 &\leq \delta_3,
\end{align*}
\]

(51)

(52)

Based on (52), we set \( \delta_1 = 1 \) and \( \delta_2 = 6Q^2 \). With \( \delta_1 \) and \( \delta_2 \) being fixed, we next choose \( \delta_3 > 0 \) such that the RHS of (53) is positive, i.e.,

\[
1 - \frac{\sigma^2}{4 \delta_3} \left(\delta_3 - \frac{480 + 2784Q^2}{(1 - \sigma^2)^2}\right) > 0.
\]

It suffices to set \( \delta_3 = \frac{4528Q^2}{(1-\sigma^2)^2} \). Now, with the previously fixed values of \( \delta_1, \delta_2, \delta_3 \), in order to make (53) hold, it suffices to choose \( \alpha \) such that \( 0 < \alpha \leq \frac{1-\sigma^2}{2\mu} \). Similarly, it can be verified that in order to make (51) hold, it suffices to make \( \alpha \) satisfy \( 0 < \alpha \leq \frac{(1-\sigma^2)^2}{187Q^2} \), which completes the proof.

We note that if the step-size \( \alpha \) satisfies the condition in Lemma 11, we have \( \rho(J_\alpha) < 1 \). Moreover, since \( J_\alpha \) is non-negative, we have that \( \sum_{t=0}^{T-1} J_\alpha^t \leq \sum_{t=0}^{\infty} J_\alpha^t = (I_3 - J_\alpha)^{-1} \). Therefore, following (49), we have:

\[
u^{(t+1)T} \leq (J_\alpha^T + (I_3 - J_\alpha)^{-1} H_\alpha) u^{tT}, \quad \forall t \geq 0.
\]

(54)

The rest of the convergence analysis is to derive the condition on the the number of each inner iterations \( T \) and the step-size \( \alpha \) of GT-SVRG such that the following inequality holds:

\[
\rho(J_\alpha^T + (I_3 - J_\alpha)^{-1} H_\alpha) < 1.
\]

We first show that \( (I_3 - J_\alpha)^{-1} H_\alpha \) is sufficiently small under an appropriate weighted matrix norm in the light of Lemma 9.

**Lemma 12.** Let Assumptions 7, 2, 3 hold. Consider the system matrices \( J_\alpha, H_\alpha \) defined in Proposition 2. If the step-size \( \alpha \) follows \( 0 < \alpha \leq \frac{(1-\sigma^2)^2}{187Q^2} \), then

\[
\| (I_3 - J_\alpha)^{-1} H_\alpha \|_q \leq 0.66,
\]

where \( q = \left[1, 1, \frac{1453}{(1-\sigma^2)^2}\right]^T \).

**Proof.** We start by deriving an entry-wise upper bound for the matrix \( (I_3 - J_\alpha)^{-1} \). Note that

\[
I - J_\alpha = \begin{bmatrix}
1 - \sigma^2 & 0 & -2\alpha^2 L_z^2 \\
-2L_z^2 \alpha & \frac{\mu \alpha}{2} & 0 \\
\frac{120}{\mu} & \frac{87}{\mu} & \frac{1 - \sigma^2}{4}
\end{bmatrix},
\]

(55)

whose determinant is given by

\[
\det(I - J_\alpha) = \frac{(1-\sigma^2)^2 \mu \alpha - 348L_z^6 \alpha^3}{16} - \frac{120 \alpha^3 \mu L_z^2}{(1 - \sigma^2)^2}.
\]

It can be verified that if \( 0 < \alpha \leq \frac{(1-\sigma^2)^2}{187Q^2} \),

\[
\det(I - J_\alpha) \geq \frac{(1-\sigma^2)^2 \mu \alpha}{32}.
\]

(56)

Then we derive an entry-wise upper bound for \( \text{adj}(I_3 - J_\alpha) \), where \( \text{adj}(\cdot) \) denotes the adjugate of the argument matrix and we denote \( \text{adj}(\cdot)_{ij} \) as its \( i,j \)th entry:

\[
\text{adj}(I - J_\alpha)_{1,1} = \frac{174L_z^6 \alpha^2}{(1 - \sigma^2)^2}, \quad \text{adj}(I - J_\alpha)_{1,3} = \frac{\mu L_z^6 \alpha^3}{1 - \sigma^2},
\]

(57)

\(
\text{adj}(I - J_\alpha)_{2,2} \leq \frac{(1-\sigma^2)^2}{8}, \quad \text{adj}(I - J_\alpha)_{2,3} = \frac{4L_z^6 \alpha^3}{(1 - \sigma^2)^2},
\]

(58)

\[
\text{adj}(I - J_\alpha)_{3,3} = \frac{87}{4} \alpha, \quad \text{adj}(I - J_\alpha)_{3,3} = \frac{\mu (1 - \sigma^2)}{4}.
\]
With the help of the above calculations, an entry-wise upper bound for \((I_3 - J_3)^{-1} H_\alpha\) can be obtained, i.e., if \(0 < \alpha \leq \frac{(1 - \sigma^2)^2}{187QL}\), we have

\[
(I_3 - J_3)^{-1} H_\alpha \leq \begin{bmatrix}
0.039 & 0.039 & 0 \\
0.23 & 0.23 & 0 \\
334 & 334 & 0 \\
\end{bmatrix}. 
\]

Using Lemma 9 in a similar way as the proof of Lemma 11, it can be verified that \((I_3 - J_3)^{-1} H_\alpha\) \(Q\) \(\leq 0.66Q\), where \(Q = [1, 1, 1453/(1 - \sigma^2)^2]^T\), which completes the proof.

Note that we use two different weighted matrix norms to bound \(J_0\) and \((I - J_0)^{-1} H_\alpha\) respectively in Lemma 11 and 12, i.e., \(\|\cdot\|_\infty^q\) and \(\|\cdot\|_q^\infty\), where \(\delta = [1, SQ^2/(1 - \sigma^2)^2]_i\), and \(Q = [1, 1, 1453/(1 - \sigma^2)^2]^T\). It can be verified that \(\|\cdot\|_\infty^q \leq 8Q^2\|\cdot\|_q^\infty\), \(\forall X \in \mathbb{R}^{3 \times 3}\).

We next show the linear convergence of the outer-loop of \(\text{GT-SVRG}\), i.e., the linear decay of the subsequence \(\{u_t^T\}_{t \geq 0}\) of \(\{u_k^T\}_{k \geq 0}\), where \(T\) is the number of inner-loop iterations.

**Lemma 13.** Let Assumption 2 and 3 hold. Consider the iterates \(\{u_k^T\}\) generated by \(\text{GT-SVRG}\) (defined in Proposition 2). If the step-size \(\alpha = \frac{(1 - \sigma^2)^2}{187QL}\) and the number of inner-loop iterations \(T = \frac{1406Q^2}{(1 - \sigma^2)^2} \log(200Q)\), then the following holds:

\[
\|u_{t+1}^T\|_q^\infty \leq \frac{1}{748Q^2} \|u_t^T\|_q^\infty, \quad t \geq 0. 
\]

**Proof.** Recall the recursion in (54): \(u_k^T = \left(J_0^T + (I - J_0)^{-1} H_\alpha\right) u_t^T\).

Note that the weighted vector norm \(\|\cdot\|_q^\infty\) induces the weighted matrix norm \(\|\cdot\|_q^\infty^q\). Then using Lemma 11 and 12, and (57), if \(\alpha = \frac{(1 - \sigma^2)^2}{187QL}\), then \(T \geq 0:\)

\[
\|u_{t+1}^T\|_\infty^q \leq \left\|J_0^T + (I - J_0)^{-1} H_\alpha\right\|_q^\infty \|u_t^T\|_\infty^q \\
\leq \left\|J_0^T\right\|_\infty^q + 0.66 \|u_t^T\|_\infty^q, \\
\leq 8Q^2 \left(\|J_0\|_\infty^q \|J_0\|_q^{\infty} + 0.66\right) \|u_t^T\|_\infty^q, \\
\leq 8Q^2 \exp \left(-\frac{(1-\sigma^2)^2}{748Q^2} T\right) + 0.66 \|u_t^T\|_\infty^q, \\
\leq \frac{x}{\|\cdot\|_\infty^q\} \exp\{x\}, \quad \forall x \in \mathbb{R}. 
\]

The proof follows by setting \(T = \frac{1406Q^2}{(1 - \sigma^2)^2} \log(200Q^2)\) in the last inequality above.

**VII. Conclusion**

In this paper, we have proposed a novel framework for constructing variance-reduced decentralized stochastic first-order methods over undirected and weight-balanced directed graphs that hinge on gradient tracking techniques. In particular, we have derived decentralized versions of SAGA and SVRG algorithms, \(\text{GT-SAGA}\) and \(\text{GT-SVRG}\), that achieve accelerated linear convergence for smooth and strongly-convex functions. We have further shown that in the “big-data” regime \(\text{GT-SAGA}\) and \(\text{GT-SVRG}\) achieve non-asymptotic linear speedups in terms of the number of nodes compared with centralized SAGA and SVRG and exhibit superior performance over existing decentralized schemes.

**References**

[1] R. Xin, U. A. Khan, and S. Kar, “Variance-reduced decentralized stochastic optimization with gradient tracking–Part I: GT-SAGA,” arXiv:1909.11774, 2019.

[2] R. Xin, U. A. Khan, and S. Kar, “Variance-reduced decentralized stochastic optimization with gradient tracking–Part II: GT-SVRG,” arXiv:1910.04057, 2019.

[3] J. Tisitsiklis, D. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 1986.

[4] A. Nedic and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Trans. on Autom. Control*, vol. 54, no. 1, pp. 48, 2009.

[5] J. Chen and A. H. Sayed, “Diffusion adaptation strategies for distributed optimization and learning over networks,” *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, 2012.

[6] S. Kar and J. M. F. Moura, “Consensus + innovations distributed inference over networks: cooperation and sensing in networked systems,” *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 99–109, 2013.

[7] J. Duchi, A. Agarwal, and M. J. Wainwright, “Dual averaging for distributed optimization: Convergence analysis and network scaling,” *IEEE Trans. on Autom. control*, vol. 57, no. 3, pp. 592–606, 2011.

[8] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, “Push-sum distributed dual averaging for convex optimization,” in *Ann. Conf. Decis. Control*. IEEE, 2012, pp. 5453–5458.

[9] E. Wei and A. Ozdaglar, “On the O(1/k) convergence of asynchronous distributed alternating direction method of multipliers,” in *2013 IEEE Global Conf. Signal Inf. Process*. IEEE, 2013, pp. 551–554.

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

[11] M. Maros and J. Jaulden, “On the Q-linear convergence of distributed generalized adam under non-strongly convex function components,” *IEEE Signal Inf. Process. Netw.*, 2019.

[12] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, “DQM: decentralized quadratically approximated alternating direction method of multipliers,” in *IEEE Global Conf. Signal Inf. Process*. IEEE, 2013, pp. 551–554.

[13] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, “Exact diffusion for distributed optimization and learning Part I: Algorithm development,” *IEEE Trans. on Signal Process.*, vol. 63, no. 22, pp. 6013–6023, 2015.

[14] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, “Exact diffusion for distributed optimization and learning Part II: Convergence analysis,” *IEEE Trans. on Signal Process.*, vol. 67, no. 3, pp. 708–723, 2018.

[15] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, “Exact diffusion for distributed optimization and learning Part II: Convergence analysis,” *IEEE Trans. on Signal Process.*, vol. 67, no. 3, pp. 724–739, 2018.

[16] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, “DL-M: decentralized linearized alternating direction method of multipliers,” *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 4051–4064, 2015.

[17] P. Di Lorenzo and G. Scutari, “Distributed nonconvex optimization over networks,” in *6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*. IEEE, 2015, pp. 229–232.
[19] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, “Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepizes,” in *Ann. Conf. Decis. Control*. IEEE, 2015, pp. 2055–2060.

[20] P. Di Lorenzo and G. Scutari, “NEXT: In-network nonconvex optimization,” *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 2, pp. 120–136, 2016.

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

[22] A. Nedic, 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.

[23] Y. Sun, A. Danezis, and G. Scutari, “Convergence rate of distributed optimization algorithms based on gradient tracking,” *arXiv preprint arXiv:1905.02637*, 2019.

[24] G. Scutari and Y. Sun, “Distributed nonconvex constrained optimization over time-varying digraphs,” *Mathematical Programming*, vol. 176, no. 1-2, pp. 497–544, 2019.

[25] C. Xi, R. Xin, and U. A. Khan, “ADD-OPT: accelerated distributed directed optimization,” *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1329–1339, 2017.

[26] G. Qu and N. Li, “Accelerated distributed nesterov gradient descent,” *IEEE Transactions on Automatic Control*, 2019.

[27] D. Jakovetić, “A unification and generalization of exact distributed first-order methods,” *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 1, pp. 31–46, 2018.

[28] S. A. Alghunaim, K. Yuan, and A. H. Sayed, “A linearly convergent proximal gradient algorithm for decentralized optimization,” *arXiv preprint arXiv:1905.07996*, 2019.

[29] B. Li, S. Cen, Y. Chen, and Y. Chi, “Communication-efficient distributed optimization in networks with gradient tracking,” *arXiv preprint arXiv:1909.05844*, 2019.

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

[31] S. Pu, W. Shi, J. Xu, and A. Nedić, “A push-pull gradient method for distributed optimization in networks,” in *Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 3380–3390.

[32] K. Scaman, F. Bach, S. Bubeck, Y. Lee, and L. Massoulié, “Optimal convergence rates for convex distributed optimization in networks,” *Journal of Machine Learning Research*, vol. 20, no. 159, pp. 1–31, 2019.

[33] M. Maros and J. Jaldén, “ECO-PANDA: a computationally economic, geometrically converging dual optimization method on time-varying undirected graphs,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5257–5261.

[34] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, “A dual approach for optimal algorithms in distributed optimization over networks,” *arXiv preprint arXiv:1809.00710*, 2018.

[35] K. Srivastava and A. Nedić, “Distributed asynchronous constrained stochastic optimization,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, 2011.

[36] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu, “Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5330–5340.

[37] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, “Stochastic gradient push for distributed deep learning,” *arXiv:1811.10792*, 2018.

[38] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, “D2: Decentralized training over decentralized data,” *arXiv:1803.07068*, 2018.

[39] S. Pu and A. Nedić, “A distributed stochastic gradient tracking method,” in *Ann. Conf. Decis. Control*. IEEE, 2018, pp. 963–968.

[40] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, “Distributed stochastic optimization with gradient tracking over strongly-connected networks,” *arXiv preprint arXiv:1903.07266*, 2019.

[41] K. Yuan, S. A. Alghunaim, B. Ying, and A. H. Sayed, “On the performance of exact diffusion over adaptive networks,” *arXiv preprint arXiv:1903.10956*, 2019.

[42] J. Zhang and K. You, “Decentralized stochastic gradient tracking for empirical risk minimization,” *arXiv preprint arXiv:1909.02712*, 2019.

[43] S. Vlaski and A. H. Sayed, “Distributed learning in non-convex environments—Part I: Agreement at a linear rate,” *arXiv:1907.01848*, 2019.

[44] S. Vlaski and A. H. Sayed, “Distributed learning in non-convex environments – Part II: Polynomial escape from saddle-points,” *arXiv:1907.01849*, 2019.

[45] A. Olshevsky, L. C. Paschalidis, and S. Pu, “A non-asymptotic analysis of network independence for distributed stochastic gradient descent,” *arXiv preprint arXiv:1906.02702*, 2019.