2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Variance reduction for faster decentralized general convex optimization

Ran Xin, Subhro Das, Soummya Kar, and Usman A. Khan

Abstract— **This paper studies decentralized stochastic empirical risk minimization over a network of nodes, where each node has access to a finite collection of risk functions. While this formulation has been well-studied when each local function is strongly convex or nonconvex, it is still not clear if acceleration (in the stochastic settings) can be achieved for general convex functions. In this paper, we show that GT-SAGA, an algorithm that combines gradient tracking and incremental variance reduction, converges to a global minimizer at a provably faster rate than the existing decentralized methods for this general convex formulation. In particular, GT-SAGA achieves a topology-independent iteration and gradient complexity when the local sample size is sufficiently large. Our proof techniques hinge on a simple linear coupling of convex descent inequality and variance bounds developed for nonconvex optimization, which can be of independent interest. To the best of our knowledge, these are the first such results in decentralized general convex empirical risk minimization.**

I. Introduction

With the unprecedented growth in the amount of data being generated and collected, it is becoming increasingly challenging to train modern machine learning models on a single machine. This is because storing and processing very large amounts of data is typically beyond the capability of existing computational devices. A workaround that has been quite popular is decentralized training where the original dataset is stored and processed over multiple machines (GPUs, workstations) and the machines communicate with each other (or, more recently, through a parameter server) to learn the corresponding model parameters.

Decentralized training of machine learning models can be equivalently thought of as decentralized optimization where the objective of the optimization problem is to minimize the (expected or empirical) loss incurred by the model in predicting labels versus the ground truth. Examples of such machine learning problems include regression, prediction, or classification, which can equivalently be written as quadratic, (strongly) convex, or nonconvex optimization problems. A strongly convex problems is where the convex loss landscape is further bounded below by a quadratic and this geometry (i.e., having a certain curvature near the unique global minimum) typically leads to a cleaner analysis and faster convergence properties.

Much of the existing work in decentralized optimization has focused on quadratic, strongly convex, or nonconvex problems. This is potentially because adding a quadratic regularizer to any convex problem makes it strongly convex and strong convexity bounds are arguably simpler to integrate in the subsequent analysis. There is no easy way however to escape nonconvexity and thus the attention dedicated to nonconvex problems is justified because of their applicability to many emergent applications. In contrast, work on decentralized convex optimization, even though it exists, has been rather scarce. Clearly, the utility of convex problems still remains significant as enforcing a specialized curvature near the stationary points of the corresponding loss functions is not always meaningful.

In this paper, we consider decentralized convex optimization and using the well known ideas of gradient tracking and variance reduction show that our proposed decentralized method **GT-SAGA** is provably faster than the existing methods for general convex problems. In particular, we develop a linear coupling of the *convex descent inequality* and *variance bounds for nonconvex optimization* that enables establishing the global convergence in terms of global function value in decentralized convex problems. This linear coupling is of independent interest and the methodology may be useful in other extensions in the corresponding problem domains. The main contributions of this work include the following.

(i) We show that the optimality gap in **GT-SAGA** iterates scales sublinearly in m. This is in contrast to linear scaling in the decentralizd batch gradient methods, e.g., [1], [2], where m is the size of local data points at each node;

(ii) We show that the addition of variance reduction leads to a convergence rate of $\mathcal{O}(1/K)$, which matches the rate of centralized stochastic variance reduced methods [3] for convex problems. Note that this rate is faster than $\mathcal{O}(1/\sqrt{K})$ for decentralized stochastic gradient methods without variance reduction, e.g., [4], [5];

(iii) We show that when m is sufficiently large, the rate of **GT-SAGA** is network-topology independent.

Related Work: Early work on decentralized optimization can be found in e.g., [6]–[9] that builds upon undirected networks and doubly stochastic weight matrices. For arbitrary directed networks, the corresponding decentralized methods rely on row and/or column stochastic weights and can be found in [10]–[13]. These methods are built on local gradient corrections at the agents and thus incur a certain steady-state error under constant stepsizes, which can be removed with a decaying stepsize however at the expense of a sublinear convergence. Linear convergence to the exact solution with

RX is with ByteDance, San Jose, CA, USA, SD is with IBM Research, Cambridge, MA, USA, SK is with Carnegie Mellon University, Pittsburgh, PA, USA, and UAK is with Tufts University, Boston, MA, USA. Email addresses: ran.xin@bytedance.com, subhro.das@ibm.com, soummyak@andrew.cmu.edu, khan@ece.tufts.edu. This work is partially supported by NSF under the award 2230630.

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

a constant stepsize is subsequently achieved with the help of gradient tracking [14]–[18]. Work on decentralized stochastic optimization can be found, e.g., in [19]–[22]. Stochastic optimization has been accelerated with the help of variance reduction where relevant work in the centralized settings cane be found in [3], [23]–[27]. Existing variance-reduced decentralized stochastic methods can be found in [28]–[40]. Most of this work is restricted to either strongly convex or nonconvex problems. In contrast, in this paper, we combine gradient tracking with variance reduction and provide an algorithm that guarantees improved performance for general convex problems.

Notation: We use lowercase bold letters to denote vectors and uppercase bold letters to denote matrices. The $d \times d$ identity matrix is denoted by $\mathbf{I}_d$, while the d-dimensional column vectors of all ones and zeros are represented by $\mathbf{1}_d$ and $\mathbf{0}_d$, respectively. We use $\mathbf{X} \otimes \mathbf{Y}$ to denote the Kronecker product of two matrices $\mathbf{X}$ and $\mathbf{Y}$. The Euclidean norm of a vector or the spectral norm of a matrix is denoted by $\|\cdot\|$. We work with a rich enough probability triple $(\Theta, \mathcal{F}, \mathbb{P})$, where all random objects are defined properly. We make a blanket assumption that each node i at every iteration k is able to obtain i.i.d. minibatch samples from its local data. The induced natural filtration is denoted by $\mathcal{F}^k$, and increasing family of sub σ-algebras that represents the historical information of the algorithm with samples up to iteration k. Subsequently, we will use conditional expectation with respect to this filtration $\mathcal{F}^k$.

We now describe the rest of the paper. Section II describes the problem formulation while Section III recaps the **GT-SAGA** algorithm and provides the main results of this paper. Section IV provides the convergence analysis with detailed proofs of the corresponding descent inequalities and linear coupling. Finally, Section V concludes the paper.

II. Problem Formulation

We consider decentralized optimization problems over a network of nodes. In particular, there are n nodes communicating over a directed graph $\mathcal{G} := \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} := \{1, \cdots, n\}$ 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 private collection of m smooth convex functions $\{f_{i,j} : \mathbb{R}^p \to \mathbb{R}\}_{j=1}^m$, that can be viewed as a cost associated with the j-th data sample at the i-th node, while $f_i := \sum_j f_{i,j}$ is the local cost function at node i. The goal of the networked nodes is to solve the following optimization problem:

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

In other words, the nodes must agree on a global minimizer of F via local computation and communication at each node that is restricted by the graph $\mathcal{G}$.

III. **GT-SAGA** Algorithm and Main Results

GT-SAGA [40] builds upon local SAGA estimators [3] and global gradient tracking [14], [15], and is formally presented in Algorithm 1. We refer the readers to [38]–[40] for detailed discussion on the development of **GT-SAGA**.

Algorithm 1 **GT-SAGA** at each node i

```
Require: $\mathbf{x}_i^0 = \overline{\mathbf{x}}^0 \in \mathbb{R}^p$; $\alpha \in \mathbb{R}^+$; $\{\underline{w}_{ir}\}_{r=1}^n$; $\mathbf{y}_i^0 = \mathbf{0}_p$;
         $\mathbf{z}_{i,j}^0 = \mathbf{x}_i^0, \forall j \in \{1, \ldots, m\}$; $\mathbf{g}_i^{-1} = \mathbf{0}_p$.
for $k = 0, 1, 2, \ldots$ do
    Pick $\tau_i^k$ uniformly at random from $\{1, \ldots, m\}$;
    Compute a local SAGA estimator $\mathbf{g}_i^k$:
        $\mathbf{g}_i^k = \nabla f_{i,\tau_i^k}(\mathbf{x}_i^k) - \nabla f_{i,\tau_i^k}(\mathbf{z}_{i,\tau_i^k}^k) + \frac{1}{m} \sum_{j=1}^m \nabla f_{i,j}(\mathbf{z}_{i,j}^k)$;
    $\mathbf{y}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir} \left(\mathbf{y}_r^k + \mathbf{g}_r^k - \mathbf{g}_r^{k-1}\right)$;
    $\mathbf{x}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir} \left(\mathbf{x}_r^k - \alpha \mathbf{y}_r^{k+1}\right)$;
    Pick $s_i^k$ uniformly at random from $\{1, \cdots, m\}$;
    Set $\mathbf{z}_{i,j}^{k+1} = \mathbf{x}_i^k, j = s_i^k$;  $\mathbf{z}_{i,j}^{k+1} = \mathbf{z}_{i,j}^k, j \neq s_i^k$;
end for
```

Before we provide the main results of this paper, we formally present our assumptions.

Assumption 1. *Each f_i is L-smooth and convex. Moreover, $-\infty < F^* := \inf_{\mathbf{x}} F(\mathbf{x})$ is achieved by some $\mathbf{x}^* \in \mathbb{R}^p$.*

Assumption 2. *The $n \times n$ weight matrix $\underline{\mathbf{W}} = \{\underline{w}_{ir}\}$ associated with 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 $\underline{\mathbf{W}}$.*

The following theorem presents the global convergence rate of **GT-SAGA** under general convexity with the help of the following variables:

$$\overline{\mathbf{x}}^k := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^k, \qquad \widehat{\mathbf{x}}_i^K := \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{x}_i^k,$$

and let $\nabla \mathbf{f}_k \in \mathbb{R}^{np}$ concatenate all local exact gradients $\nabla f_i(\mathbf{x}_i^k)$'s.

Theorem 1. *Let Assumptions 1 and 2 hold. If the positive step-size α in **GT-SAGA** is such that*

$$\alpha \leq \min \left\{ \frac{1}{8}, \frac{(1-\lambda^2)^{3/4}}{18 \lambda^{1/2} m^{1/2}}, \frac{(1-\lambda^2)^2}{12 \lambda}, \frac{n^{1/3}}{16 m^{2/3}} \right\} \frac{1}{L},$$

then we have

$$\sum_{k=0}^{K-1} \mathbb{E}\Big[F(\overline{\mathbf{x}}^k) - F^*\Big] \lesssim \frac{\|\overline{\mathbf{x}}^0 - \mathbf{x}^*\|^2}{\alpha} + \frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{nL}.$$

Since f is convex, we have by Jensen's inequality:

$$\mathbb{E}\left[F\left(\frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{x}}_i^K\right) - F^*\right] \lesssim \frac{\|\overline{\mathbf{x}}^0 - \mathbf{x}^*\|^2}{\alpha K} + \frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{nLK}.$$

If α attains its upper bound, then

$$\mathbb{E}\left[F\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{x}}_i^K\right)-F^*\right]$$
$$\lesssim \max\left\{\frac{m^{1/2}}{(1-\lambda)^{3/4}}, \frac{1}{(1-\lambda)^2}, \frac{m^{2/3}}{n^{1/3}}\right\}\frac{L\|\overline{\mathbf{x}}^0-\mathbf{x}^*\|^2}{K}$$
$$+\frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{nLK}.$$

If the local sample size m is large enough, the rate becomes topology-independent, i.e.,

$$\mathbb{E}\left[F\left(\frac{1}{n}\textstyle\sum_{i=1}^{n}\widehat{\mathbf{x}}_i^K\right)-F^*\right]$$
$$\lesssim \frac{Lm^{2/3}\|\overline{\mathbf{x}}^0-\mathbf{x}^*\|^2}{n^{1/3}K}+\frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{nLK}.$$

We have the following observations from Theorem 1.

- The dependence of local sample size m on the convergence rates shown in Theorem 1 is sublinear, i.e., $m^{2/3}$. This observation is critical in that the dependence of m in existing decentralized batch gradient methods [1] is linear, i.e., m. As a result, **GT-SAGA** is provably faster than the batch gradient methods when the local sample size is large.
- The convergence rate of **GT-SAGA** is $O(1/K)$, which is strictly faster than the rate $O(1/\sqrt{K})$ of decentralized stochastic gradient methods for convex problems without variance reduction, e.g., [4].
- When the local sample size m is sufficiently large, the rate of **GT-SAGA** is topology-independent. This is the first such result for decentralized stochastic convex problems with variance reduction.

In the next section, we analyze **GT-SAGA** for smooth convex functions and derive its convergence properties.

IV. Convergence Analysis

In this section, we present the proof of Theorem 1. Our proof techniques rely on a simple linear coupling of the convex descent inequality and variance bounds developed in our earlier work for nonconvex optimization [40], which can be of independent interest and useful for other decentralized algorithms based on similar principles. Before we proceed, we define network-wide averages of certain variable from Algorithm 1 that will be useful in the subsequent analysis:

$$\overline{\mathbf{g}}^k=\sum_{i=1}^{n}\frac{\mathbf{g}_i^k}{n},\ \overline{\nabla\mathbf{f}}(\mathbf{x}^k)=\sum_{i=1}^{n}\frac{\nabla f_i(\mathbf{x}_i^k)}{n},\ \mathbf{J}=\frac{\mathbf{1}\mathbf{1}^\top\otimes\mathbf{I}_p}{n}.$$

Moreover, we let $\mathbf{x}^k$ and $\mathbf{g}^k$ stack all of the local variables $\mathbf{x}_i^k$'s and $\mathbf{g}_i^k$'s, respectively.

A. Descent Inequality with Convexity

In this subsection, we provide a descent lemma for decentralized convex problems. Although the following lemma is developed for **GT-SAGA**, we emphasize that the same proof holds true for any decentralized stochastic gradient tracking based methods with unbiased gradient estimators.

Lemma 1 (Convex Descent). *If $0<\alpha\leq\frac{1}{8L}$, then we have, $\forall k\geq 0$,*

$$\mathbb{E}\Big[\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2|\mathcal{F}^k\Big]\leq\|\overline{\mathbf{x}}^k-\mathbf{x}^*\|^2-\alpha\big(F(\overline{\mathbf{x}}^k)-F^*\big)$$
$$+\alpha^2\mathbb{E}[\|\overline{\mathbf{g}}^k-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k]$$
$$+\tfrac{2\alpha L}{n}\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2.$$

Proof: From the $\mathbf{x}_i^k$-update in Algorithm 1, we get

$$\overline{\mathbf{x}}^{k+1}=\overline{\mathbf{x}}^k-\alpha\overline{\mathbf{g}}^k,\qquad\forall k\geq 0,$$

by summing both sides over $i=1,\ldots,n$ and noting that $\sum_r\mathbf{y}_r^k=\sum_r\mathbf{g}_r^k$. We now consider the following recursion: $\forall k\geq 0$,

$$\begin{aligned}&\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2\\&=\|\overline{\mathbf{x}}^k-\alpha\overline{\mathbf{g}}^k-\mathbf{x}^*\|^2\\&=\|\overline{\mathbf{x}}^k-\mathbf{x}^*\|^2-2\alpha\langle\overline{\mathbf{g}}^k,\overline{\mathbf{x}}^k-\mathbf{x}^*\rangle+\alpha^2\|\overline{\mathbf{g}}^k\|^2.\end{aligned}$$

Taking the conditional expectation over the filtration $\mathcal{F}^k$, we have: $\forall k\geq 0$,

$$\begin{aligned}&\mathbb{E}[\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2|\mathcal{F}^k]\\&=\|\overline{\mathbf{x}}^k-\mathbf{x}^*\|^2-2\alpha\langle\mathbb{E}[\overline{\mathbf{g}}^k|\mathcal{F}^k],\overline{\mathbf{x}}^k-\mathbf{x}^*\rangle+\alpha^2\mathbb{E}[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k]\\&=\|\overline{\mathbf{x}}^k-\mathbf{x}^*\|^2-2\alpha\langle\overline{\nabla\mathbf{f}}(\mathbf{x}^k),\overline{\mathbf{x}}^k-\mathbf{x}^*\rangle+\alpha^2\mathbb{E}[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k].\end{aligned}\tag{1}$$

For the last term on the RHS of (1), we have: $\forall k\geq 0$,

$$\begin{aligned}&\mathbb{E}[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k]\\&=\mathbb{E}[\|\overline{\mathbf{g}}^k-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)+\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k]\\&=\mathbb{E}[\|\overline{\mathbf{g}}^k-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k]+\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\\&=\mathbb{E}[\|\overline{\mathbf{g}}^k-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k]\\&\quad+\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)-\nabla F(\overline{\mathbf{x}}^k)+\nabla F(\overline{\mathbf{x}}^k)\|^2\\&\leq\mathbb{E}[\|\overline{\mathbf{g}}^k-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k]+\tfrac{2L^2}{n}\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2\\&\quad+2\|\nabla F(\overline{\mathbf{x}}^k)\|^2,\end{aligned}\tag{2}$$

where the second equality uses the fact that the conditional expectation of the inner product in the cross term is zero and the last inequality uses the triangle inequality and the definition of L-smoothness. Subsequently, using (2) in (1) gives: $\forall k\geq 0$,

$$\begin{aligned}&\mathbb{E}[\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2|\mathcal{F}^k]\\&\leq\|\overline{\mathbf{x}}^k-\mathbf{x}^*\|^2-2\alpha\langle\overline{\nabla\mathbf{f}}(\mathbf{x}^k),\overline{\mathbf{x}}^k-\mathbf{x}^*\rangle\\&\quad+2\alpha^2\|\nabla F(\overline{\mathbf{x}}^k)\|^2+\alpha^2\mathbb{E}[\|\overline{\mathbf{g}}^k-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k]\\&\quad+\frac{2\alpha^2L^2}{n}\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\|^2.\end{aligned}\tag{3}$$

To proceed, we focus on the second term on the right hand side of the above equation. To this aim, define

$$A^k:=\langle\overline{\nabla\mathbf{f}}(\mathbf{x}^k),\overline{\mathbf{x}}^k-\mathbf{x}^*\rangle=\frac{1}{n}\sum_{i=1}^{n}\langle\nabla f_i(\mathbf{x}_i^k),\overline{\mathbf{x}}^k-\mathbf{x}^*\rangle.$$

We next handle A^k with the help of the smoothness and convexity of each f_i. We first note that

$$A^k = \frac{1}{n}\sum_{i=1}^{n} \Big(\underbrace{\langle \nabla f_i(\mathbf{x}_i^k), \mathbf{x}_i^k - \mathbf{x}^* \rangle}_{B_i^k} + \underbrace{\langle \nabla f_i(\mathbf{x}_i^k), \overline{\mathbf{x}}^k - \mathbf{x}_i^k \rangle}_{C_i^k} \Big), \tag{4}$$

and then consider its two terms B_i^k and C_i^k separately. Towards B_i^k, we recall that since each f_i is convex, it lies above all of its tangents and thus satisfies a *linear lower bound*:

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle,$$

for all $\mathbf{x}, \mathbf{y}, i$, that is,

$$\langle \nabla f_i(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle \geq f_i(\mathbf{x}) - f_i(\mathbf{y}),$$

for all $\mathbf{x}, \mathbf{y}, i$. Setting $\mathbf{x} = \mathbf{x}_i^k$ and $\mathbf{y} = \mathbf{x}^*$ in the inequality above leads to

$$B_i^k := \langle \nabla f_i(\mathbf{x}_i^k), \mathbf{x}_i^k - \mathbf{x}^* \rangle \geq f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}^*), \tag{5}$$

for all $\mathbf{x}, \mathbf{y}, i$. Towards C_i^k, we recall that since each f_i is L-smooth, f_i satisfies a *quadratic upper bound*, i.e.,

$$f_i(\mathbf{y}) \leq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \tfrac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2,$$

for all $\mathbf{x}, \mathbf{y}, i$, that is

$$\langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq f_i(\mathbf{y}) - f_i(\mathbf{x}) - \tfrac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2,$$

for all $\mathbf{x}, \mathbf{y}, i$. Setting $\mathbf{x} = \mathbf{x}_i^k$ and $\mathbf{y} = \overline{\mathbf{x}}^k$ in the equation for C_k^i leads to

$$\begin{aligned} C_i^k &:= \langle \nabla f_i(\mathbf{x}_i^k), \overline{\mathbf{x}}^k - \mathbf{x}_i^k \rangle \\ &\geq f_i(\overline{\mathbf{x}}^k) - f_i(\mathbf{x}_i^k) - \tfrac{L}{2}\|\overline{\mathbf{x}}^k - \mathbf{x}_i^k\|^2, \end{aligned} \tag{6}$$

for all $\mathbf{x}, \mathbf{y}, i$. Now, we use (5) and (6) in (4) to obtain: $\forall k \geq 0$,

$$\begin{aligned} A^k &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \big(B_i^k + C_i^k\big) \\ &\geq \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \big(f_i(\overline{\mathbf{x}}^k) - f_i(\mathbf{x}^*) - \tfrac{L}{2}\|\overline{\mathbf{x}}^k - \mathbf{x}_i^k\|^2\big) \\ &= F(\overline{\mathbf{x}}^k) - F^* - \tfrac{L}{2n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2. \end{aligned} \tag{7}$$

Finally, we use (7) in (3) to obtain: $\forall k \geq 0$,

$$\begin{aligned} &\mathbb{E}[\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 | \mathcal{F}^k] \\ &\leq \|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 - 2\alpha\left(F(\overline{\mathbf{x}}^k) - F^* - \tfrac{L}{2n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right) \\ &\quad + 2\alpha^2\|\nabla F(\overline{\mathbf{x}}^k)\|^2 \\ &\quad + \alpha^2\mathbb{E}[\|\overline{\mathbf{g}}^k - \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k] + \tfrac{2\alpha^2 L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 \\ &= \|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \alpha\big(F(\overline{\mathbf{x}}^k) - F^*\big) + 2\alpha^2\|\nabla F(\overline{\mathbf{x}}^k)\|^2 \\ &\quad + \alpha^2\mathbb{E}[\|\overline{\mathbf{g}}^k - \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k] \\ &\quad + (1 + 2\alpha L)\,\tfrac{\alpha L}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 \\ &\leq \|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 - 2\alpha(1 - 2\alpha L)\big(F(\overline{\mathbf{x}}^k) - F^*\big) \\ &\quad + \alpha^2\mathbb{E}[\|\overline{\mathbf{g}}^k - \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k] \\ &\quad + (1 + 2\alpha L)\,\tfrac{\alpha L}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2, \end{aligned}$$

where we use $\|\nabla F(\mathbf{x})\|^2 \leq 2L(F(\mathbf{x}) - F^*)$ in the last inequality. The proof follows by $0 < \alpha \leq \frac{1}{8L}$. □

The above lemma is akin to the standard descent inequality for gradient descent algorithms that is further specialized to decentralized problems. In particular: the first two terms in Lemma 1's statement establish descent, i.e., the average iterate $\overline{\mathbf{x}}^k$ gets closer and closer to $\mathbf{x}^*$ because $\alpha(F(\overline{\mathbf{x}}^k) - F^*)$ is strictly positive; the third term is the error due to estimated gradients and incorporate variance reduction; while the last term is the consensus error, i.e., how far are the local iterates (concatenated in $\mathbf{x}^k$) from the average iterate at time k since $\mathbf{J}$ is the averaging matrix. The next step is to linearly couple the third and fourth terms and then refine Lemma 1. We describe this in the next subsection. Before we proceed, we provide a useful result.

Lemma 2. *The following holds true for all $k \geq 0$.*

$$\begin{aligned} \|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 &= \|\overline{\nabla\mathbf{f}}(\mathbf{x}^k) - \nabla F(\overline{\mathbf{x}}^k) + \nabla F(\overline{\mathbf{x}}^k)\|^2 \\ &\leq 2\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k) - \nabla F(\overline{\mathbf{x}}^k)\|^2 + 2\|\nabla F(\overline{\mathbf{x}}^k)\|^2 \\ &\leq \tfrac{2L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 4L(F(\overline{\mathbf{x}}^k) - F^*). \end{aligned} \tag{8}$$

B. Refined Descent Inequality with Convexity and Variance Reduction

We first construct two auxiliary $\mathcal{F}^k$-adapted sequences: $\forall i \in \mathcal{V}$, $\forall k \geq 0$,

$$Q_i^k := \tfrac{1}{m}\textstyle\sum_{j=1}^{m}\|\overline{\mathbf{x}}^k - \mathbf{z}_{i,j}^k\|^2, \qquad Q^k := \tfrac{1}{n}\textstyle\sum_{i=1}^{n} Q_i^k,$$

and recall that the bound on the variance of the gradient estimator $\mathbf{g}_i$ is given by the following lemma [40].

Lemma 3 **([40]).** *The following holds: $\forall k \geq 0$,*

$$\begin{aligned} \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k)\|^2|\mathcal{F}^k\big] \leq \tfrac{2L^2}{n}&\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 \\ &+ 2L^2 Q^k. \end{aligned} \tag{9}$$

As explained earlier, we now use Lemma 3 to further refine Lemma 1.

Lemma 4 **(Convex Descent with Variance Reduction).** *If $0 < \alpha \leq \frac{1}{8L}$, then $\forall k \geq 0$,*

$$\begin{aligned} &\mathbb{E}\Big[\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\|^2|\mathcal{F}^k\Big] \\ &\leq \|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \alpha\big(F(\overline{\mathbf{x}}^k) - F^*\big) \\ &\quad + \tfrac{2L^2\alpha^2}{n}Q^k + \tfrac{4L\alpha}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2. \end{aligned}$$

Proof: Applying Lemma 3 in Lemma 1, we have the following: $\forall k \geq 0$,

$$\begin{aligned} &\mathbb{E}\Big[\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\|^2|\mathcal{F}^k\Big] \\ &\leq \|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \alpha\big(F(\overline{\mathbf{x}}^k) - F^*\big) \\ &\quad + \alpha^2\mathbb{E}\Big[\|\overline{\mathbf{g}}^k - \overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k\Big] + \tfrac{2\alpha L}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 \\ &= \|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \alpha\big(F(\overline{\mathbf{x}}^k) - F^*\big) \\ &\quad + \tfrac{\alpha^2}{n^2}\mathbb{E}\Big[\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2|\mathcal{F}^k\Big] + \tfrac{2\alpha L}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 \\ &\leq \|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \alpha\big(F(\overline{\mathbf{x}}^k) - F^*\big) + \tfrac{2L^2\alpha^2}{n}Q^k \\ &\quad + \left(\tfrac{2L^2\alpha^2}{n^2} + \tfrac{2L\alpha}{n}\right)\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2, \end{aligned}$$

where the last line uses Lemma 3. The proof follows by using the fact that $0 < \alpha \le \frac{1}{8L}$. □

The following corollary is immediate from Lemma 4 and will become the key result in order to prove Theorem 1.

Corollary 1. *If $0 < \alpha \le \frac{1}{8L}$, then*

$$\begin{aligned}&\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big]\\ &\le \tfrac{1}{\alpha}\|\overline{\mathbf{x}}^0 - \mathbf{x}^*\|^2 + \tfrac{2L^2\alpha}{n}\textstyle\sum_{k=0}^{K-1}\mathbb{E}[Q^k]\\ &+ 4L\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\big].\end{aligned} \tag{10}$$

Proof: Taking the telescoping sum of both sides in Lemma 4 over $k = 0, \dots, K-1$, gives: if $0 < \alpha \le \frac{1}{8L}$, then $\forall k \ge 0$,

$$\begin{aligned}\mathbb{E}\Big[\|\overline{\mathbf{x}}^K - \mathbf{x}^*\|^2\Big] \le \|\overline{\mathbf{x}}^0 - \mathbf{x}^*\|^2&\\ - \alpha\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big]&\\ + \tfrac{2L^2\alpha^2}{n}\textstyle\sum_{k=0}^{K-1}\mathbb{E}[Q^k]&\\ + 4L\alpha\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\big]&.\end{aligned}$$

The proof follows by rearranging the above terms and by dropping the negative norm from the upper bound. □

Here, we briefly explain Corollary 1. First, note that the left side of (10) is the target metric, i.e., the optimality gap $\sum_{k=0}^{K-1}\mathbb{E}[F(\overline{\mathbf{x}}^k) - F^*]$, of Theorem 1. In order to arrive at the result of Theorem 1, it remains to bound $\sum_{k=0}^{K-1}\mathbb{E}[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2]$ and $\sum_{k=0}^{K-1}\mathbb{E}[Q^k]$ in terms of the optimality gap. Although these bounds are not explicitly derived in [41], some of the results in [41] can be easily refined for convex problems as we describe in the next subsection.

C. Auxiliary Results

We recall the following lemma from [40] that was established without convexity.

Lemma 5 ([40])**.** *If $0 < \alpha \le \min\Big\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\Big\}\frac{1}{L}$, then we have: $\forall K \ge 1$,*

$$\begin{aligned}\textstyle\sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \le \tfrac{16\lambda^4\alpha^2}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}&\\ + \Big(97m^2 + \tfrac{8\lambda^2}{1-\lambda^2}\Big)\tfrac{32\lambda^2\alpha^4L^2}{(1-\lambda^2)^3}\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big],&\end{aligned} \tag{11}$$

and, $\forall K \ge 1$,

$$\begin{aligned}\textstyle\sum_{k=0}^{K}\mathbb{E}\left[Q^k\right] \le \tfrac{114\lambda^4\alpha^2}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}&\\ + 33m^2\alpha^2\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big].&\end{aligned} \tag{12}$$

We first refine the consensus error bound.

Lemma 6 (Variance-Reduced Consensus Bound). *If the step-size α is such that*

$$0 < \alpha \le \min\left\{\tfrac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}, \tfrac{1-\lambda^2}{12\lambda}, \tfrac{1}{8}, \tfrac{\sqrt{n}}{\sqrt{8m}}\right\}\tfrac{1}{L},$$

then we have

$$\begin{aligned}4L\textstyle\sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \le \tfrac{128\lambda^4\alpha^2L}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}&\\ + \tfrac{1}{4}\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big].&\end{aligned}$$

Proof: We recall from (8) that, $\forall k \ge 0$,

$$\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 \le \tfrac{2L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 4L(F(\overline{\mathbf{x}}^k) - F^*).$$

Using this inequality in (11) gives:

$$\begin{aligned}&\textstyle\sum_{k=0}^{K}\mathbb{E}\big[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\big]\\ &\le \tfrac{16\lambda^4\alpha^2}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}\\ &+ \Big(97m^2 + \tfrac{8\lambda^2}{1-\lambda^2}\Big)\tfrac{64\lambda^2\alpha^4L^4}{(1-\lambda^2)^3}\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\big]\\ &+ \Big(97m^2 + \tfrac{8\lambda^2}{1-\lambda^2}\Big)\tfrac{128\lambda^2\alpha^4L^3}{(1-\lambda^2)^3}\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big].\end{aligned}$$

We note that if $0 < \alpha \le \min\Big\{\frac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}, \frac{1-\lambda^2}{12\lambda}\Big\}\frac{1}{L}$, then $\big(97m^2 + \frac{8\lambda^2}{1-\lambda^2}\big)\frac{64\lambda^2\alpha^4L^4}{(1-\lambda^2)^3} \le \frac{1}{64}$, and therefore

$$\begin{aligned}\textstyle\sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \le \tfrac{16\lambda^4\alpha^2}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}&\\ + \tfrac{1}{2}\mathbb{E}\left[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] + \tfrac{1}{32L}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big],&\end{aligned}$$

that is

$$\begin{aligned}\textstyle\sum_{k=0}^{K}\mathbb{E}\left[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \le \tfrac{32\lambda^4\alpha^2}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}&\\ + \tfrac{1}{16L}\textstyle\sum_{k=0}^{K-1}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big],&\end{aligned} \tag{13}$$

and the proof follows. □

Next, we refine the error bound of Q^k.

Lemma 7 (**Bound on the Q^k Sum**)**.** *If the step-size α is such that $0 < \alpha \le \min\Big\{\frac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}, \frac{1-\lambda^2}{12\lambda}, \frac{1}{8}, \frac{n^{1/3}}{8m^{2/3}}\Big\}\frac{1}{L}$, then we have*

$$\begin{aligned}\tfrac{4L^2\alpha}{n}\textstyle\sum_{k=0}^{K}\mathbb{E}\left[Q^k\right] \le \tfrac{200\lambda^4\alpha^2L}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}&\\ + \tfrac{1}{4}\textstyle\sum_{k=0}^{K-1}\left(F(\overline{\mathbf{x}}^k) - F^*\right).&\end{aligned}$$

Proof: Using (8) in (12) gives:

$$\begin{aligned}\textstyle\sum_{k=0}^{K}\mathbb{E}\left[Q^k\right] &\le \tfrac{114\lambda^4\alpha^2}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}\\ &\quad + 66m^2L^2\alpha^2\textstyle\sum_{k=0}^{K-1}\mathbb{E}\left[\tfrac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right]\\ &\quad + 132m^2\alpha^2L\textstyle\sum_{k=0}^{K-1}\left(F(\overline{\mathbf{x}}^k) - F^*\right)\\ &\le \left(20m^2L^2\alpha^2 + 1\right)\tfrac{114\lambda^4\alpha^2}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}\\ &\quad + 136m^2\alpha^2L\textstyle\sum_{k=0}^{K-1}\left(F(\overline{\mathbf{x}}^k) - F^*\right),\end{aligned}$$

where the last line is due to (13). Hence, if $0 < \alpha \le \min\Big\{\frac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}, \frac{1-\lambda^2}{12\lambda}, \frac{1}{8}, \frac{n^{1/3}}{16m^{2/3}}\Big\}\frac{1}{L}$, then we have

$$\begin{aligned}&\tfrac{4L^2\alpha}{n}\textstyle\sum_{k=0}^{K}\mathbb{E}\left[Q^k\right]\\ &\le \tfrac{4L\alpha}{n}\Big(20m^2L^2\alpha^2 + 1\Big)\tfrac{114\lambda^4\alpha^2L}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}\\ &\quad + \tfrac{600m^2\alpha^3L^3}{n}\textstyle\sum_{k=0}^{K-1}\left(F(\overline{\mathbf{x}}^k) - F^*\right)\\ &\le \tfrac{200\lambda^4\alpha^2L}{(1-\lambda^2)^3}\tfrac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n} + \tfrac{1}{4}\textstyle\sum_{k=0}^{K-1}\left(F(\overline{\mathbf{x}}^k) - F^*\right),\end{aligned}$$

and the proof follows. □

D. Proof of Theorem 1

Finally, it is straightforward to obtain Theorem 1 by applying Lemma 6 and Lemma 7 to (10) in Corollary 1 after some standard algebraic manipulations.

V. Conclusions

In this paper, we prove that **GT-SAGA** achieves the best known rate for decentralized general convex optimization. Our convergence analysis uses a novel linear coupling of descent inequality and the variance bounds. We show a sublinear scaling $\mathcal{O}(m^{2/3})$ of the optimality gap in terms of the number of samples m at each node in contrast to a linear scaling in the existing work. Moreover, **GT-SAGA** converges at $\mathcal{O}(1/K)$, in contrast to $\mathcal{O}(1/\sqrt{K})$ for stochastic gradient methods without variance reduction in convex problems, and its convergence is network-topology independent in the big-data (large m) regime.

References

[1] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control. Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, 2017.

[2] J. Xu, Y. Tian, Y. Sun, and G. Scutari, "Distributed algorithms for composite optimization: Unified and tight convergence analysis," *arXiv:2002.11534*, 2020.

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

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

[5] S. A. Alghunaim and K. Yuan, "An enhanced gradient-tracking bound for distributed online stochastic convex optimization," *arXiv preprint arXiv:2301.02855*, 2023.

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

[7] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1663–1707, 2010.

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

[9] 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, 2012.

[10] K. I. Tsianos, *The role of the Network in Distributed Optimization Algorithms: Convergence Rates, Scalability, Communication/Computation Tradeoffs and Communication Delays*, Ph.D. thesis, Dept. Elect. Comp. Eng. McGill University, 2013.

[11] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. on Autom. Control*, vol. 60, no. 3, pp. 601–615, Mar. 2015.

[12] C. Xi and U. A. Khan, "DEXTRA: A fast algorithm for optimization over directed graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 4980–4993, Oct. 2017.

[13] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," *IEEE Trans. on Autom. Control*, vol. 62, no. 8, pp. 3986–3992, Oct. 2016.

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

[15] 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.

[16] 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, May 2018.

[17] C. Xi, V. S. Mai, R. Xin, E. Abed, and U. A. Khan, "Linear convergence in optimization over directed graphs with row-stochastic matrices," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3558–3565, Oct. 2018.

[18] 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.

[19] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545, 2010.

[20] 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 *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5330–5340.

[21] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2: Decentralized training over decentralized data," in *International Conference on Machine Learning*, 2018, pp. 4848–4856.

[22] 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.*, 2020.

[23] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, 2013, pp. 315–323.

[24] 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.

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

[26] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8194–8244, 2017.

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

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

[29] K. Yuan, B. Ying, J. Liu, and A. H. Sayed, "Variance-reduced stochastic learning by networked agents under random reshuffling," *IEEE Trans. Signal Process.*, vol. 67, no. 2, pp. 351–366, 2018.

[30] Z. Shen, A. Mokhtari, T. Zhou, P. Zhao, and H. Qian, "Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication," in *International Conference on Machine Learning*, 2018, pp. 4624–4633.

[31] Z. Wang and H. Li, "Edge-based stochastic gradient algorithm for distributed optimization," *IEEE Trans. Network Science and Engineering*, 2019.

[32] H. Hendrikx, F. Bach, and L. Massoulié, "An accelerated decentralized stochastic proximal algorithm for finite sums," in *Advances in Neural Information Processing Systems*, 2019, pp. 952–962.

[33] 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 Processing Magazine*, vol. 37, no. 3, pp. 102–113, 2020.

[34] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *arXiv preprint arXiv:1912.04230*, 2019.

[35] H. Sun, S. Lu, and M. Hong, "Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking," in *Proceedings of the 37th International Conference on Machine Learning*, 13–18 Jul 2020, vol. 119, pp. 9217–9228.

[36] B. Li, S. Cen, Y. Chen, and Y. Chi, "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1662–1672.

[37] M. I. Qureshi, R. Xin, S. Kar, and U. A. Khan, "Push-SAGA: A decentralized stochastic algorithm with variance reduction over directed graphs," *arXiv preprint arXiv:2008.06082*, 2020.

[38] 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.

[39] 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, 2020.

[40] R. Xin, U. A. Khan, and S. Kar, "A fast randomized incremental gradient method for decentralized non-convex optimization," *IEEE Transactions in Automatic Conrol*, vol. 67, pp. 5150–5165.

[41] 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.