An improved convergence analysis for decentralized online stochastic non-convex optimization

Ran Xin, Usman A. Khan, and Soummya Kar

***Abstract*—In this paper, we study decentralized online stochastic non-convex optimization over a network of nodes. Integrating a technique called gradient tracking in decentralized stochastic gradient descent, we show that the resulting algorithm, `GT-DSGD`, enjoys certain desirable characteristics towards minimizing a sum of smooth non-convex functions. In particular, for general smooth non-convex functions, we establish non-asymptotic characterizations of `GT-DSGD` and derive the conditions under which it achieves network-independent performances that match the centralized minibatch `SGD`. In contrast, the existing results suggest that `GT-DSGD` is always network-dependent and is therefore strictly worse than the centralized minibatch `SGD`. When the global non-convex function additionally satisfies the Polyak-Łojasiewics (PL) condition, we establish the linear convergence of `GT-DSGD` up to a steady-state error with appropriate constant step-sizes. Moreover, under stochastic approximation step-sizes, we establish, for the first time, the optimal global sublinear convergence rate on almost every sample path, in addition to the asymptotically optimal sublinear rate in expectation. Since strongly convex functions are a special case of the functions satisfying the PL condition, our results are not only immediately applicable but also improve the currently known best convergence rates and their dependence on problem parameters.**

***Index Terms*—Decentralized optimization, stochastic gradient methods, non-convex problems, multi-agent systems.**

I. Introduction

This paper considers decentralized non-convex optimization where n nodes cooperate to solve the following problem:

$$\text{P1:} \qquad \min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

such that each function $f_i : \mathbb{R}^p \to \mathbb{R}$ is local and private to node i and the nodes communicate over a balanced 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 $(i, j), i, j \in \mathcal{V}$, such that node j sends information to node i. Throughout the paper, we assume that each local f_i is smooth and non-convex. We focus on an *online*[1] setup where data samples are collected in real-time and hence each node i only has access to a noisy sample $\mathbf{g}_i$ of the true gradient at each iteration, such that $\mathbf{g}_i$ is an unbiased estimate of ∇f_i with bounded variance. Problems of this nature have found significant interest in signal processing, machine learning, and control. See e.g., [1], [2], for comprehensive surveys on these problems.

RX and SK are with the ECE Dept. at Carnegie Mellon University, {ranx, soummyak}@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 #1903972 and #1935555.

[1]We note that "online" sometimes also refers to time-varying objective functions, which is different from the problem setup in this paper.

Based on the classical stochastic gradient descent (**SGD**) [1], a well-known solution to Problem P1 is decentralized **SGD** (**DSGD**) [3], [4]. However, the convergence of **DSGD** for non-convex problems has only been established under certain regularity assumptions such as uniformly bounded difference between local and global gradients [5]–[7], or coercivity of each local function [8]. It has also been observed that if the data distributions across the nodes are heterogeneous, the practical performance of **DSGD** degrades significantly [2], [9], [10]. One notable line of work towards improving the performance of **DSGD** is EXTRA [11] and Exact Diffusion [12], where the convergence under the stochastic non-convex setting is established without the aforementioned regularity assumptions [13]; however, they require the weight matrix to be symmetric and the smallest eigenvalue is lower bounded by $-1/3$. Another family of algorithms to eliminate the performance limitation of **DSGD** is based on gradient tracking, introduced in [14], [15], where the basic idea is to replace the local gradients with a tracker of the global gradient ∇F. Decentralized first-order methods with gradient tracking have been well studied under exact gradients, where relevant work can be found, e.g., in [16]–[20]. However, the convergence behavior of gradient tracking methods has many unanswered questions when it comes to non-convex online stochastic problems [21], [22].

Main contributions. This paper considers **GT-DSGD** [9], that adds gradient tracking to **DSGD**, for online stochastic non-convex problems and rigorously develops novel results, key insights, and new analysis techniques that fill the theory gaps in the existing literature on gradient tracking methods [9], [21], [22]. The main contributions are described in the following:
(1) General smooth non-convex problems: We explicitly characterize the non-asymptotic, transient and steady-state performance of **GT-DSGD** and derive the conditions under which they are comparable to that of the centralized minibatch **SGD**. In particular, we show that its non-asymptotic mean-squared rate is network-independent and further matches the centralized minibatch **SGD** when the number of iterations is large enough. In sharp contrast, the existing results in [21], [22] suggest that the convergence rate and steady-state performance of **GT-DSGD** are always network-dependent and therefore are strictly worse than that of the centralized minibatch **SGD**; see Section III-A for details.
(2) Problems satisfying the global Polyak-Łojasiewicz (PL) condition: We analyze **GT-DSGD** when the global (smooth non-convex) function F further satisfies the PL condition. For both constant and decaying step-sizes, we explicitly characterize the non-asymptotic, transient and steady-state behaviors in expectation, and establish the conditions under which they

are comparable to that of the centralized minibatch **SGD**. We further establish global sublinear convergence rates on almost every sample path. The obtained sample path-wise rates are order-optimal (in the sense of polynomial time decay). To the best of our knowledge, these are the first results on path-wise convergence rate for online decentralized stochastic optimization under non-convexity, thus generalizing prior results in the decentralized stochastic approximation literature, e.g., [23], where the convergence analysis is mostly performed under assumptions of local convexity. As special cases, these results improve the current state-of-the-art on exact gradient methods under the PL condition [24] and stochastic strongly convex problems [9]; see Section III-B for details.

(3) Convergence analysis: We emphasize that the analysis techniques in this work are substantially different from the existing ones [9], [21], [22] and may be applied to other gradient methods built upon similar principles. We describe a few key features in the following. We establish tighter bounds on the stochastic gradient tracking process, by exploiting the unbiasedness of the online stochastic gradients, based on which all convergence theorems are derived; see Section V-B. To prove the convergence under general non-convexity, we characterize a descent inequality explicitly with network consensus errors and further show that the cumulative consensus errors along the algorithm path are dominated by the cumulative descent effect of the local gradients; see Section V-C. Towards the convergence analysis under the global PL condition, we derive the uniform boundedness of gradient tracking errors that is crucial in simplifying the ensuing analysis; see Lemma 18. Subsequently, we construct an appropriate stochastic process that forms an almost supermartingale [25] to prove sublinear rates on almost every sample path; see Section VII. To develop the convergence results in mean under the global PL condition, we use the analytical tools developed for recursive processes with time-varying step-sizes; see Section VIII.

Road map and notation. The rest of the paper is organized as follows. Section II describes the assumptions and the **GT-DSGD** algorithm. In Section III, we present the main results and discuss the contributions of this work in the context of the current state-of-the-art, whereas Section III-A and III-B respectively focus on the general non-convex and the PL case. We present detailed numerical experiments in Section IV to demonstrate the main theoretical results in this paper. Section V establishes general bounds on the stochastic gradient tracking process and proves the convergence for smooth non-convex functions. Sections VI, VII and VIII provide the convergence analysis under the PL condition on top of the results obtained in Section V. In particular, Sections VI and VIII focus on the convergence in mean with constant and decaying step-sizes respectively while Section VII focuses on the almost sure convergence. Section IX concludes the paper.

We use lowercase bold letters to denote vectors and uppercase bold letters for matrices. The matrix, $\mathbf{I}_d$ (resp. $\mathbf{O}_d$), represents the $d \times d$ identity (resp. zero matrix); $\mathbf{1}_d$ and $\mathbf{0}_d$ are the d-dimensional column vectors of all ones and zeros, respectively. We denote $[\mathbf{x}]_i$ as the i-th entry of a vector $\mathbf{x}$. The Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \otimes \mathbf{B}$. We use $\|\cdot\|$ to denote the Euclidean norm of a vector or the spectral norm of a matrix. For a matrix $\mathbf{X}$, we use $\rho(\mathbf{X})$ to denote its spectral radius, $\mathbf{X}^*$ to denote its adjugate, $\det(\mathbf{X})$ to denote its determinant, $[\mathbf{X}]_{i,j}$ to denote its (i,j)th element and $\mathrm{diag}(\mathbf{X})$ as the diagonal matrix that consists of the diagonal entries of $\mathbf{X}$. Matrix-vector inequalities are interpreted in the entry-wise sense. We use $\sigma(\cdot)$ to denote the σ-algebra generated by the random variables and/or sets in its argument.

II. Assumptions and the GT-DSGD Algorithm

We are interested in finding a first-order stationary point of Problem P1 via local computation and communication at each node. We first enlist the necessary assumptions that are standard in the literature [1], [9], [10], [26].

Assumption 1 (Objective functions). *Each f_i is L-smooth, i.e., $\exists L > 0$ s.t. $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Moreover, F is bounded below, i.e., $F^* := \inf_{\mathbf{x}} F(\mathbf{x}) > -\infty$.*

Assumption 2 (Network model). *The directed communication network is strongly-connected and admits a primitive doubly-stochastic weight matrix $\underline{\mathbf{W}} = \{\underline{w}_{ir}\} \in \mathbb{R}^{n\times n}$.*

We consider iterative processes that generate at each node i a sequence of state vectors $\{\mathbf{x}_k^i : k \geq 0\}$, where $\mathbf{x}_0^i$ is assumed to be a constant. At each iteration k, each node i is able to call the local oracle that returns a stochastic gradient $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$, where $\boldsymbol{\xi}_k^i$ is a random vector in $\mathbb{R}^q$ and $\mathbf{g}_i : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^p$ is a Borel-measurable function. For example, $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$ may be considered as the stochastic gradient evaluated at the state $\mathbf{x}_k^i$ with the data sample $\boldsymbol{\xi}_k^i$ observed at node i and iteration k. We work with a rich enough probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and define the natural filtration (an increasing family of sub-σ-algebras of $\mathcal{F}$) as, $\forall k \geq 1$,

$$\mathcal{F}_k := \sigma\left(\left\{\boldsymbol{\xi}_t^i : 0 \leq t \leq k-1, i \in \mathcal{V}\right\}\right), \qquad \mathcal{F}_0 := \{\Omega, \phi\},$$

where ϕ is the empty set. The intuitive meaning of $\mathcal{F}_k$ is that it contains the historical information of the algorithm iterates in question up to iteration $k-1$.

Assumption 3 (Oracle model). *The stochastic gradient process $\{\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i) : \forall k \geq 0, \forall i \in \mathcal{V}\}$ satisfies:*

- $\mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)|\mathcal{F}_k\right] = \nabla f_i(\mathbf{x}_k^i), \forall k \geq 0, \forall i \in \mathcal{V}$;
- $\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i) - \nabla f_i(\mathbf{x}_k^i)\right\|^2 |\mathcal{F}_k\right] \leq \nu_i^2, \forall k \geq 0, \forall i \in \mathcal{V}$, *for some constant $\nu_i > 0$;*
- *The family $\left\{\boldsymbol{\xi}_k^i : \forall k \geq 0, \forall i \in \mathcal{V}\right\}$ of random vectors is independent.*

We denote $\nu_a^2 := \frac{1}{n}\sum_{i=1}^n \nu_i^2$, the average of the variance of local stochastic gradients. We are also interested in the case when *the global objective function F* further satisfies the Polyak-Łojasiewicz (PL) condition that was introduced in [26].

Assumption 4. *$\exists \mu > 0$ s.t. the global function $F : \mathbb{R}^p \to \mathbb{R}$ satisfies $2\mu\left(F(\mathbf{x}) - F^*\right) \leq \|\nabla F(\mathbf{x})\|^2, \forall \mathbf{x} \in \mathbb{R}^p$.*

When Assumption 4 holds, we denote $\kappa := \frac{L}{\mu} \geq 1$, which can be interpreted as the condition number of F; see Lemma 12. Note that under the PL condition, every stationary point $\mathbf{x}^*$ of F is a global minimum of F, while F is not necessarily convex. Assumption 4 holds, e.g., in certain reinforcement learning problems [27], see [26], [28] for more details.

Algorithm. **GT-DSGD**, introduced in [9] for smooth strongly convex problems and formally described in Algorithm 1, recursively descends in the direction of an auxiliary variable $\mathbf{y}_k^i$ at each node, instead of the local stochastic gradient $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$. The auxiliary variable $\mathbf{y}_k^i$ is constructed under the dynamic average consensus principle [29] and tracks a time-varying signal $\sum_i \mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$, which mimics the global gradient; see [2], [9] for further intuition and explanation. We note that **GT-DSGD** uses the adapt-then-combine (ATC) structure [4] resulting in improved stability of the algorithm.

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

Require: $\mathbf{x}_0^i$; $\{\alpha_k\}$; $\{\underline{w}_{ir}\}$; $\mathbf{y}_i^0 = \mathbf{0}_p$; $\mathbf{g}_r(\mathbf{x}_{-1}^r, \boldsymbol{\xi}_{-1}^r) := \mathbf{0}_p$.

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

$$\mathbf{y}_{k+1}^i = \sum_{r=1}^n \underline{w}_{ir}\big(\mathbf{y}_k^r + \mathbf{g}_r(\mathbf{x}_k^r, \boldsymbol{\xi}_k^r) - \mathbf{g}_r(\mathbf{x}_{k-1}^r, \boldsymbol{\xi}_{k-1}^r)\big)$$

$$\mathbf{x}_{k+1}^i = \sum_{r=1}^n \underline{w}_{ir}\big(\mathbf{x}_k^r - \alpha_k \mathbf{y}_{k+1}^r\big)$$

2: **end for**

III. Main Results

In this section, we present our main convergence results for **GT-DSGD** and compare them with the corresponding state-of-the-art. For analysis purposes and the ease of presentation of main results, we let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{g}_k$, all in $\mathbb{R}^{np}$, respectively concatenate $\mathbf{x}_k^i$'s, $\mathbf{y}_k^i$'s, $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$'s, and write **GT-DSGD** in the following matrix form: $\forall k \geq 0$,

$$\mathbf{y}_{k+1} = \mathbf{W}\left(\mathbf{y}_k + \mathbf{g}_k - \mathbf{g}_{k-1}\right), \tag{1a}$$

$$\mathbf{x}_{k+1} = \mathbf{W}\left(\mathbf{x}_k - \alpha_k \mathbf{y}_{k+1}\right), \tag{1b}$$

where $\mathbf{W} = \underline{\mathbf{W}} \otimes \mathbf{I}_p$. We denote the exact averaging matrix as $\mathbf{J} := (\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes \mathbf{I}_p$ and $\lambda := \|\mathbf{W} - \mathbf{J}\|$, which characterizes the network connectivity. Under Assumption 2, we have $\lambda \in [0, 1)$; see [30]. For convenience, we let $\nabla \mathbf{f}_k \in \mathbb{R}^{np}$ concatenate all local exact gradients $\nabla f_i(\mathbf{x}_k^i)$'s and denote

$$\overline{\mathbf{x}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{x}_k, \quad \overline{\mathbf{y}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{y}_k,$$
$$\overline{\nabla \mathbf{f}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\nabla \mathbf{f}_k, \quad \overline{\mathbf{g}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{g}_k.$$

We assume without loss of generality that $\mathbf{x}_0^i = \mathbf{x}_0^r, \forall i, r \in \mathcal{V}$.

A. General smooth non-convex functions

In this subsection, we are concerned with the convergence of **GT-DSGD** for general smooth non-convex functions.

Theorem 1. *Let Assumptions 1, 2, and 3 hold and consider* ***GT-DSGD*** *under a constant step-size $\alpha_k = \alpha, \forall k \geq 0$, such that $0 < \alpha \leq \min\{1, \frac{1-\lambda^2}{12\lambda}, \frac{(1-\lambda^2)^2}{4\sqrt{6}\lambda^2}\}\frac{1}{2L}$, then, $\forall K > 1$,*

$$\underbrace{\frac{1}{n}\sum_{i=1}^n \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right]}_{\text{Mean-squared stationary gap}} \leq \underbrace{\frac{4(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha K} + \frac{2\alpha\nu_a^2 L}{n}}_{\text{Centralized minibatch SGD}}$$
$$+ \underbrace{\frac{448\alpha^2 L^2\lambda^2\nu_a^2}{(1-\lambda^2)^3} + \frac{64\alpha^2 L^2\lambda^4}{(1-\lambda^2)^3 K}\frac{\|\nabla \mathbf{f}_0\|^2}{n}}_{\text{Decentralized network effect}}.$$

Further, $\frac{1}{n}\sum_{i=1}^n \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\big[\|\nabla F(\mathbf{x}_k^i)\|^2\big]$ decays at the rate of $\mathcal{O}(\frac{1}{K})$ up to a steady-state error such that

$$\limsup_{K\to\infty} \frac{1}{n}\sum_{i=1}^n \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \leq \underbrace{\frac{2\alpha\nu_a^2 L}{n}}_{\text{Centralized minibatch SGD}} + \underbrace{\frac{448\alpha^2 L^2\lambda^2\nu_a^2}{(1-\lambda^2)^3}}_{\text{Decentralized network effect}}.$$

Theorem 1 is proved in Section V.

Remark 1 (Transient and steady-state performance). Theorem 1 explicitly characterizes the non-asymptotic performance of **GT-DSGD** for general smooth non-convex functions with an appropriate constant step-size. In particular, the stationary gap of **GT-DSGD** for any finite number of iterations K is bounded by the sum of four terms. The first two terms are independent of the network spectral gap $1 - \lambda$ and match the complexity of the centralized minibatch **SGD** up to constant factors [1]. The third and the fourth terms depend on $1 - \lambda$ reflecting the decentralized network and are in the order of $\mathcal{O}(\alpha^2)$. This is a much tighter characterization compared with the existing results [21], [22] on **GT-DSGD** and leads to provably faster non-asymptotic rate, see Remark 2 below. Theorem 1 also shows that as $K \to \infty$, the stationary gap of **GT-DSGD** decays sublinearly at the rate of $\mathcal{O}(1/K)$ up to a steady-state error. It can be observed that if $\alpha = \mathcal{O}(\frac{(1-\lambda)^3}{\lambda^2 nL})$, then the steady state stationary gap of **GT-DSGD** matches that of the centralized minibatch **SGD** up to constant factors. The existing analysis [22], however, suggests that under the same choice of the step-size α, the steady state stationary gap of **GT-DSGD** is strictly worse than the centralized minibatch **SGD**.

The following corollary of Theorem 1 is concerned with the non-asymptotic convergence rate of **GT-DSGD** over a finite time horizon for general smooth non-convex functions.

Corollary 1. *Let Assumptions 1, 2, and 3 hold and suppose that $\|\nabla \mathbf{f}_0\|^2 = \mathcal{O}(n)$. Setting $\alpha = \sqrt{n/K}$ in Theorem 1, for $K \geq 4nL^2 \max\left\{1, \frac{144\lambda^2}{(1-\lambda^2)^2}, \frac{96\lambda^4}{(1-\lambda^2)^4}\right\}$, we obtain:*

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \leq \underbrace{\frac{4(F(\overline{\mathbf{x}}_0) - F^*)}{\sqrt{nK}} + \frac{2\nu_a^2 L}{\sqrt{nK}}}_{\text{Centralized minibatch SGD}}$$
$$+ \underbrace{\frac{448n\lambda^2\nu_a^2 L^2}{(1-\lambda^2)^3 K} + \frac{64L^2\lambda^4\|\nabla \mathbf{f}_0\|^2}{(1-\lambda^2)^3 K^2}}_{\text{Decentralized network effect}}.$$

Thus, if K further satisfies that $K \geq K_{nc} := \mathcal{O}\left(\frac{n^3\lambda^4 L^2}{(1-\lambda)^6}\right)$, then we have

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] = \mathcal{O}\left(\frac{\nu_a^2 L}{\sqrt{nK}}\right).$$

Remark 2 (Non-asymptotic mean-squared rate and transient time for network independence). Corollary 1 shows that if the number of iterations is large enough, i.e., $K \geq K_{nc}$, by setting $\alpha = \frac{\sqrt{n}}{\sqrt{K}}$, the non-asymptotic rate of **GT-DSGD** matches that of the centralized minibatch **SGD** up to factors of

universal constants. This discussion shows that, in the regime that $K \geq K_{nc}$, **GT-DSGD** achieves a network-independent linear speedup compared with the centralized minibatch **SGD** that processes all data at a single node. In other words, the number of stochastic gradient computations required to achieve an approximate stationary point is reduced by a factor of $1/n$ at each node in the network. These results significantly improve the existing convergence guarantees of **GT-DSGD** for general smooth non-convex functions [21], [22]. In particular, references [21], [22] show that if $\alpha = \frac{c_0}{\sqrt{K}}$, where K is large enough and c_0 is some positive constant, **GT-DSGD** achieves the convergence rate of $\frac{c_1}{\sqrt{K}}$, where c_1 is a function of the network spectral gap $(1-\lambda)$. The convergence results in [21], [22] thus suggest that the rate of **GT-DSGD** is always network-dependent and is strictly worse than that of the centralized minibatch **SGD** and hence fail to characterize the network-independent performance of **GT-DSGD**.

Remark 3 (Comparison with DSGD). We observe from Corollary 1 that the convergence of **GT-DSGD** is robust to the difference between the local and the global functions. In other words, **GT-DSGD** outperforms **DSGD** when data distributions across the nodes are significantly heterogeneous, since the convergence rate of the latter explicitly depends on a factor that measures the heterogeneity between the local and the global functions [5]. However, the transient time for **GT-DSGD** to achieve network independent performance has a network dependence of $\mathcal{O}((1-\lambda)^{-6})$ which is worse than that of **DSGD** where the dependence is $\mathcal{O}((1-\lambda)^{-4})$. Moreover, we note that **GT-DSGD** requires two consecutive rounds of communication per node per iteration to update the state and the gradient tracker variables respectively, compared to **DSGD**.

B. Smooth non-convex functions under PL condition

In this subsection, we discuss the performance of **GT-DSGD** when the global objective function F further satisfies the PL condition. We begin with the case of constant step-size.

Theorem 2. *Let Assumption 1, 2, 3 and 4 hold. If the step-size $\alpha_k = \alpha, \forall k \geq 0$, satisfies that*

$$0 < \alpha \leq \overline{\alpha} := \min\left\{\frac{1}{2L}, \frac{(1-\lambda^2)^2}{42\lambda^2 L}, \frac{1-\lambda^2}{24\lambda L\kappa^{1/4}}, \frac{1-\lambda^2}{2\mu}\right\},$$

then $\mathbb{E}[\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2]$ and $\mathbb{E}[F(\overline{\mathbf{x}}_k) - F^*]$ decay linearly at the rate of $\mathcal{O}((1-\mu\alpha)^k)$ up to a steady-state error such that

$$\limsup_{k\to\infty} \mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] \leq \frac{288\lambda^4\alpha^5 L^3\kappa\nu_a^2}{n(1-\lambda^2)^4} + \frac{144\lambda^2\alpha^2\nu_a^2}{(1-\lambda^2)^3},$$

$$\limsup_{k\to\infty} \mathbb{E}\left[F(\overline{\mathbf{x}}_k) - F^*\right] \leq \frac{3\alpha\kappa\nu_a^2}{2n} + \frac{72\lambda^2\alpha^2\kappa L\nu_a^2}{(1-\lambda^2)^3}.$$

Moreover, $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{x}_k^i) - F^*\right]$ decays linearly at the rate of $\mathcal{O}((1-\mu\alpha)^k)$ up to a steady-state error such that

$$\limsup_{k\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{x}_k^i) - F^*\right] = \underbrace{\mathcal{O}\left(\frac{\alpha\kappa\nu_a^2}{n}\right)}_{\text{Centralized minibatch SGD}} + \underbrace{\mathcal{O}\left(\frac{\lambda^2\alpha^2\kappa L\nu_a^2}{(1-\lambda)^3}\right)}_{\text{Decentralized network effect}}.$$

Theorem 2 is proved in Section VI.

Remark 4 (Transient and steady-state performance). Theorem 2 shows that when the global objective function F satisfies the PL condition and the constant step-size α is less than $\overline{\alpha}$, the optimality gap of **GT-DSGD** decays linearly up to a steady-state error that is the sum of two terms. The first term is independent of the network and matches that of the centralized minibatch **SGD** up to constant factors, while the second term is due to the network and is controlled by $\mathcal{O}(\alpha^2)$. In contrast to [9], which requires a stronger assumption that the global objective function is strongly convex, we note that our stability range of the step-size α is larger by a factor of $\mathcal{O}(\kappa^{5/12})$; this relaxed upper bound on α further leads to a faster linear convergence when exact gradients are available, see Remark 5. Next, it can be verified from Theorem 2 that to match the steady-state error performance of the centralized minibatch **SGD** (up to constant factors), it suffices to choose the step-size α in **GT-DSGD** such that $\alpha = \mathcal{O}(\frac{(1-\lambda)^3}{\lambda^2 nL})$, which is larger by a factor of $\mathcal{O}(\kappa)$ than the corresponding result in [9]; in other words, Theorem 2 demonstrates a tighter and faster convergence rate to achieve the same steady-state error.

Remark 5 (Global linear convergence under exact gradient oracle). Theorem 2 further shows that when the exact gradient oracle is available at each node, i.e., $\nu_i^2 = 0, \forall i \in \mathcal{V}$, **GT-DSGD** reduces to its deterministic counterpart [14], [16], [17] and achieves global linear convergence to an optimal solution with an appropriate constant step-size. In other words, when $\alpha = \overline{\alpha}$, it achieves an q-accurate optimal solution in $\mathcal{O}\big(\max\{\kappa, \frac{\lambda^2\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{1}{1-\lambda}\}\log\frac{1}{q}\big)$ iterations. This result improves upon the state-of-the-art gradient computation and communication complexity under the PL condition [24]. The gradient computation complexity can be further improved to $\mathcal{O}\big(\kappa\log\frac{1}{\epsilon}\big)$ by performing $\mathcal{O}\big(\frac{1}{1-\lambda}\log\frac{\kappa}{1-\lambda}\big)$ rounds of consensus communication at each iteration. This gradient computation complexity result matches the state-of-the-art [31] on decentralized exact gradient methods (without Nesterov acceleration), which further requires a stronger assumption that each local function is convex and the global function is strongly convex. In contrast, we only require the PL condition on the global objective F.

We now proceed to the case of decaying step-sizes. The next result shows the sample path-wise performance of **GT-DSGD** under a family of stochastic approximation step-sizes [32], i.e., $\alpha_k > 0$, $\sum_{k=0}^\infty \alpha_k = \infty$, and $\sum_{k=0}^\infty \alpha_k^2 < \infty$, which enables the exact sublinear convergence in contrast to the inexact linear convergence under a constant step-size.

Theorem 3. *Let Assumptions 1, 2, 3, and 4 hold. Consider the step-size sequence $\{\alpha_k\}$ such that $\alpha_k = \delta(k+\varphi)^{-\epsilon}, \forall k \geq 0$, where $\epsilon \in (0.5, 1]$, $\delta \geq 1/\mu$, and $\varphi \geq \max\{(\delta/\overline{\alpha})^{1/\epsilon}, \frac{4}{1-\lambda^2}\}$ for $\overline{\alpha}$ given in Theorem 2. Then $\forall i, j \in \mathcal{V}$ and for arbitrarily small $\epsilon_1 > 0$, we have:*

$$\mathbb{P}\left(\sum_{k=0}^\infty k^{2\epsilon-1-\epsilon_1}\left\|\mathbf{x}_k^i - \mathbf{x}_k^j\right\|^2 < \infty\right) = 1,$$

$$\mathbb{P}\Big(\lim_{k\to\infty} k^{2\epsilon-1-\epsilon_1}\left(F(\mathbf{x}_k^i) - F^*\right) = 0\Big) = 1.$$

Theorem 3 is proved in Section VII.

Remark 6 (Global sublinear rate on almost every sample path). Theorem 3 guarantees that **GT-DSGD** exhibits a global sublinear convergence on almost every sample path, under decaying step-sizes, when the global function F satisfies the PL condition. This result is of significant practical value in that it is applicable to every instantiation of the algorithm while the expectation type convergence only characterizes, roughly speaking, the performance on average. Furthermore, in the case of general non-degenerate variances (see Assumption 3), these path-wise rates are order-optimal, in the sense of polynomial time decay; this follows by considering the stochastic approximation reformulation of the optimization problem (i.e., the problem of obtaining zeros of the gradient function $\nabla F(\mathbf{x})$) and invoking standard central limit type arguments, see [32].) To the best of our knowledge, Theorem 3 is the first to show path-wise convergence for online decentralized stochastic optimization under non-convexity, thus generalizing prior results in the decentralized stochastic approximation and optimization literature, such as [23], where such analysis is performed under assumptions of local convexity.

Finally, we consider the convergence rate of **GT-DSGD** in expectation when $\alpha_k = \mathcal{O}(1/k), \forall k \geq 0$.

Theorem 4. *Let Assumptions 1, 2, 3, and 4 hold. Consider the step-size sequence $\{\alpha_k\}$ such that $\alpha_k = \beta(k+\gamma)^{-1}, \forall k \geq 0$, where $\beta > 2/\mu$, and $\gamma \geq \max\left\{\frac{\beta}{\overline{\alpha}}, \frac{8}{1-\lambda^2}\right\}$ for $\overline{\alpha}$ given in Theorem 2. We have: $\forall k \geq 0$,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[F(\mathbf{x}_k^i)-F^*\right] \leq \underbrace{\frac{2L\nu_a^2\beta^2}{n(\mu\beta-1)(k+\gamma)}}_{\text{Centralized minibatch SGD}} + \underbrace{\frac{2\left(F(\overline{\mathbf{x}}_0)-F^*\right)}{(k/\gamma+1)^{\mu\beta}} + \frac{3L^2\widehat{x}\beta^3}{n(\mu\beta-2)(k+\gamma)^2}}_{\text{Decentralized network effect}},$$

where $\widehat{x}$ is a positive constant given in (62).

The non-asymptotic rate in Theorem 4 shows that **GT-DSGD** asymptotically achieves network independent $\mathcal{O}(1/k)$ rate in mean when the global objective function F satisfies the PL condition, matching the $\Omega(1/k)$ oracle lower bound [1]. The following corollary examines the number of transient iterations required to achieve network-independence under specific choices of parameter β and γ in Theorem 4.

Corollary 2. *Let Assumptions 1, 2, 3, and 4 hold. Set $\beta = 6/\mu$ and $\gamma = \max\left\{\frac{6}{\mu\overline{\alpha}}, \frac{8}{1-\lambda^2}\right\}$ in Theorem 4 and suppose that $\|\nabla \mathbf{f}_0\|^2 = \mathcal{O}(n)$. Then we have:*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[F(\mathbf{x}_k^i)-F^*\right] = \mathcal{O}\left(\frac{\kappa^2\left(F(\overline{\mathbf{x}}_0)-F^*\right)}{k^2} + \frac{\kappa\nu_a^2}{n\mu k}\right),$$

if k is large enough such that $k \gtrsim K_{PL}$, where

$$K_{PL} := \frac{\lambda^2 n\kappa}{(1-\lambda)^3} + \frac{\lambda\kappa^{5/4}}{1-\lambda} + \kappa + \frac{\lambda^{3/2}\kappa^{11/8}}{(1-\lambda)^{3/2}} + \frac{\kappa^{-1/2}}{(1-\lambda)^{3/2}} + \frac{\lambda^2 n\kappa^{1/2}L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda)^2\nu_a^2}.$$

Theorem 4 and Corollary 2 are proved in Section VIII.

Remark 7 (Transient time for network independent rate). Corollary 2 shows after K_{PL} iterations, the convergence rate of **GT-DSGD** matches that of the centralized minibatch **SGD** [1] up to constant factors and therefore achieves an asymptotic linear speedup. We now compare this transient time with the existing literature. First, Ref. [9] shows that, under the strong convexity of F, **GT-DSGD** asymptotically converges at $\mathcal{O}(1/k)$; however, the convergence rate derived in [9] depends on arbitrary constants and therefore the transient time is not clear. Second, recent work [33], [34] shows that when each local function f_i is strongly convex, the corresponding transient time of **DSGD** is $\mathcal{O}\big(n\kappa^6(1-\lambda)^{-2}\big)$. Our results on the transient time K_{PL} therefore significantly improve upon the dependence of the condition number κ under weaker assumptions on the objective functions, while being moderately worse in terms of the network dependence, i.e. $1-\lambda$.

IV. Numerical Experiments

In this section, we present numerical experiments to demonstrate the main theoretical results in Section III with the help of learning problems on real-world datasets, summarized in Table I, and minimizing certain synthetic functions to illustrate the PL condition. We consider three different graph topologies, i.e., a directed exponential graph with 16 nodes, an undirected grid graph with 16 nodes, and an undirected geometric graph with 100 nodes; see Fig. 1. The primitive doubly stochastic weights are set to be equal for the exponential graph and are generated by the Metroplis rule [35] for the grid and the geometric graphs. The second largest singular values λ associated with the weight matrices of these graphs are $0.6, 0.93$ and 0.99, respectively. Towards the stochastic gradient oracle, we consider two different setups: (i) each node has access to a finite collection of data samples and the stochastic gradient is computed with respect to one randomly selected data sample at each iteration; (ii) each node has access to the gradient of its local function subject to random noise, with zero-mean and bounded variance, at each iteration. The performance metric of interest is the average of global function values across the nodes $\frac{1}{n}\sum_{i=1}^{n}F(\mathbf{x}_k^i)$, which we refer to as *loss*, versus the number of epochs[2] in (i) and the number of iterations in (ii). We manually optimize the parameters of all algorithms across all experiments to achieve their best performances.

TABLE I
A SUMMARY OF THE DATASETS USED IN NUMERICAL EXPERIMENTS, AVAILABLE AT HTTPS://WWW.OPENML.ORG/.

Dataset	train	dimension	classes
a9a	48,832	124	2
w8a	60,000	301	2
creditcard	100,000	30	2
Fashion-MNIST	60,000	785	10
CIFAR-10	50,000	3073	10
STL-10	5,000	27649	10

To study the convergence behavior of **GT-DSGD**, we conduct three different experiments: binary classification with non-convex logistic regression [36], multiclass classification

[2]Each epoch is one effective pass of local data samples at each node.

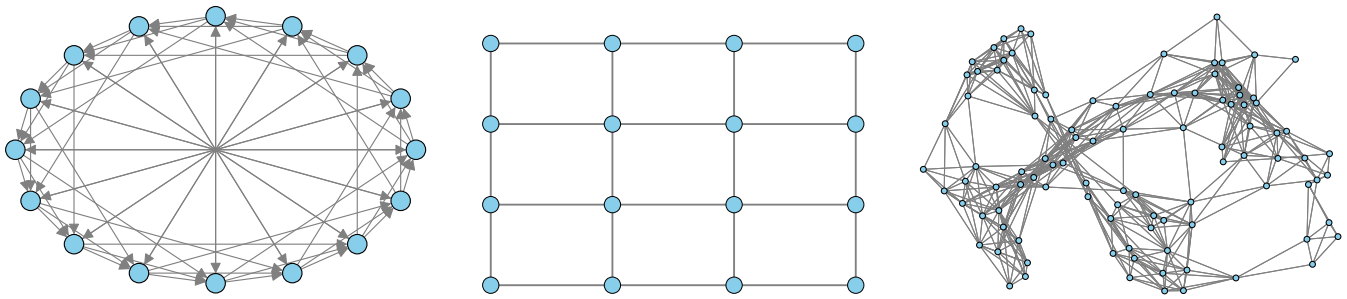


Fig. 1. A directed exponential graph with 16 nodes, an undirected grid graph with 16 nodes, and an undirected geometric graph with 100 nodes.

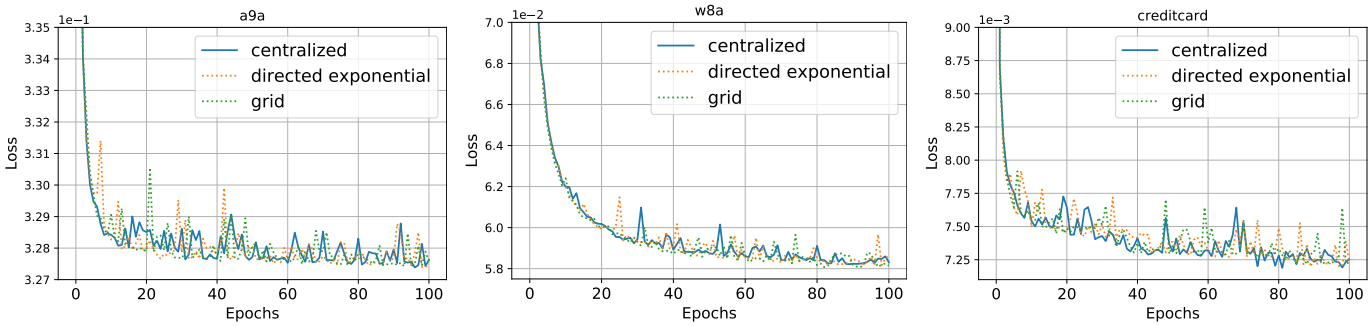


Fig. 2. The performance of **GT−DSGD** for non-convex logistic regression over different graphs and comparison with the centralized minibatch **SGD** on the a9a, w8a and creditcard datasets.

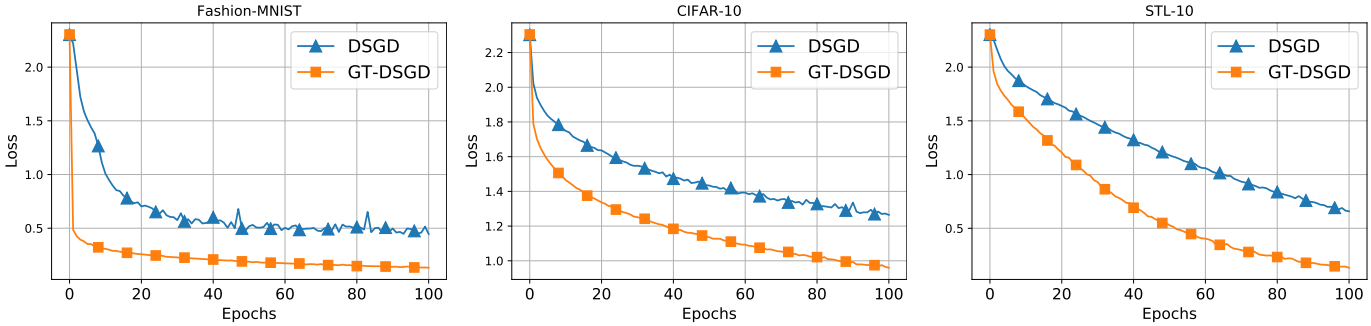


Fig. 3. Performance comparison between **GT−DSGD** and **DSGD** for one-hidden-layer neural network under heterogeneous data distributions across the nodes on the Fashion-MNIST, CIFAR-10 and STL-10 datasets.

with neural networks, and minimizing synthetic non-convex functions that satisfy the global PL condition. We compare the performance of **GT−DSGD** with **DSGD** [5] to illustrate the advantages of the former in the setting of heterogeneous data distributions across the nodes; moreover, we use the centralized minibatch **SGD** as the benchmark to illustrate the scenarios in which **GT−DSGD** achieves a network-independent performance. The experimental results are described in the next subsections. It can be verified that the numerical results of **GT−DSGD** are consistent with the theory in this paper.

A. Non-convex logistic regression for binary classification

We first consider a binary classification problem with the help of a non-convex logistic regression model [36]. Specifically, the decentralized optimization problem of interest is given by $\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}) + r(\mathbf{x})$, such that

$$f_i(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^m \log\left[1 + e^{-(\mathbf{x}^\top \boldsymbol{\theta}_{ij})\xi_{ij}}\right], \; r(\mathbf{x}) = \sum_{d=1}^p \frac{R[\mathbf{x}]_d^2}{1+[\mathbf{x}]_d^2},$$

where $\boldsymbol{\theta}_{i,j}$ is the feature vector, $\xi_{i,j}$ is the corresponding binary label, and $r(\mathbf{x})$ is a non-convex regularizer with $R = 10^{-4}$.

We compare the performance of **GT−DSGD** over the directed exponential and the grid graphs, both with 16 nodes, to the centralized **SGD** with a minibatch size of 16. We consider the best possible constant step-size for both algorithms. The numerical results over the a9a, w8a, and creditcard datasets are shown in Fig. 2. It can be observed that, across all datasets, the convergence behavior of **GT−DSGD** matches that of the centralized minibatch **SGD** and is independent of the underlying graph topology, as long as the total number of iterations is large enough. This observation is consistent with Corollary 1, demonstrating the network-independent convergence of **GT−DSGD** under an appropriate constant step-size for general smooth non-convex functions.

B. Neural network for multiclass classification

We next compare the performance of **DSGD** (without gradient tracking) and **GT−DSGD**, both with a constant step-size, when the data distributions across the nodes are significantly

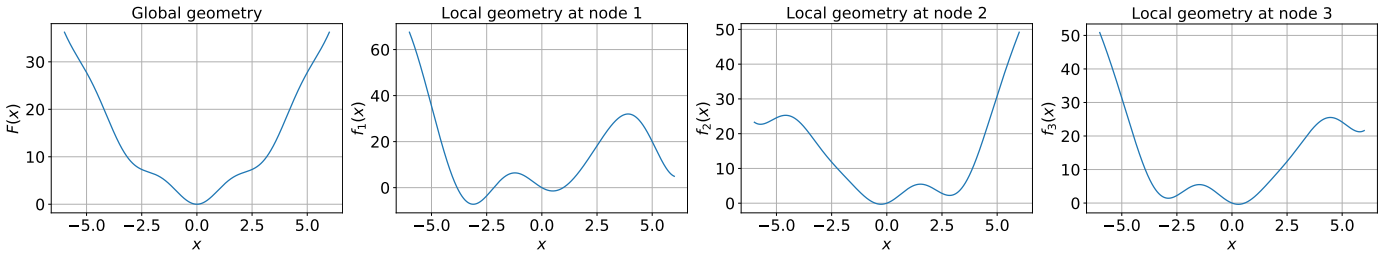
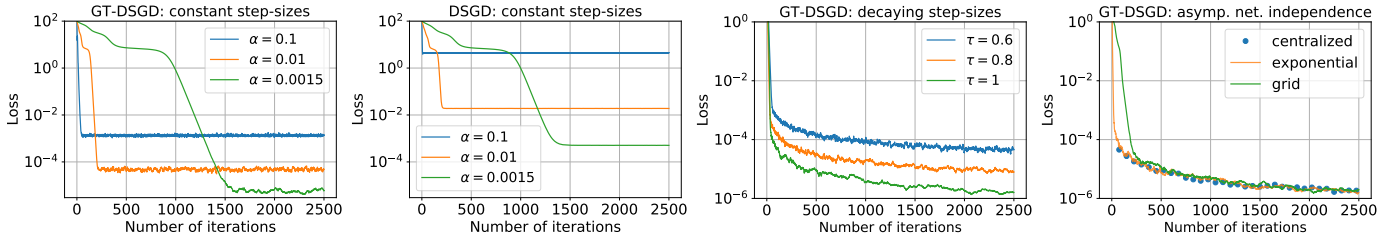


Fig. 4. The global and local geometries in the experiment with synthetic functions that satisfy the global PL condition.



(a) (b) (c) (d)

Fig. 5. Convergence of **GT-DSGD** and **DSGD** under the global PL condition: (a)(b) Inexact linear convergence with different constant step-sizes α. (c) Exact sublinear convergence of **GT-DSGD** with decaying step-sizes $\alpha_k = (k+3)^{-\tau}$ under different values of τ. (d) Exact sublinear convergence of **GT-DSGD** over different graphs in comparison with the centralized minibatch **SGD**, all with the decaying step-size $\alpha_k = (k+3)^{-1}$.

heterogeneous. To this aim, we consider a harsh problem setup where the data samples are distributed over the 100-node geometric graph in Fig. 1 such that each node has the same number of data samples and the samples belong to only one or two classes (out of 10 possible classes). We consider decentralized training of a neural network with one fully connected hidden layer of 64 neurons and sigmoid activation. The experimental results over the Fashion-MNIST, CIFAR-10, and STL-10 datasets are shown in Fig. 3. We observe that **GT-DSGD** significantly outperforms **DSGD** in this setting, demonstrating the robustness of **GT-DSGD** to heterogeneous data across the nodes; see also Remark 3.

C. Synthetic functions that satisfy the global PL condition

Finally, we show the performance of **GT-DSGD** when the global function satisfies the PL condition and compare it with **DSGD** and the centralized minibatch **SGD**. In particular, each local function is chosen as $f_i(x) = x^2 + 3\sin^2(x) + a_i x\cos(x)$, such that $\sum_{i=1}^n a_i = 0$ and $a_i \neq 0, \forall i \in \mathcal{V}$, leading to the global function $F(x) = x^2 + 3\sin^2(x)$, which is clearly non-convex and further satisfies the PL condition [28]. It can be verified that each local function is highly nonlinear and significantly different from the global function; see Fig. 4. We inject random Gaussian noise with mean 0 and the standard deviation 0.5 to the gradient computation at each node. The corresponding numerical results can be found in Fig. 5, where the experiments in Fig. 5(a)-(c) are performed over the directed exponential graph with 16 nodes. It can be observed from Fig. 5(a) that **GT-DSGD** achieves inexact linear convergence under constant step-sizes; moreover, a smaller step-size leads to a smaller steady-state error but at a slower rate. Compared with the convergence of **DSGD** under constant step-sizes shown in Fig. 5(b), **GT-DSGD** achieves a smaller steady-state error much faster benefiting from gradient tracking that effectively exploits the global geometry. Fig. 5(c) shows that **GT-DSGD** achieves exact sublinear convergence to the optimal solution with decaying step-sizes of the form $\alpha_k = (k+3)^{-\tau}$ under different values of τ chosen in $(0.5, 1]$. Clearly, a larger τ leads to a faster rate as Theorem 3 suggests. Finally, we observe from Fig. 5(d) that the convergence rate of **GT-DSGD** with $\tau = 1$ matches that of the centralized minibatch **SGD** with the same decaying step-size after a small number of transient iterations over different graphs. This phenomenon demonstrates the asymptotically network-independent and optimal $\mathcal{O}(1/k)$ rate achieved by **GT-DSGD**. This observation is consistent with Theorem 4.

V. Convergence Analysis: The General Non-Convex Case

It is straightforward to verify that the random variables generated by **GT-DSGD** are square-integrable and that $\mathbf{x}_k, \mathbf{y}_k$ are $\mathcal{F}_k$-measurable and $\mathbf{g}(\mathbf{x}_k, \boldsymbol{\xi}_k)$ is $\mathcal{F}_{k+1}$-measurable, $\forall k$. In this section, we derive general bounds on the stochastic gradient tracking process, which may be of independent interest, and prove Theorem 1. We start by presenting some standard results on decentralized stochastic gradient tracking algorithms; their proofs can be found, e.g., in [9], [16], [37].

Lemma 1. *Under Assumption 1-3, We have the following:*

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

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

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

(d) $\mathbb{E}[\langle \mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i) - \nabla f_i(\mathbf{x}_k^i), \mathbf{g}_r(\mathbf{x}_k^r, \boldsymbol{\xi}_k^r) - \nabla f_r(\mathbf{x}_k^r)\rangle | \mathcal{F}_k] = 0$, $\forall k \geq 0, \forall i, r \in \mathcal{V}$ *such that* $i \neq r$.

(e) $\mathbb{E}\left[\|\overline{\mathbf{g}}_k - \overline{\nabla \mathbf{f}}_k\|^2 | \mathcal{F}_k\right] \leq \nu_a^2/n, \forall k \geq 0$.

As a consequence of the state update of **GT-DSGD** described in (1b) and Lemma 1(b), we have: $\forall k \geq 0$,

$$\overline{\mathbf{x}}_{k+1} = \overline{\mathbf{x}}_k - \alpha_k \overline{\mathbf{y}}_{k+1} = \overline{\mathbf{x}}_k - \alpha_k \overline{\mathbf{g}}_k, \tag{2}$$

i.e., the mean state $\overline{\mathbf{x}}_k$ of the network proceeds in the direction of the average of local stochastic gradients $\overline{\mathbf{g}}_k$. The following lemma provides several useful relations on the consensus process of the state vectors across the network [37].

Lemma 2. *Let Assumption 2 hold. We have the following inequalities:* $\forall k \geq 0$,

$$\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\|^2 \leq \frac{1+\lambda^2}{2}\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + \frac{2\alpha_k^2\lambda^2}{1-\lambda^2}\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2.$$

$$\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\|^2 \leq 2\lambda^2\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + 2\alpha_k^2\lambda^2\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2.$$

$$\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\| \leq \lambda\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + \alpha_k\lambda\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|.$$

A. A descent inequality

In this subsection, we establish a key descent inequality that characterizes the expected decrease of the value of the global objective function F over each iteration in light of (2).

Lemma 3. *Let Assumptions 1-3 hold. If* $0 < \alpha_k \leq \frac{1}{2L}$, *then we have:* $\forall k \geq 0$,

$$\mathbb{E}\left[F(\overline{\mathbf{x}}_{k+1})|\mathcal{F}_k\right] \leq F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 - \frac{\alpha_k}{4}\left\|\overline{\nabla\mathbf{f}}_k\right\|^2 + \frac{\alpha_k L^2}{2}\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n} + \frac{\alpha_k^2 L\nu_a^2}{2n}.$$

Proof. Since F is L-smooth, we have [26]: $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle\nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{3}$$

Setting $\mathbf{y} = \overline{\mathbf{x}}_{k+1}$ and $\mathbf{x} = \overline{\mathbf{x}}_k$ in (3) to obtain: $\forall k \geq 0$,

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

Conditioning on $\mathcal{F}_k$, by $\mathbb{E}[\overline{\mathbf{g}}_k|\mathcal{F}_k] = \overline{\nabla\mathbf{f}}_k$, obtains: $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}[F(\overline{\mathbf{x}}_{k+1})|\mathcal{F}_k]\\
&\leq F(\overline{\mathbf{x}}_k) - \alpha_k\left\langle\nabla F(\overline{\mathbf{x}}_k), \overline{\nabla\mathbf{f}}_k\right\rangle + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right]\\
&= F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 - \frac{\alpha_k}{2}\left\|\overline{\nabla\mathbf{f}}_k\right\|^2\\
&\quad + \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k) - \nabla\mathbf{f}_k\|^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right]\\
&\leq F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 - \frac{\alpha_k}{2}\|\overline{\nabla\mathbf{f}}_k\|^2\\
&\quad + \frac{\alpha_k L^2}{2n}\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right],
\end{aligned} \tag{4}$$

where the equality above uses $\langle\mathbf{x}, \mathbf{y}\rangle = \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2), \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, and the last inequality is due to Lemma 1(c). For the last term in (4), note that: $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right] &= \mathbb{E}\left[\left\|\overline{\mathbf{g}}_k - \overline{\nabla\mathbf{f}}_k + \overline{\nabla\mathbf{f}}_k\right\|^2|\mathcal{F}_k\right]\\
&= \mathbb{E}\left[\left\|\overline{\mathbf{g}}_k - \overline{\nabla\mathbf{f}}_k\right\|^2|\mathcal{F}_k\right] + \|\overline{\nabla\mathbf{f}}_k\|^2\\
&\leq \nu_a^2/n + \left\|\overline{\nabla\mathbf{f}}_k\right\|^2,
\end{aligned} \tag{5}$$

where the second equality uses that $\overline{\nabla\mathbf{f}}_k$ is $\mathcal{F}_k$-measurable and $\mathbb{E}[\overline{\mathbf{g}}_k|\mathcal{F}_k] = \overline{\nabla\mathbf{f}}_k$, and the last inequality uses Lemma 1(e). We now use (5) in (4) to obtain: $\forall k \geq 0$,

$$\mathbb{E}[F(\overline{\mathbf{x}}_{k+1})|\mathcal{F}_k] \leq F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 + \frac{\alpha_k^2 L\nu_a^2}{2n} - \frac{\alpha_k(1-\alpha_k L)}{2}\|\overline{\nabla\mathbf{f}}_k\|^2 + \frac{\alpha_k L^2}{2n}\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2.$$

The proof follows by noting that $1 - \alpha_k L \geq \frac{1}{2}$, if $0 < \alpha_k \leq \frac{1}{2L}$, $\forall k \geq 0$, in the inequality above. □

Compared with the corresponding descent inequality for the centralized stochastic gradient descent, see, e.g., [1], [26], the descent inequality for **GT-DSGD** derived in Lemma 3 has an additional network consensus error term $\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|$. We therefore seeks for means to control this perturbation in order to establish the convergence of **GT-DSGD**. We will bound the consensus and the gradient tracking error jointly.

B. Bounding the gradient tracking error

In this subsection, we analyze the gradient tracking process.

Lemma 4. *Let Assumption 1-3 hold. We have:* $\forall k \geq 0$,

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

Proof. Using the gradient tracking update (1a), and the fact that $\mathbf{W}\mathbf{J} = \mathbf{J}\mathbf{W} = \mathbf{J}$, we have: $\forall k \geq 0$,

$$\begin{aligned}
&\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\|^2\\
&= \|\mathbf{W}(\mathbf{y}_{k+1} + \mathbf{g}_{k+1} - \mathbf{g}_k) - \mathbf{J}(\mathbf{y}_{k+1} + \mathbf{g}_{k+1} - \mathbf{g}_k)\|^2\\
&= \|\mathbf{W}\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1} + (\mathbf{W} - \mathbf{J})(\mathbf{g}_{k+1} - \mathbf{g}_k)\|^2\\
&= \|\mathbf{W}\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2 + \|(\mathbf{W} - \mathbf{J})(\mathbf{g}_{k+1} - \mathbf{g}_k)\|^2\\
&\quad + 2\langle(\mathbf{W} - \mathbf{J})\mathbf{y}_{k+1}, \ (\mathbf{W} - \mathbf{J})(\mathbf{g}_{k+1} - \mathbf{g}_k)\rangle\\
&\leq \lambda^2\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2 + \lambda^2\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2\\
&\quad + 2\underbrace{\langle(\mathbf{W} - \mathbf{J})\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})(\mathbf{g}_{k+1} - \mathbf{g}_k)\rangle}_{C_1},
\end{aligned} \tag{6}$$

where the last inequality is due to Lemma 1(a). Towards C_1, since $\mathbf{y}_{k+1}$ and $\mathbf{g}_k$ are $\mathcal{F}_{k+1}$-measurable, we have: $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}[C_1|\mathcal{F}_{k+1}]\\
&= \langle(\mathbf{W} - \mathbf{J})\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})(\nabla\mathbf{f}_{k+1} - \mathbf{g}_k)\rangle\\
&= \langle(\mathbf{W} - \mathbf{J})\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})(\nabla\mathbf{f}_k - \mathbf{g}_k)\rangle\\
&\quad + \langle(\mathbf{W} - \mathbf{J})\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})(\nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k)\rangle.
\end{aligned} \tag{7}$$

The proof then follows by taking the expectation on (6) and using (7) in the resulting inequality. □

Next, we bound the terms in Lemma 4 respectively. For the second term in Lemma 4, we have the following.

Lemma 5. *Let Assumption 1-3 hold. We have:* $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2\right]\\
&\leq 18L^2\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2\right] + 6n\alpha_k^2 L^2\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2\right]\\
&\quad + 12\alpha_k^2 L^2\lambda^2\mathbb{E}\left[\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2\right] + 3n\nu_a^2.
\end{aligned}$$

Proof. Since both $\nabla \mathbf{f}_{k+1}$ and $\mathbf{g}_k$ are $\mathcal{F}_{k+1}$-measurable and $\mathbb{E}[\mathbf{g}_{k+1}|\mathcal{F}_{k+1}] = \nabla \mathbf{f}_{k+1}$, we have: $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{g}_{k+1} - \mathbf{g}_k\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{g}_{k+1} - \nabla \mathbf{f}_{k+1}\right\|^2\right] + \mathbb{E}\left[\left\|\nabla \mathbf{f}_{k+1} - \mathbf{g}_k\right\|^2\right], \\
&\leq n\nu_a^2 + \mathbb{E}\left[\left\|\nabla \mathbf{f}_{k+1} - \mathbf{g}_k\right\|^2\right] \\
&\leq n\nu_a^2 + 2\mathbb{E}\left[\left\|\nabla \mathbf{f}_{k+1} - \nabla \mathbf{f}_k\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla \mathbf{f}_k - \mathbf{g}_k\right\|^2\right] \\
&\leq 3n\nu_a^2 + 2L^2 \underbrace{\mathbb{E}\left[\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2\right]}_{C_2}
\end{aligned} \tag{8}$$

where the first inequality uses Assumption 3 and the last inequality uses Assumption 3 and the L-smoothness of each f_i. Towards C_2, we have: $\forall k \geq 0$,

$$\begin{aligned}
C_2 &= \mathbb{E}\left[\left\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1} + \mathbf{J}\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_k + \mathbf{J}\mathbf{x}_k - \mathbf{x}_k\right\|^2\right] \\
&\leq 3\mathbb{E}\left[\left\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\right\|^2\right] + 3n\alpha_k^2\mathbb{E}\left[\left\|\overline{\mathbf{g}}_k\right\|^2\right] \\
&\quad + 3\mathbb{E}\left[\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2\right] \\
&\leq 9\mathbb{E}\left[\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2\right] + 3n\alpha_k^2\mathbb{E}\left[\left\|\overline{\mathbf{g}}_k\right\|^2\right] \\
&\quad + 6\alpha_k^2\lambda^2\mathbb{E}\left[\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2\right],
\end{aligned} \tag{9}$$

where the second inequality uses (2) and the last inequality uses Lemma 2. The proof follows by using (9) in (8). □

For the third term in Lemma 4, we have the following.

Lemma 6. *Let Assumption 1-3 hold. We have:* $\forall k \geq 0$,

$$\mathbb{E}\left[\left\langle (\mathbf{W} - \mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_k - \mathbf{g}_k\right)\right\rangle\right] \leq \nu_a^2.$$

Proof. Using the fact that $\mathbf{J}(\mathbf{W} - \mathbf{J}) = \mathbf{O}_{np}$ and the gradient tracking update (1a), we have: $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle (\mathbf{W} - \mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_k - \mathbf{g}_k\right)\right\rangle | \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{W}\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_k - \mathbf{g}_k\right)\right\rangle | \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{W}^2\left(\mathbf{y}_k + \mathbf{g}_k - \mathbf{g}_{k-1}\right), (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_k - \mathbf{g}_k\right)\right\rangle | \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{W}^2\mathbf{g}_k, (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_k - \mathbf{g}_k\right)\right\rangle | \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{W}^2\left(\mathbf{g}_k - \nabla \mathbf{f}_k\right), (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_k - \mathbf{g}_k\right)\right\rangle | \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left(\mathbf{g}_k - \nabla \mathbf{f}_k\right)^\top (\mathbf{J} - \mathbf{W}^\top\mathbf{W}^2)\left(\mathbf{g}_k - \nabla \mathbf{f}_k\right) | \mathcal{F}_k\right],
\end{aligned} \tag{10}$$

where the third and the fourth equality exploit the fact that the random vectors $\mathbf{y}_k$, $\mathbf{g}_{k-1}$ and $\nabla \mathbf{f}_k$ are $\mathcal{F}_k$-measurable and that $\mathbb{E}[\mathbf{g}_k|\mathcal{F}_k] = \nabla \mathbf{f}_k$. In light of Lemma 1(d), (10) reduces to

$$\begin{aligned}
&\mathbb{E}\left[\left\langle (\mathbf{W} - \mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_k - \mathbf{g}_k\right)\right\rangle | \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left(\mathbf{g}_k - \nabla \mathbf{f}_k\right)^\top \mathrm{diag}(\mathbf{J} - \mathbf{W}^\top\mathbf{W}^2)\left(\mathbf{g}_k - \nabla \mathbf{f}_k\right) | \mathcal{F}_k\right] \\
&\leq \mathbb{E}\left[\left(\mathbf{g}_k - \nabla \mathbf{f}_k\right)^\top \mathrm{diag}(\mathbf{J})\left(\mathbf{g}_k - \nabla \mathbf{f}_k\right) | \mathcal{F}_k\right], \\
&= \mathbb{E}\left[\left\|\mathbf{g}_k - \nabla \mathbf{f}_k\right\|^2 | \mathcal{F}_k\right]/n
\end{aligned} \tag{11}$$

where the inequality holds since $\mathrm{diag}(\mathbf{W}^\top\mathbf{W}^2)$ is nonnegative. The proof follows by using Assumption 3 in (11) and taking the expectation on the resulting inequality. □

For the last term in Lemma 4, we have the following.

Lemma 7. *Let Assumption 1-3 hold. We have:* $\forall k \geq 0$,

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

where η_1 and η_2 are arbitrary positive constants[3].

Proof. Using $(\mathbf{W} - \mathbf{J})\mathbf{J} = \mathbf{O}_{np}$ and the Cauchy-Schwarz inequality, we have: $\forall k \geq 0$,

$$\begin{aligned}
&\left\langle (\mathbf{W} - \mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_{k+1} - \nabla \mathbf{f}_k\right)\right\rangle \\
&= \left\langle (\mathbf{W} - \mathbf{J})\left(\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right), (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_{k+1} - \nabla \mathbf{f}_k\right)\right\rangle \\
&\leq \lambda^2 L\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|,
\end{aligned} \tag{12}$$

where the last inequality uses $\|\mathbf{W} - \mathbf{J}\| = \lambda$ and the L-smoothness of each f_i. We note that, $\forall k \geq 0$,

$$\begin{aligned}
&\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| \\
&= \left\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1} + \mathbf{J}\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_k + \mathbf{J}\mathbf{x}_k - \mathbf{x}_k\right\| \\
&\leq \left\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\right\| + \alpha_k\sqrt{n}\left\|\overline{\mathbf{g}}_k\right\| + \left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\| \\
&\leq 2\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\| + \alpha_k\sqrt{n}\left\|\overline{\mathbf{g}}_k\right\| + \alpha_k\lambda\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|.
\end{aligned} \tag{13}$$

where the last inequality uses Lemma 2. We use (13) in (12) to obtain: $\forall k \geq 0$,

$$\begin{aligned}
&\left\langle (\mathbf{W} - \mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\left(\nabla \mathbf{f}_{k+1} - \nabla \mathbf{f}_k\right)\right\rangle \\
&\leq \lambda^3\alpha_k L\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2 \\
&\quad + \underbrace{\left(\lambda\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|\right)\left(\lambda\alpha_k L\sqrt{n}\left\|\overline{\mathbf{g}}_k\right\|\right)}_{C_3} \\
&\quad + \underbrace{2(\lambda\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|)(\lambda L\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|)}_{C_4}.
\end{aligned} \tag{14}$$

By Young's inequality, we have that

$$C_3 \leq 0.5\eta_1\lambda^2\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2 + 0.5\eta_1^{-1}\lambda^2\alpha_k^2 L^2 n\left\|\overline{\mathbf{g}}_k\right\|^2,$$

where $\eta_1 > 0$ is arbitrary, and that,

$$C_4 \leq \eta_2\lambda^2\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2 + \eta_2^{-1}\lambda^2 L^2\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2,$$

where $\eta_2 > 0$ is arbitrary. The proof follows by Using the bounds on C_3 and C_4 in (14). □

With the help of auxiliary Lemmas 5-7, we now prove an upper bound on the gradient tracking error.

Lemma 8. *Let Assumption 1-3 hold. If* $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda L}$, *then we have:* $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\right\|^2}{nL^2}\right] &\leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2}{nL^2}\right] \\
&\quad + \frac{24\lambda^2}{1-\lambda^2}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right] \\
&\quad + \frac{6\lambda^2\alpha_k^2}{1-\lambda^2}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{6\nu_a^2}{L^2}.
\end{aligned}$$

Proof. We apply the upper bounds in Lemma 5, 6 and 7 to Lemma 4 to obtain: $\forall k \geq 0, \forall \eta_1 > 0, \forall \eta_2 > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\right\|^2\right] \\
&\leq \lambda^2(1 + 12\lambda^2\alpha_k^2 L^2 + 2\lambda\alpha_k L + \eta_1 + 2\eta_2) \\
&\qquad \times \mathbb{E}\left[\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2\right] \\
&\quad + (3\lambda^2 n + 2)\nu_a^2 \\
&\quad + \left(18 + 2\eta_2^{-1}\right)\lambda^2 L^2\mathbb{E}\left[\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2\right] \\
&\quad + \left(6 + \eta_1^{-1}\right)\lambda^2\alpha_k^2 L^2 n\mathbb{E}\left[\left\|\overline{\mathbf{g}}_k\right\|^2\right].
\end{aligned} \tag{15}$$

[3]We note that η_1 and η_2 will be fixed later.

We set $\eta_1 = \frac{1-\lambda^2}{6\lambda^2}$ and $\eta_2 = \frac{1-\lambda^2}{12\lambda^2}$ in (15). It is straightforward to verify that if $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda^2 L}, \forall k \geq 0$, then we have:

$$\lambda^2(1 + 12\lambda^2\alpha_k^2 L^2 + 2\lambda\alpha_k L + \eta_1 + 2\eta_2) \leq \frac{1+\lambda^2}{2}. \quad (16)$$

Moreover, recall from (5) that

$$\mathbb{E}\big[\|\overline{\mathbf{g}}_k\|^2\big] \leq \mathbb{E}\big[\|\overline{\nabla \mathbf{f}}_k\|^2\big] + \nu_a^2/n. \quad (17)$$

Using (16), (17), $\eta_1 = \frac{1-\lambda^2}{6\lambda^2}$ and $\eta_2 = \frac{1-\lambda^2}{12\lambda^2}$ in (15), we have: if $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda^2 L}$, then

$$\begin{aligned}
&\mathbb{E}\big[\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\|^2\big] \\
&\leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2\big] + \left(\frac{6\lambda^2\alpha_k^2 L^2}{1-\lambda^2} + 5n\right)\nu_a^2 \\
&\quad + \frac{24\lambda^2 L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2\big] + \frac{6\lambda^2\alpha_k^2 L^2 n}{1-\lambda^2}\mathbb{E}\big[\|\overline{\nabla \mathbf{f}}_k\|^2\big].
\end{aligned}$$

The proof follows by $\frac{6\lambda^2\alpha_k^2 L^2}{1-\lambda^2} \leq 1$ if $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda L}, \forall k$. □

C. LTI dynamics

In this subsection, we establish the convergence rate of **GT-DSGD** for general smooth non-convex functions under an appropriate constant step-size such that $\alpha_k = \alpha, \forall k \geq 0$. To this end, we now jointly write Lemma 2 and 8 in the following linear-time-invariant system that characterizes the convergence of consensus and gradient tracking process.

Proposition 1. *Let Assumption 1-3 hold. If $0 < \alpha \leq \frac{1-\lambda^2}{24\lambda L}$, then we have the following (entry-wise) matrix-vector inequality hold: $\forall k \geq 0$,*

$$\mathbf{u}_{k+1} \leq \mathbf{G}\mathbf{u}_k + \mathbf{b}_k, \quad (18)$$

where the state vector $\mathbf{u}_k \in \mathbb{R}^2$, the system matrix $\mathbf{G} \in \mathbb{R}^{2\times 2}$ and the perturbation vector $\mathbf{b}_k \in \mathbb{R}^2$ are given by

$$\mathbf{u}_k = \begin{bmatrix} \mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] \\ \mathbb{E}\left[\frac{\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2}{nL^2}\right] \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \frac{1+\lambda^2}{2} & \frac{2\alpha^2\lambda^2 L^2}{1-\lambda^2} \\ \frac{24\lambda^2}{1-\lambda^2} & \frac{1+\lambda^2}{2} \end{bmatrix},$$

$$\mathbf{b}_k = \begin{bmatrix} 0 \\ \frac{6\lambda^2\alpha^2}{1-\lambda^2}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{6\nu_a^2}{L^2} \end{bmatrix}.$$

In light of Proposition 1, we first solve the range of α such that $\rho(\mathbf{G}) < 1$, using the following lemma from [30].

Lemma 9. *Let $\mathbf{X} \in \mathbb{R}^{d\times d}$ be a non-negative matrix and $\mathbf{x} \in \mathbb{R}^d$ be a positive vector. If $\mathbf{X}\mathbf{x} < \mathbf{x}$, then $\rho(\mathbf{X}) < 1$. Moreover, if $\mathbf{X}\mathbf{x} \leq z\mathbf{x}$, for some $z > 0$, then $\rho(\mathbf{X}) \leq z$.*

Lemma 10. *If $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{24\lambda}, \frac{(1-\lambda^2)^2}{15\lambda^2}\right\}\frac{1}{L}$, then we have $\rho(\mathbf{G}) < 1$ and hence $\sum_{k=0}^{\infty}\mathbf{G}^k = (\mathbf{I}_2 - \mathbf{G})^{-1}$.*

Proof. In the light of Lemma 9, we solve the range of α and a positive vector $\mathbf{s} = [s_1, s_2]^\top$ such that $\mathbf{G}\mathbf{s} < \mathbf{s}$, which is equivalent to the following two inequalities:

$$\begin{cases} \frac{1+\lambda^2}{2}s_1 + \frac{2\alpha^2\lambda^2 L^2}{1-\lambda^2}s_2 < s_1 \\ \frac{24\lambda^2}{1-\lambda^2}s_1 + \frac{1+\lambda^2}{2}s_2 < s_2 \end{cases} \iff \begin{cases} \alpha^2 < \frac{(1-\lambda^2)^2}{4\lambda^2 L^2}\frac{s_1}{s_2} \\ \frac{s_1}{s_2} < \frac{(1-\lambda^2)^2}{48\lambda^2} \end{cases}$$

We set $s_1/s_2 = (1-\lambda^2)^2/(50\lambda^2)$ and the proof follows by using it to solve for the range of α such that the first inequality above holds. □

Now, we prove an upper bound on the accumulated consensus errors along the algorithm path as follows.

Lemma 11. *Let Assumption 1-3 hold. If $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{24\lambda}, \frac{(1-\lambda^2)^2}{8\sqrt{6}\lambda^2}\right\}\frac{1}{L}$, then we have the following inequality.*

$$\begin{aligned}
\sum_{k=0}^{K}\mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] &\leq \frac{96\alpha^4\lambda^4 L^2}{(1-\lambda^2)^4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] \\
&\quad + \frac{16\alpha^2\lambda^4}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}_0\|^2}{n} + \frac{112\alpha^2\lambda^2\nu_a^2 K}{(1-\lambda^2)^3}.
\end{aligned}$$

Proof. We recursively apply (18) to obtain: $\forall k \geq 1$,

$$\mathbf{u}_k \leq \mathbf{G}^k\mathbf{u}_0 + \sum_{t=0}^{k-1}\mathbf{G}^t\mathbf{b}_{k-1-t}. \quad (19)$$

Summing up (19) over k from 1 to K, we obtain: $\forall K \geq 1$,

$$\begin{aligned}
\sum_{k=0}^{K}\mathbf{u}_k &\leq \sum_{k=0}^{K}\mathbf{G}^k\mathbf{u}_0 + \sum_{k=1}^{K}\sum_{t=0}^{k-1}\mathbf{G}^t\mathbf{b}_{k-1-t} \\
&\leq \left(\sum_{k=0}^{\infty}\mathbf{G}^k\right)\mathbf{u}_0 + \left(\sum_{k=0}^{\infty}\mathbf{G}^k\right)\sum_{k=0}^{K-1}\mathbf{b}_k \\
&= (\mathbf{I}_2 - \mathbf{G})^{-1}\mathbf{u}_0 + (\mathbf{I}_2 - \mathbf{G})^{-1}\sum_{k=0}^{K-1}\mathbf{b}_k. \quad (20)
\end{aligned}$$

In light of (20), we next compute an (entry-wise) upper bound on $(\mathbf{I}_2 - \mathbf{G})^{-1}$ as follows. We note that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{8\sqrt{6}\lambda^2 L}$,

$$\det(\mathbf{I}_2 - \mathbf{G}) = \frac{(1-\lambda^2)^2}{4} - \frac{48\alpha^2\lambda^4 L^2}{(1-\lambda^2)^2} \geq \frac{(1-\lambda^2)^2}{8}.$$

Using the lower bound on $\det(\mathbf{I}_2 - \mathbf{G})$ above, we have that

$$(\mathbf{I}_2 - \mathbf{G})^{-1} = \frac{(\mathbf{I}_2 - \mathbf{G})^*}{\det(\mathbf{I}_2 - \mathbf{G})} \leq \begin{bmatrix} \frac{4}{1-\lambda^2} & \frac{16\alpha^2\lambda^2 L^2}{(1-\lambda^2)^3} \\ \frac{192\lambda^2}{(1-\lambda^2)^3} & \frac{4}{1-\lambda^2} \end{bmatrix}. \quad (21)$$

We use (21) in (20) with $\|\mathbf{x}_0 - \mathbf{J}\mathbf{x}_0\| = 0$ to obtain: $\forall K \geq 1$,

$$\begin{aligned}
\sum_{k=0}^{K}\mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] &\leq \frac{16\alpha^2\lambda^2}{(1-\lambda^2)^3}\mathbb{E}\left[\frac{\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\|^2}{n}\right] \\
&\quad + \frac{96\alpha^4\lambda^4 L^2}{(1-\lambda^2)^4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{96\alpha^2\lambda^2\nu_a^2 K}{(1-\lambda^2)^3}. \quad (22)
\end{aligned}$$

Finally, we use the gradient tracking update (1a) to obtain:

$$\begin{aligned}
&\mathbb{E}\big[\,\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\|^2\,\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\,\|(\mathbf{W} - \mathbf{J})\mathbf{g}_0\|^2\,\big]|\mathcal{F}_0\big] \\
&= \mathbb{E}\big[\,\|(\mathbf{W} - \mathbf{J})(\mathbf{g}_0 - \nabla \mathbf{f}_0)\|^2\,\big] + \mathbb{E}\big[\,\|(\mathbf{W} - \mathbf{J})\nabla \mathbf{f}_0\|^2\,\big] \\
&\leq \lambda^2 n\nu_a^2 + \lambda^2\,\|\nabla \mathbf{f}_0\|^2, \quad (23)
\end{aligned}$$

where the second equality uses $\mathbb{E}[\mathbf{g}_0|\mathcal{F}_0] = \nabla \mathbf{f}_0$ and that $\nabla \mathbf{f}_0$ is constant and the last inequality uses $\|\mathbf{W} - \mathbf{J}\| = \lambda$. The proof follows by using (23) in (22). □

Lemma 11 states that the accumulated consensus error may be bounded by the accumulated average of local exact gradients and the accumulated variance of stochastic gradients. We next show that this bound leads to the convergence of **GT-DSGD** for general smooth non-convex functions, i.e., Theorem 1.

Proof of Theorem 1. We take the expectation of the descent inequality in Lemma 3 and sum up the resulting inequality over k from 0 to $K-1$, $\forall K \geq 1$, to obtain: if $0 < \alpha \leq \frac{1}{2L}$,

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{x}}_K)\right] \leq \mathbb{E}\left[F(\overline{\mathbf{x}}_0)\right] - \frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right] \\ - \frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{\alpha^2\nu_a^2 LK}{2n} \\ + \frac{\alpha L^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right]. \end{aligned} \tag{24}$$

Rearranging (24) and using that F is bounded below by F^* obtains: if $0 < \alpha \leq \frac{1}{2L}$, $\forall K \geq 1$,

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right] \leq \frac{2(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha} + \frac{\alpha\nu_a^2 LK}{n} \\ - \frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + L^2\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right]. \end{aligned} \tag{25}$$

Moreover, we observe: $\forall K \geq 1$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \\ &\leq \frac{2}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\left(\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i) - \nabla F(\overline{\mathbf{x}}_k)\right\|^2 + \left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right]\right) \\ &\leq 2L^2\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right] + 2\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right], \end{aligned}$$

where the last inequality uses the L-smoothness of F. Using (25) in the inequality above obtains: $\forall K \geq 1$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \leq \frac{4(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha} + \frac{2\alpha\nu_a^2 LK}{n} \\ - \sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + 4L^2\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right]. \end{aligned} \tag{26}$$

We finally apply the upper bound derived in Lemma 11 on the term of (26) to obtain: If $0 < \alpha \leq \min\left\{\frac{1}{2}, \frac{1-\lambda^2}{24\lambda}, \frac{(1-\lambda^2)^2}{8\sqrt{6}\lambda^2}\right\}\frac{1}{L}$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \\ &\leq \frac{4(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha} + \frac{2\alpha\nu_a^2 LK}{n} + \frac{448\alpha^2 L^2\lambda^2\nu_a^2 K}{(1-\lambda^2)^3} \\ &- \left(1 - \frac{384\alpha^4 L^4\lambda^4}{(1-\lambda^2)^4}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{64\alpha^2 L^2\lambda^4}{(1-\lambda^2)^3}\frac{\left\|\nabla \mathbf{f}_0\right\|^2}{n}. \end{aligned}$$

Clearly, if $0 < \alpha \leq \frac{1-\lambda^2}{5L\lambda}$, then $1 - \frac{384\alpha^4 L^4\lambda^4}{(1-\lambda^2)^4} \geq 0$, and the proof follows by dropping the negative term. □

VI. Convergence analysis under PL condition: constant step-size

In this section, we, built on top of the results established in Section V, develop general bounds on the iterates of **GT-DSGD** when the global function F further satisfies the PL condition and prove Theorem 2. The following is a useful inequality that may be found in [26].

Lemma 12. *Let Assumption 1 hold. We have:* $\forall \mathbf{x} \in \mathbb{R}^p$.

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

Proof. By (3) and the fact that F is bounded below by F^*, we have $F^* \leq F\left(\mathbf{x} - L^{-1}\nabla F(\mathbf{x})\right) \leq F(\mathbf{x}) - \frac{1}{2L}\left\|\nabla F(\mathbf{x})\right\|^2$, which yields the desired inequality. □

We conclude from Lemma 12 that, under Assumption 1 and 4, $\mu \leq L$ and recall $\kappa := \frac{L}{\mu} \geq 1$. The following lemma is helpful in establishing the performance of **GT-DSGD** at each node.

Lemma 13. *Let Assumption 1 hold. We have*

$$\frac{1}{n}\sum_{i=1}^{n}(F(\mathbf{x}_k^i) - F^*) \leq 2\left(F(\overline{\mathbf{x}}_k) - F^*\right) + L\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}.$$

Proof. Setting $\mathbf{y} = \mathbf{x}_k^i$ and $\mathbf{x} = \overline{\mathbf{x}}_k$ in (3), we obtain

$$\begin{aligned}&F(\mathbf{x}_k^i) - F^* \\ &\leq F(\overline{\mathbf{x}}_k) - F^* + \left\langle\nabla F(\overline{\mathbf{x}}_k), \mathbf{x}_k^i - \overline{\mathbf{x}}_k\right\rangle + \tfrac{1}{2}L\left\|\mathbf{x}_k^i - \overline{\mathbf{x}}_k\right\|^2, \\ &\leq F(\overline{\mathbf{x}}_k) - F^* + \left\|\nabla F(\overline{\mathbf{x}}_k)\right\|\left\|\mathbf{x}_k^i - \overline{\mathbf{x}}_k\right\| + \tfrac{1}{2}L\left\|\mathbf{x}_k^i - \overline{\mathbf{x}}_k\right\|^2, \\ &\leq F(\overline{\mathbf{x}}_k) - F^* + \tfrac{1}{2}L^{-1}\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2 + L\left\|\mathbf{x}_k^i - \overline{\mathbf{x}}_k\right\|^2 \\ &\leq 2\left(F(\overline{\mathbf{x}}_k) - F^*\right) + L\left\|\mathbf{x}_k^i - \overline{\mathbf{x}}_k\right\|^2, \end{aligned} \tag{27}$$

where the third inequality uses Young's inequality and the last inequality is due to Lemma 12. Averaging (27) over i from 1 to n proves the lemma. □

In the following, we refine several results developed in Section V. We first use the PL inequality to in Lemma 3.

Lemma 14. *Let Assumptions 1-4 hold. If* $0 < \alpha_k \leq \frac{1}{2L}$, *then we have:* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\frac{F(\overline{\mathbf{x}}_{k+1}) - F^*}{L}\middle|\mathcal{F}_k\right] \leq (1 - \mu\alpha_k)\frac{F(\overline{\mathbf{x}}_k) - F^*}{L} \\ + \frac{\alpha_k L}{2}\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n} + \frac{\alpha_k^2\nu_a^2}{2n}. \end{aligned}$$

Proof. The proof follows by using the PL condition in the descent inequality in Lemma 3 and then substracting F^* from both sides of the resulting inequality. □

We next use Lemma 12 to refine Lemma 8 as follows.

Lemma 15. *Let Assumption 1-3 hold. If* $0 < \alpha_k \leq \min\left\{\frac{1-\lambda^2}{12\lambda}, 1\right\}\frac{1}{2L}$, *then we have:* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\right\|^2}{nL^2}\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2}{nL^2}\right] \\ + \frac{24\lambda^2\alpha_k^2 L^2}{1-\lambda^2}\mathbb{E}\left[\frac{F(\overline{\mathbf{x}}_k) - F^*}{L}\right] \\ + \frac{27\lambda^2}{1-\lambda^2}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right] + \frac{6\nu_a^2}{L^2}. \end{aligned}$$

Proof. By Lemma 1(c) and Lemma 12, we have: $\forall k \geq 0$,

$$\begin{aligned}\left\|\overline{\nabla \mathbf{f}}_k\right\|^2 &\leq 2\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2 + 2\left\|\nabla F(\overline{\mathbf{x}}_k) - \overline{\nabla \mathbf{f}}_k\right\|^2 \\ &\leq 4L\left(F(\overline{\mathbf{x}}_k) - F^*\right) + 2L^2 n^{-1}\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2. \quad (28)\end{aligned}$$

Using the inequality above in Lemma 8 to obtain: $\forall k \geq 0$,

$$\begin{aligned}&\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\right\|^2}{nL^2}\right] \\ &\leq \left(\frac{24\lambda^2}{1-\lambda^2} + \frac{12\lambda^2\alpha_k^2 L^2}{1-\lambda^2}\right)\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right] + \frac{6\nu_a^2}{L^2} \\ &+ \frac{24\lambda^2\alpha_k^2 L}{1-\lambda^2}\mathbb{E}\left[F(\overline{\mathbf{x}}_k) - F^*\right] + \frac{1+\lambda^2}{2}\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2}{nL^2}\right].\end{aligned}$$

The proof follows by $\frac{12\lambda^2\alpha_k^2 L^2}{1-\lambda^2} \leq \frac{3\lambda^2}{1-\lambda^2}$ if $0 < \alpha_k \leq \frac{1}{2L}$. □

We now write the inequalities in Lemma 2, 14 and 15 jointly in a linear dynamics as follows.

Proposition 2. *Let Assumption 1-4 hold. If* $0 < \alpha_k \leq \min\left\{1, \frac{1-\lambda^2}{12\lambda}, \frac{(1-\lambda^2)^2}{4\sqrt{6}\lambda^2}\right\}\frac{1}{2L}$, *then we have the following (entry-wise) matrix-vector inequality:* $\forall k \geq 0$,

$$\mathbf{v}_{k+1} \leq \mathbf{H}_k \mathbf{v}_k + \mathbf{u}_k, \quad (29)$$

where the state vector $\mathbf{v}_k \in \mathbb{R}^3$, *the system matrix* $\mathbf{H} \in \mathbb{R}^{3\times 3}$ *and the perturbation vector* $\mathbf{u}_k \in \mathbb{R}^3$ *are given by*

$$\mathbf{v}_k = \begin{bmatrix} \mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right] \\ \mathbb{E}\left[\frac{F(\overline{\mathbf{x}}_k) - F^*}{L}\right] \\ \mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2}{nL^2}\right] \end{bmatrix}, \quad \mathbf{u}_k = \begin{bmatrix} 0 \\ \frac{\alpha_k^2\nu_a^2}{2n} \\ \frac{6\nu_a^2}{L^2} \end{bmatrix},$$

$$\mathbf{H}_k = \begin{bmatrix} \frac{1+\lambda^2}{2} & 0 & \frac{2\alpha_k^2\lambda^2 L^2}{1-\lambda^2} \\ \frac{\alpha_k L}{2} & 1-\mu\alpha_k & 0 \\ \frac{27\lambda^2}{1-\lambda^2} & \frac{24\lambda^2\alpha_k^2 L^2}{1-\lambda^2} & \frac{1+\lambda^2}{2} \end{bmatrix}.$$

In the following lemma, we find the range of the step-size α_k such that $\rho(\mathbf{H}_k) < 1, \forall k \geq 0$, with the help of Lemma 9.

Lemma 16. *Let Assumption 1-4 hold. If the step-size sequence* α_k *satisfies for all* k *that*

$$0 < \alpha_k \leq \overline{\alpha} := \min\left\{\frac{1}{2L}, \frac{(1-\lambda^2)^2}{42\lambda^2 L}, \frac{1-\lambda^2}{24\lambda L\kappa^{1/4}}, \frac{1-\lambda^2}{2\mu}\right\}, \quad (30)$$

then we have: $\rho(\mathbf{H}_k) \leq 1 - \frac{\mu\alpha_k}{2} < 1$, $\forall k \geq 0$.

Proof. In the light of Lemma 9, we solve for the range of the step-size α_k and a positive vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \delta_3]$ such that $\mathbf{H}_k\boldsymbol{\delta} \leq \left(1 - \frac{\mu\alpha_k}{2}\right)\boldsymbol{\delta}$, which may be written as

$$\frac{\mu\alpha_k}{2} + \frac{2\alpha_k^2\lambda^2 L^2}{1-\lambda^2}\frac{\delta_3}{\delta_1} \leq \frac{1-\lambda^2}{2}, \quad (31)$$

$$\kappa\delta_1 \leq \delta_2, \quad (32)$$

$$\frac{\mu\alpha_k}{2} \leq \frac{1-\lambda^2}{2} - \frac{27\lambda^2}{1-\lambda^2}\frac{\delta_1}{\delta_3} - \frac{24\lambda^2\alpha_k^2 L^2}{1-\lambda^2}\frac{\delta_2}{\delta_3}. \quad (33)$$

According to (32), we fix $\delta_1 = 1$ and $\delta_2 = \kappa$. We now impose that $0 < \alpha_k \leq \frac{1-\lambda^2}{2\mu}, \forall k \geq 0$. Then, according to (33), we choose $\delta_3 > 0$ such that $\frac{27\lambda^2}{1-\lambda^2}\frac{1}{\delta_3} + \frac{24\lambda^2\alpha_k^2 L^2}{1-\lambda^2}\frac{\kappa}{\delta_3} \leq \frac{1-\lambda^2}{4}$. It suffices to fix $\delta_3 = \frac{108\lambda^2}{(1-\lambda^2)^2} + \frac{96\lambda^2\alpha_k^2 L^2\kappa}{(1-\lambda^2)^2}$. Now, we use the fixed values of $\delta_1, \delta_2, \delta_3$ and the requirement that $0 < \alpha_k \leq \frac{1-\lambda^2}{2\mu}$ to solve the range of α_k such that (31) holds, i.e.,

$$\frac{216\alpha_k^2\lambda^4 L^2}{(1-\lambda^2)^3} + \frac{192\alpha_k^4\lambda^4 L^4\kappa}{(1-\lambda^2)^3} \leq \frac{1-\lambda^2}{4}.$$

It therefore suffices to choose α_k such that

$$0 < \alpha_k \leq \min\left\{\frac{1-\lambda^2}{6\lambda L\kappa^{1/4}}, \frac{(1-\lambda^2)^2}{42\lambda^2 L}\right\}.$$

Summarizing the obtained upper bounds on α_k in the discussion completes the proof. □

We note that $\overline{\alpha}$ defined in (30) is the same as the one given in Theorem 2. The following lemma drives upper bounds on several important quantities.

Lemma 17. *Let Assumption 1-4 hold. If* $0 < \alpha_k \leq \overline{\alpha}$, *where* $\overline{\alpha}$ *is given in (30), then we have:* $\forall k \geq 0$,

$$\left[\left(\mathbf{I}_3 - \mathbf{H}_k\right)^{-1}\mathbf{u}_k\right]_1 \leq \frac{288\lambda^4\alpha_k^5 L^3\kappa\nu_a^2}{n(1-\lambda^2)^4} + \frac{144\alpha_k^2\lambda^2\nu_a^2}{(1-\lambda^2)^3},$$

$$\left[\left(\mathbf{I}_3 - \mathbf{H}_k\right)^{-1}\mathbf{u}_k\right]_2 \leq \frac{3\alpha_k\nu_a^2}{2\mu n} + \frac{72\lambda^2\alpha_k^2\kappa\nu_a^2}{(1-\lambda^2)^3}.$$

Proof. By the definition of $\mathbf{H}_k$ in Proposition 2, we first compute the determinant of $(\mathbf{I}_3 - \mathbf{H}_k)$: $\forall k \geq 0$,

$$\begin{aligned}\det(\mathbf{I}_3 - \mathbf{H}_k) &= \frac{\mu\alpha_k(1-\lambda^2)^2}{4} - \frac{24\alpha_k^5 L^5\lambda^4}{(1-\lambda^2)^2} - \frac{54\mu\alpha_k^3 L^2\lambda^4}{(1-\lambda^2)^2} \\ &\geq \frac{\mu\alpha_k(1-\lambda^2)^2}{12}.\end{aligned}$$

if $0 < \alpha_k \leq \overline{\alpha}$, where $\overline{\alpha}$ is given in (30). Moreover, the adjugate of $\mathbf{I}_3 - \mathbf{H}_k$, denoted as $\underline{\mathbf{H}}^*$, is given by

$$[\underline{\mathbf{H}}^*]_{1,2} = \frac{48\lambda^4\alpha_k^4 L^4}{(1-\lambda^2)^2}, \qquad [\underline{\mathbf{H}}^*]_{1,3} = \frac{2\mu\alpha_k^3\lambda^2 L^2}{1-\lambda^2},$$

$$[\underline{\mathbf{H}}^*]_{2,2} \leq \frac{(1-\lambda^2)^2}{4}, \qquad [\underline{\mathbf{H}}^*]_{2,3} = \frac{\alpha_k^3 L^3\lambda^2}{1-\lambda^2}.$$

The proof follows by $(\mathbf{I}_3 - \mathbf{H}_k)^{-1} = \underline{\mathbf{H}}^* / \det(\mathbf{I}_3 - \mathbf{H}_k)$ and the definition of $\mathbf{u}_k$ given in Proposition 2. □

We are now ready to prove Theorem 2 that characterizes the performance of **GT-DSGD** under a constant step-size.

Proof of Theorem 2. We consider a constant step-size such that $\alpha_k = \alpha, \forall k \geq 0$, with $0 < \alpha \leq \overline{\alpha}$ where $\overline{\alpha}$ is given in (30). We denote $\mathbf{H}_k := \mathbf{H}$ and $\mathbf{u}_k := \mathbf{u}, \forall k \geq 0$, and recursively apply (29) from k to 1 to obtain: $\forall k \geq 1$,

$$\mathbf{v}_k \leq \mathbf{H}^k\mathbf{v}_0 + \sum_{t=0}^{k-1}\mathbf{H}^t\mathbf{u} \leq \mathbf{H}^k\mathbf{v}_0 + \left(\mathbf{I}_3 - \mathbf{H}\right)^{-1}\mathbf{u}. \quad (34)$$

It is then clear that the first two statements in Theorem 2 follow by using Lemma 16 and 17 in (34) and the third statement in Theorem 2 follows by Lemma 13. □

VII. Convergence analysis under PL condition: almost sure convergence

In this section, we prove Theorem 3, i.e., the almost sure sublinear convergence rates of **GT-DSGD** when the global function satisfies the PL condition under a family of stochastic approximation step-sizes. We first establish a key fact that under appropriate step-sizes, the stochastic gradient tracking errors are uniformly bounded in mean squared across all iterations. This fact will also be used in Section VIII.

Lemma 18. *Let Assumptions 1-4 hold. If $0 < \alpha_k \leq \overline{\alpha}$, for $\overline{\alpha}$ given in (30), then we have: $\sup_{k\geq 0} \mathbb{E}\big[\, \|\mathbf{y}_k - \mathbf{J}\mathbf{y}_k\|^2 \,\big] \leq \widehat{y}$, where $\widehat{y}$ is a positive constant given by*

$$\widehat{y} := \frac{30\lambda^2\overline{\alpha}^3 L^3\kappa\nu_a^2}{(1-\lambda^2)^2} + \frac{60n\lambda^2\overline{\alpha}^2L^3(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{16n\nu_a^2}{1-\lambda^2} + \lambda^2\|\nabla\mathbf{f}_0\|^2. \tag{35}$$

Proof. We prove by mathematical induction that for the state vector $\mathbf{v}_k$ defined in Proposition 2, there exists some positive constant vector $\widehat{\mathbf{v}} = [\widehat{v}_1, \widehat{v}_2, \widehat{v}_3]^\top$ such that

$$\mathbf{v}_k \leq \widehat{\mathbf{v}}, \qquad \forall k \geq 0. \tag{36}$$

if $0 < \alpha_k \leq \overline{\alpha}$, where $\overline{\alpha}$ is given in (30). We first note that in order to make (36) hold when $k = 0$, according to the definition of $\mathbf{v}_0$ and (23), it suffices to choose $\widehat{\mathbf{v}}$ such that

$$\widehat{\mathbf{v}}^\top \geq \left[0, \frac{F(\overline{\mathbf{x}}_0)-F^*}{L}, \frac{\lambda^2\nu_a^2}{L^2} + \frac{\lambda^2\|\nabla\mathbf{f}_0\|^2}{nL^2}\right]. \tag{37}$$

Next, we show that if $\mathbf{v}_k \leq \widehat{\mathbf{v}}$ for some $k \geq 0$ and then we also have $\mathbf{v}_{k+1} \leq \widehat{\mathbf{v}}$ with an appropriate choice of $\widehat{\mathbf{v}}$. In light of Proposition 2, we have $\mathbf{v}_{k+1} \leq \mathbf{H}_k\mathbf{v}_k + \mathbf{u}_k \leq \mathbf{H}_k\widehat{\mathbf{v}} + \mathbf{u}_k$, and hence it suffices to choose $\widehat{\mathbf{v}}$ such that $\mathbf{H}_k\widehat{\mathbf{v}} + \mathbf{u}_k \leq \widehat{\mathbf{v}}, \forall k$, which is equivalent to the following set of inequalities:

$$\frac{2\alpha_k^2\lambda^2L^2}{1-\lambda^2}\widehat{v}_3 \leq \frac{1-\lambda^2}{2}\widehat{v}_1, \tag{38}$$

$$\frac{\kappa}{2}\widehat{v}_1 + \frac{\alpha_k\nu_a^2}{2\mu n} \leq \widehat{v}_2, \tag{39}$$

$$\frac{27\lambda^2}{1-\lambda^2}\widehat{v}_1 + \frac{24\lambda^2\alpha_k^2L^2}{1-\lambda^2}\widehat{v}_2 + \frac{6\nu_a^2}{L^2} \leq \frac{1-\lambda^2}{2}\widehat{v}_3, \tag{40}$$

where $0 < \alpha_k \leq \overline{\alpha}$ and $\kappa = L/\mu$. First, we note that to make (38) hold, it suffices to choose $\widehat{v}_1$ as

$$\widehat{v}_1 = \frac{4\overline{\alpha}^2\lambda^2L^2}{(1-\lambda^2)^2}\widehat{v}_3. \tag{41}$$

Second, based on (37), (39), and (41), we choose $\widehat{v}_2$ as

$$\widehat{v}_2 = \frac{2\overline{\alpha}^2\lambda^2L^2\kappa}{(1-\lambda^2)^2}\widehat{v}_3 + \frac{\overline{\alpha}\nu_a^2}{2\mu n} + \frac{F(\overline{\mathbf{x}}_0)-F^*}{L}. \tag{42}$$

Third, to make (40) hold, it suffices to choose $\widehat{v}_3$ such that

$$\widehat{v}_3 \geq \frac{54\lambda^2}{(1-\lambda^2)^2}\widehat{v}_1 + \frac{48\lambda^2\overline{\alpha}^2L^2}{(1-\lambda^2)^2}\widehat{v}_2 + \frac{12\nu_a^2}{L^2(1-\lambda^2)}, \tag{43}$$

which, using (41) and (42), is equivalent to

$$\widehat{v}_3 \geq \frac{216\overline{\alpha}^2\lambda^4L^2}{(1-\lambda^2)^4}\widehat{v}_3 + \frac{96\lambda^4\overline{\alpha}^4L^4\kappa}{(1-\lambda^2)^4}\widehat{v}_3 + \frac{24\lambda^2\overline{\alpha}^3L\kappa\nu_a^2}{n(1-\lambda^2)^2} + \frac{48\lambda^2\overline{\alpha}^2L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{12\nu_a^2}{L^2(1-\lambda^2)}. \tag{44}$$

By the definition of $\overline{\alpha}$ in (30), we have $\frac{216\overline{\alpha}^2\lambda^4L^2}{(1-\lambda^2)^4} \leq \frac{6}{49}$ and that $\frac{96\overline{\alpha}^4\lambda^4L^4\kappa}{(1-\lambda^2)^4} \leq \frac{1}{3456}$; therefore, to make (44) hold, it suffices to choose $\widehat{v}_3$ such that

$$\widehat{v}_3 \geq \frac{30\lambda^2\overline{\alpha}^3L\kappa\nu_a^2}{n(1-\lambda^2)^2} + \frac{60\lambda^2\overline{\alpha}^2L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{15\nu_a^2}{L^2(1-\lambda^2)}.$$

Based on the above inequality and (37), we choose $\widehat{v}_3$ as

$$\widehat{v}_3 = \frac{30\lambda^2\overline{\alpha}^3L\kappa\nu_a^2}{n(1-\lambda^2)^2} + \frac{60\lambda^2\overline{\alpha}^2L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{16\nu_a^2}{L^2(1-\lambda^2)} + \frac{\lambda^2\|\nabla\mathbf{f}_0\|^2}{nL^2}.$$

The induction is complete and the proof then follows by the definition of $\mathbf{v}_k$ in Proposition 2. □

We prove Theorem 3 using the Robbins-Siegmund almost supermartingale convergence theorem [25], presented as follows.

Lemma 19 (Robbins-Siegmund). *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}, \mathbb{P})$ be a filtered space. Suppose that Z_k, B_k, C_k and D_k are nonnegative and $\mathcal{F}_k$-measurable random variables such that*

$$\mathbb{E}\left[Z_{k+1}|\mathcal{F}_k\right] \leq (1+B_k)\,Z_k + C_k - D_k, \ \forall k \geq 0.$$

Then on the event $\{\sum_{k=0}^\infty B_k < \infty, \ \sum_{k=0}^\infty C_k < \infty\}$, we have that $\lim_{k\to\infty} Z_k$ exists and is finite almost surely, and that $\sum_{k=0}^\infty D_k < \infty$ almost surely.

We are now ready to present the proof of Theorem 3, where we construct appropriate almost supermartingales that characterize the sample path-wise convergence rate of **GT-DSGD** under a family of stochastic approximation step-sizes.

Proof of Theorem 3. We consider the step-size sequence $\{\alpha_k\}$ of the following form: $\forall k \geq 0$,

$$\alpha_k = \delta(k+\varphi)^{-\epsilon}, \quad \text{where } \delta \geq 1/\mu \text{ and } \epsilon \in (0.5, 1], \tag{45}$$

such that $\varphi \geq \max\{(\delta/\overline{\alpha})^{1/\epsilon}, \frac{4}{1-\lambda^2}\}$. Hence, $0 < \alpha_k \leq \overline{\alpha}$ for $\overline{\alpha}$ given in (30). We construct $\mathcal{F}_k$-adapted processes: $\forall k \geq 0$,

$$R_k := (k+\varphi)^\tau\widetilde{x}_k := (k+\varphi)^\tau n^{-1}\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2,$$
$$Q_k := (k+\varphi)^\tau\Delta_k := (k+\varphi)^\tau L^{-1}(F(\overline{\mathbf{x}}_k) - F^*),$$

where $\tau = 2\epsilon - 1 - \epsilon_1$, where $\epsilon_1 \in (0, 2\epsilon-1)$ is an arbitrarily small constant. By $1+x \leq e^x, \forall x \in \mathbb{R}$, we have $(k+\varphi+1)^\tau = (k+\varphi)^\tau\big(1+\frac{1}{k+\varphi}\big)^\tau \leq (k+\varphi)^\tau e^{\frac{\tau}{k+\varphi}}$. Since $0 < \frac{\tau}{k+\varphi} \leq 1$, we have: $\forall k \geq 0$,

$$(k+\varphi+1)^\tau \leq e(k+\varphi)^\tau. \tag{46}$$

Further, by $e^x \leq 1+x+x^2$ for $0 \leq x \leq 1$,[4] we have: $\forall k \geq 0$,

$$(k+\varphi+1)^\tau \leq \left(1 + \frac{\tau}{k+\varphi} + \frac{\tau^2}{(k+\varphi)^2}\right)(k+\varphi)^\tau. \tag{47}$$

Recursion of R_k. We use Lemma 18 in Lemma 2 with the definition of α_k in (45) to obtain: $\forall k \geq 0$,

$$\mathbb{E}\left[\widetilde{x}_{k+1}\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\widetilde{x}_k\right] + \frac{2\lambda^2\widehat{y}}{n(1-\lambda^2)}\frac{\delta^2}{(k+\varphi)^{2\epsilon}}, \tag{48}$$

[4] Note that $e^x = 1 + x + x^2\sum_{k=2}^\infty \frac{x^{k-2}}{k!}, \forall x \in \mathbb{R}$. If $0 \leq x \leq 1$, then we have $e^x \leq 1 + x + x^2\sum_{k=2}^\infty \frac{1}{k!} = 1 + x + (e-2)x^2 \leq 1 + x + x^2$.

where $\widehat{y}$ is given in (35). We multiply (48) by $(k+\varphi+1)^\tau$ and then apply (46) and (47) to obtain: $\forall k \geq 0$,

$$\mathbb{E}\left[R_{k+1}\right] \leq \underbrace{\frac{1+\lambda^2}{2}\left(1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}\right)}_{T_k}\mathbb{E}\left[R_k\right] + \frac{2e\lambda^2\widehat{y}}{n(1-\lambda^2)}\frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{49}$$

Since $\varphi \geq \frac{4}{1-\lambda^2}$, i.e., $\frac{\tau}{k+\varphi} \leq \frac{1-\lambda^2}{4}, \forall k \geq 0$, we have

$$\begin{aligned} T_k &= \left(1-\frac{1-\lambda^2}{2}\right)\left(1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}\right) \\ &\leq 1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}-\frac{1-\lambda^2}{2} \\ &\leq 1+\frac{\tau^2}{(k+\varphi)^2}-\frac{1-\lambda^2}{4}. \end{aligned} \tag{50}$$

Using (50) in (49), we have: $\forall k \geq 0$,

$$\mathbb{E}\left[R_{k+1}\right] \leq \left(1+\frac{\tau^2}{(k+\varphi)^2}\right)\mathbb{E}\left[R_k\right]-\frac{1-\lambda^2}{4}\mathbb{E}\left[R_k\right] + \frac{2e\lambda^2\widehat{y}}{n(1-\lambda^2)}\frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{51}$$

Note that $\sum_{k=0}^{\infty}(k+\varphi)^{-2} < \infty$ and $\sum_{k=0}^{\infty}(k+\varphi)^{\tau-2\epsilon} < \infty$ since $2\epsilon-\tau > 1$. Applying a special case of Lemma 19 for deterministic recursions in (51) leads to $\sum_{k=0}^{\infty}\mathbb{E}\left[R_k\right] < \infty$. Since R_k is nonnegative, by monotone convergence theorem, we have $\mathbb{E}\left[\sum_{k=0}^{\infty}R_k\right] = \sum_{k=0}^{\infty}\mathbb{E}\left[R_k\right] < \infty$ which implies

$$\mathbb{P}\left(\sum_{k=0}^{\infty}R_k < \infty\right) = 1. \tag{52}$$

The first statement in Theorem 3 then follows by (52).

Recursion of Q_k. We recall from Lemma 14: $\forall k \geq 0$,

$$\mathbb{E}\left[\Delta_{k+1}|\mathcal{F}_k\right] \leq \left(1-\frac{\mu\delta}{(k+\varphi)^\epsilon}\right)\Delta_k+\frac{L\delta}{2(k+\varphi)^\epsilon}\widetilde{x}_k + \frac{\nu_a^2}{2n}\frac{\delta^2}{(k+\varphi)^{2\epsilon}}. \tag{53}$$

We multiply (53) by $(k+\varphi+1)^\tau$ and then use (46) and (47) to obtain: $\forall k \geq 0$,

$$\mathbb{E}[Q_{k+1}|\mathcal{F}_k] \leq \underbrace{\left(1-\frac{\mu\delta}{(k+\varphi)^\epsilon}\right)\left(1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}\right)}_{P_k}Q_k + \frac{eL\delta}{2(k+\varphi)^\epsilon}R_k+\frac{e\nu_a^2}{2n}\frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{54}$$

We observe that

$$\begin{aligned} P_k &\leq 1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}-\frac{\mu\delta}{(k+\varphi)^\epsilon} \\ &\leq 1+\frac{\tau^2}{(k+\varphi)^2}-\frac{\mu\delta-\tau}{(k+\varphi)^\epsilon}. \end{aligned} \tag{55}$$

We use (55) in (54) to obtain: $\forall k \geq 0$,

$$\mathbb{E}\left[Q_{k+1}|\mathcal{F}_k\right] \leq \left(1+\frac{\tau^2}{(k+\varphi)^2}\right)Q_k-\frac{\mu\delta-\tau}{(k+\varphi)^\epsilon}Q_k + \frac{eL\delta}{2(k+\varphi)^\epsilon}R_k+\frac{e\nu_a^2}{2n}\frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{56}$$

Recall that $\sum_{k=0}^{\infty}(k+\varphi)^{-2} < \infty$ and $\sum_{k=0}^{\infty}(k+\varphi)^{\tau-2\epsilon} < \infty$ since $2\epsilon-\tau > 1$. Note that $\delta \geq 1/\mu$, i.e., $\mu\delta > \tau$, applying Lemma 19 in (56) with the help of (52) gives:

$$\mathbb{P}\Big(\lim_{k\to\infty}Q_k = Q\Big) = 1, \tag{57}$$

where Q is some almost surely finite random variable, and

$$\mathbb{P}\left(\sum_{k=0}^{\infty}\frac{\mu\delta-\tau}{(k+\varphi)^\epsilon}Q_k < \infty\right) = 1. \tag{58}$$

Since $\sum_{k=0}^{\infty}\frac{\mu\delta-\tau}{(k+\varphi)^\epsilon} = \infty$, where $\epsilon \in (0.5, 1]$, we have

$$\left\{\sum_{k=0}^{\infty}\frac{\mu\delta-\tau}{(k+\varphi)^\epsilon}Q_k < \infty\right\} \subseteq \left\{\liminf_{k\to\infty}Q_k = 0\right\}, \tag{59}$$

where "$\subseteq$" denotes the inclusion relation for two events. By the monotonicity of $\mathbb{P}(\cdot)$, (58) and (59) lead to

$$\mathbb{P}\Big(\liminf_{k\to\infty}Q_k = 0\Big) = 1. \tag{60}$$

From (60) and (57), we conclude that $\mathbb{P}\left(Q=0\right)=1$ and then the proof follows by (52) and Lemma 13. □

VIII. Convergence analysis under PL condition: asymptotically optimal rate in mean

In this section, we prove Theorem 4 and Corollary 2, i.e., the asymptotically optimal convergence rate of **GT-DSGD** in expectation and the corresponding transient time to achieve network-independent performance, when the global function F satisfies the PL condition. Recall that in this context we focus on the following step-size sequence [1]:

$$\alpha_k = \frac{\beta}{k+\gamma}, \qquad \forall k \geq 0, \tag{61}$$

where $\beta > 0$ and $\gamma > 0$ are parameters to be restricted later. We require $\gamma \geq \beta/\overline{\alpha}$ so that $0 < \alpha_k \leq \overline{\alpha}$ for $\overline{\alpha}$ in (30). We first prove a non-asymptotic rate on the consensus errors.

Lemma 20. *Let Assumption 1-4 hold. If $\gamma \geq \max\left\{\frac{\beta}{\overline{\alpha}}, \frac{8}{1-\lambda^2}\right\}$ for $\overline{\alpha}$ given in (30), then we have: $\forall k \geq 0$,*

$$\mathbb{E}\left[\left\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\right\|^2\right] \leq \frac{\widehat{x}\beta^2}{(k+\gamma)^2}. \tag{62}$$

where $\widehat{x} := 8\lambda^2\widehat{y}(1-\lambda^2)^{-2}$ for $\widehat{y}$ given in (35).

Proof. We prove by induction that there exists a constant $\widehat{x}$ such that (62) holds. First, since $\mathbf{x}_0^i = \mathbf{x}_0^r, \forall i, r \in \mathcal{V}$, (62) holds trivially when $k=0$. We next show that if (62) holds for some $k \geq 0$ and then it also holds for $k+1$. From Lemma 2 and 18, we have: $\forall k \geq 0$,

$$\mathbb{E}[\|\mathbf{x}_{k+1}-\mathbf{J}\mathbf{x}_{k+1}\|^2] \leq \frac{1+\lambda^2}{2}\mathbb{E}[\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\|^2]+\frac{2\lambda^2\widehat{y}\alpha_k^2}{1-\lambda^2}.$$

Therefore, it suffices to choose $\widehat{x}$ such that $\forall k \geq 0$,

$$\frac{1+\lambda^2}{2}\frac{\widehat{x}\beta^2}{(k+\gamma)^2}+\frac{2\lambda^2\widehat{y}}{1-\lambda^2}\frac{\beta^2}{(k+\gamma)^2} \leq \frac{\widehat{x}\beta^2}{(k+\gamma+1)^2},$$

which is equivalent to

$$\frac{2\lambda^2\widehat{y}}{1-\lambda^2} \leq \left(\frac{(k+\gamma)^2}{(k+\gamma+1)^2}-\frac{1+\lambda^2}{2}\right)\widehat{x}. \tag{63}$$

Since the RHS of (63) monotonically increases with k, we suffice to choose $\widehat{x}$ such that (63) holds when $k=0$, i.e.,

$$\frac{2\lambda^2\widehat{y}}{1-\lambda^2} \leq \left(\frac{\gamma^2}{(\gamma+1)^2} - \frac{1+\lambda^2}{2}\right)\widehat{x} = \left(\frac{1-\lambda^2}{2} - \frac{2\gamma+1}{(\gamma+1)^2}\right)\widehat{x}.$$

Since $\frac{2\gamma+1}{(\gamma+1)^2} \leq \frac{2}{\gamma}$, it suffices to choose $\widehat{x}$ such that $\frac{2\lambda^2\widehat{y}}{1-\lambda^2} \leq \left(\frac{1-\lambda^2}{2} - \frac{2}{\gamma}\right)\widehat{x}$. Finally, if $\gamma \geq \frac{8}{1-\lambda^2}$, it can be observed that the induction is complete by setting $\widehat{x} := 8\lambda^2\widehat{y}(1-\lambda^2)^{-2}$. □

We next present a useful lemma adapted from [23], [32], [33].

Lemma 21. *Consider the step-size sequence $\{\alpha_k\}$ in (61). We have: for any nonnegative integers a, b such that $0 \leq a \leq b$,*

$$\prod_{s=a}^{b}(1-\mu\alpha_s) \leq \frac{(a+\gamma)^{\mu\beta}}{(b+\gamma+1)^{\mu\beta}}.$$

Proof. By (61) and $1+x \leq e^x, \forall x \in \mathbb{R}$, we have: $0 \leq a \leq b$,

$$\prod_{s=a}^{b}(1-\mu\alpha_s) = \prod_{s=a}^{b}\left(1-\frac{\mu\beta}{s+\gamma}\right) \leq \exp\left\{-\sum_{s=a}^{b}\frac{\mu\beta}{s+\gamma}\right\}. \tag{64}$$

Since $\frac{1}{s+\gamma} \geq \int_{s+\gamma}^{s+\gamma+1}\frac{1}{x}dx, \forall s \geq 0$, we have: $0 \leq a \leq b$,

$$\sum_{s=a}^{b}\frac{1}{s+\gamma} \geq \sum_{s=a}^{b}\int_{s+\gamma}^{s+\gamma+1}\frac{1}{x}dx = \log\left(\frac{b+\gamma+1}{a+\gamma}\right). \tag{65}$$

Applying (65) to (64) completes the proof. □

Now we are ready to prove Theorem 4 through a non-asymptotic analysis inspired by [9], [23], [33], [34], [38].

Proof of Theorem 4. We denote $\Psi_k := \mathbb{E}[L^{-1}(F(\overline{\mathbf{x}}_k) - F^*)]$. Using Lemma 20 in Lemma 14 gives: if $\gamma \geq \max\left\{\frac{\beta}{\overline{\alpha}}, \frac{8}{1-\lambda^2}\right\}$,

$$\Psi_{k+1} \leq (1-\mu\alpha_k)\Psi_k + \widehat{u}\alpha_k^2 + \widehat{z}\alpha_k^3, \quad \forall k \geq 0, \tag{66}$$

where $\widehat{u}$ and $\widehat{z}$ are defined as, for $\widehat{x}$ given in (62),

$$\widehat{u} := \frac{\nu_a^2}{2n} \quad \text{and} \quad \widehat{z} := \frac{L\widehat{x}}{2n}. \tag{67}$$

We recursively apply (66) from k to 0 to obtain[5]: $\forall k \geq 1$,

$$\begin{aligned}
&\Psi_k \\
&\leq \Psi_0\prod_{t=0}^{k-1}(1-\mu\alpha_t) + \sum_{t=0}^{k-1}\left(\left(\widehat{u}\alpha_t^2 + \widehat{z}\alpha_t^3\right)\prod_{l=t+1}^{k-1}(1-\mu\alpha_l)\right) \\
&\leq \Psi_0\frac{\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \sum_{t=0}^{k-1}\left(\frac{\widehat{u}\beta^2}{(t+\gamma)^2} + \frac{\widehat{z}\beta^3}{(t+\gamma)^3}\right)\frac{(t+1+\gamma)^{\mu\beta}}{(k+\gamma)^{\mu\beta}} \\
&= \Psi_0\frac{\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \frac{\widehat{u}\beta^2}{(k+\gamma)^{\mu\beta}}\sum_{t=0}^{k-1}\frac{(t+1+\gamma)^{\mu\beta}}{(t+\gamma)^2} \\
&\quad + \frac{\widehat{z}\beta^3}{(k+\gamma)^{\mu\beta}}\sum_{t=0}^{k-1}\frac{(t+1+\gamma)^{\mu\beta}}{(t+\gamma)^3},
\end{aligned} \tag{68}$$

where the second inequality is due to Lemma 21. Furthermore, by $1+x \leq e^x, \forall x \in \mathbb{R}$, we have: for $0 \leq t \leq k-1$,

$$\frac{(t+1+\gamma)^{\mu\beta}}{(t+\gamma)^{\mu\beta}} = \left(1+\frac{1}{t+\gamma}\right)^{\mu\beta} \leq \exp\left\{\frac{\mu\beta}{\gamma}\right\} \leq \sqrt{e}, \tag{69}$$

[5]For a sequence $\{s_k\}$, we adopt the convention $\prod_{k=x}^{y}s_k = 1$ if $y < x$.

where the last inequality uses $\mu\beta/\gamma \leq \mu\overline{\alpha} \leq 0.5$. We use (69) in (68) to obtain: $\forall k \geq 1$,

$$\begin{aligned}
\Psi_k \leq \Psi_0\frac{\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \frac{\sqrt{e}\widehat{u}\beta^2}{(k+\gamma)^{\mu\beta}}\sum_{s=\gamma}^{k-1+\gamma}s^{\mu\beta-2} \\
+ \frac{\sqrt{e}\widehat{z}\beta^3}{(k+\gamma)^{\mu\beta}}\sum_{s=\gamma}^{k-1+\gamma}s^{\mu\beta-3}.
\end{aligned} \tag{70}$$

By $s^{\mu\beta-2} \leq \max\left\{\int_s^{s+1}x^{\mu\beta-2}dx, \int_{s-1}^{s}x^{\mu\beta-2}dx\right\}$, we have: if $\beta > 1/\mu$, then $\forall k \geq 1$,

$$\sum_{s=\gamma}^{k-1+\gamma}s^{\mu\beta-2} \leq \int_{\gamma-1}^{k+\gamma}x^{\mu\beta-2}dx \leq \frac{(k+\gamma)^{\mu\beta-1}}{\mu\beta-1}. \tag{71}$$

Likewise, by $s^{\mu\beta-3} \leq \max\left\{\int_s^{s+1}x^{\mu\beta-3}dx, \int_{s-1}^{s}x^{\mu\beta-3}dx\right\}$, we have: if $\beta > 2/\mu$, then $\forall k \geq 1$,

$$\sum_{s=\gamma}^{k-1+\gamma}s^{\mu\beta-3} \leq \int_{\gamma-1}^{k+\gamma}x^{\mu\beta-3}dx \leq \frac{(k+\gamma)^{\mu\beta-2}}{\mu\beta-2}. \tag{72}$$

Now, we apply (71) and (72) in (70) to obtain: $\forall k \geq 1$,

$$\Psi_k \leq \frac{\Psi_0\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \frac{\sqrt{e}\widehat{u}\beta^2}{(\mu\beta-1)(k+\gamma)} + \frac{\sqrt{e}\widehat{z}\beta^3}{(\mu\beta-2)(k+\gamma)^2}. \tag{73}$$

Using (73) and Lemma 20 in Lemma 13, we obtain: $\forall k \geq 1$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[F(\mathbf{x}_k^i) - F^*] \leq &\frac{2(F(\overline{\mathbf{x}}_0)-F^*)}{(k/\gamma+1)^{\mu\beta}} + \frac{2\sqrt{e}L\widehat{u}\beta^2}{(\mu\beta-1)(k+\gamma)} \\
&+ \frac{2\sqrt{e}L\widehat{z}\beta^3}{(\mu\beta-2)(k+\gamma)^2} + \frac{2\widehat{z}\beta^2}{(k+\gamma)^2}.
\end{aligned}$$

The proof follows by that $\frac{\widehat{z}\beta^2}{(k+\gamma)^2} \leq \frac{L\widehat{z}\beta^3}{(\mu\beta-2)(k+\gamma)^2}$ and by recalling the definitions of $\widehat{u}$ and $\widehat{z}$ given in (67). □

Proof of Corollary 2. We derive the conditions under which the rate expression in Theorem 4 is network-independent. We first solve for the lower bound on k such that

$$\frac{L\nu_a^2\beta^2}{n(\mu\beta-1)(k+\gamma)} \geq \frac{L^2\widehat{x}\beta^3}{n(\mu\beta-2)(k+\gamma)^2},$$

which may be written equivalently as

$$k+\gamma \geq \frac{\mu\beta-1}{\mu\beta-2}\frac{L\widehat{x}\beta}{\nu_a^2}. \tag{74}$$

We suppose that $\|\nabla\mathbf{f}_0\|^2 = \mathcal{O}(n)$, $\beta = \theta/\mu$, where $\theta > 2$. Since $\overline{\alpha}L = \mathcal{O}\left(\frac{1-\lambda}{\lambda\kappa^{1/4}}\right)$, for $\overline{\alpha}$ defined in (30), we have

$$\widehat{x} = \mathcal{O}\left(\frac{\lambda^2 n\nu_a^2}{(1-\lambda)^3} + \frac{\lambda\kappa^{1/4}\nu_a^2}{1-\lambda} + \frac{\lambda^2 nL(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda)^2\kappa^{1/2}}\right),$$

where $\widehat{x}$ is defined in (62). Therefore, to make (74) hold, it suffices to let

$$k \gtrsim \frac{\lambda^2 n\kappa}{(1-\lambda)^3} + \frac{\lambda\kappa^{5/4}}{1-\lambda} + \frac{\lambda^2 n\kappa^{1/2}L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda)^2\nu_a^2}. \tag{75}$$

Next, we solve for the range of k such that for some $\delta \in [1, \theta)$, $(\frac{k}{\gamma}+1)^\theta \geq (\frac{k+1}{\kappa})^\delta$, i.e., $\frac{(k+\gamma)^\theta}{(k+1)^\delta} \geq \frac{\gamma^\theta}{\kappa^\delta}$. Since $\gamma > 1$, it suffices choose k such that

$$k \geq \gamma^{\frac{\theta}{\theta-\delta}}\kappa^{-\frac{\delta}{\theta-\delta}}. \tag{76}$$

We fix $\gamma = \max\{\frac{\theta}{\mu\overline{\alpha}}, \frac{8}{1-\lambda^2}\} \asymp \max\{\kappa, \frac{\lambda^2\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{1}{1-\lambda}\}$. Using (75) and (76) in Theorem 4, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(F(\mathbf{x}_k^i) - F^*\right) = \mathcal{O}\left(\frac{\kappa^\delta\left(F(\overline{\mathbf{x}}_0) - F^*\right)}{k^\delta} + \frac{\kappa\nu_a^2}{n\mu k}\right),$$

if $k \gtrsim \max\{K_1, K_2\}$, where K_1 and K_2 are given by

$$K_1 = \frac{\lambda^2 n\kappa}{(1-\lambda)^3} + \frac{\lambda\kappa^{5/4}}{1-\lambda} + \frac{\lambda^2 n\kappa^{1/2}L(F(\overline{\mathbf{x}}_0) - F^*)}{(1-\lambda)^2\nu_a^2},$$

$$K_2 = \max\left\{\kappa, \frac{\lambda^2\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{1}{1-\lambda}\right\}^{\frac{\theta}{\theta-\delta}}\kappa^{-\frac{\delta}{\theta-\delta}}.$$

The proof follows by setting $\delta = 2$ and $\theta = 6$ in the above. □

IX. Conclusion

In this paper, we comprehensively improve the existing convergence results of stochastic first-order methods based on gradient tracking for online stochastic nonconvex problems. In particular, for both constant and decaying step-sizes, we systematically develop the conditions under which the performance of **GT-DSGD** matches that of the centralized minibatch **SGD** for both general non-convex functions and non-convex functions that further satisfy the PL condition. Our results significantly improve upon the existing theory, which suggests that **GT-DSGD** is strictly worse than centralized minibatch **SGD**. For a family of stochastic approximation step-sizes, we establish the global sublinear convergence to an optimal solution on almost every sample path of **GT-DSGD** when the global objective function satisfies the PL condition.

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

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

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

[4] 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.

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

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

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

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

[9] S. Pu and A. Nedich, "Distributed stochastic gradient tracking methods," *Math. Program.*, pp. 1–49, 2020.

[10] K. Yuan, S. A. Alghunaim, B. Ying, and A. H. Sayed, "On the influence of bias-correction on distributed stochastic optimization," *IEEE Transactions on Signal Processing*, 2020.

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

[12] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning–Part I: Algorithm development," *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 708–723, 2018.

[13] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2: Decentralized training over decentralized data," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4848–4856.

[14] 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.

[15] 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.

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

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

[18] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Math. Program.*, vol. 176, no. 1-2, pp. 497–544, 2019.

[19] 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.

[20] S. Pu, W. Shi, J. Xu, and A. Nedich, "A push-pull gradient method for distributed optimization in networks," in *IEEE Conference on Decision and Control*, 2018, pp. 3385–3390.

[21] S. Lu, X. Zhang, H. Sun, and M. Hong, "GNSD: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization," in *2019 IEEE Data Science Workshop, DSW 2019*, 2019, pp. 315–321.

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

[23] S. Kar and José M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 674–690, 2011.

[24] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," *IEEE Trans. Control Netw. Syst.*, 2020.

[25] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971.

[26] B. T Polyak, "Introduction to optimization. 1987," *Optimization Software, Inc, New York*.

[27] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. 35th Int. Conf. Mach. Learn.*, 10–15 Jul 2018, pp. 1467–1476.

[28] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. 2016, pp. 795–811, Springer.

[29] M. Zhu and S. Martínez, "Discrete-time dynamic average consensus," *Automatica*, vol. 46(2), pp. 322–329, 2010.

[30] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 2012.

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

[32] M. B. Nevelson and R. Z. Hasminskii, *Stochastic approximation and recursive estimation*, vol. 47, American Mathematical Soc., 1976.

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

[34] S. Pu, A. Olshevsky, and I. C. Paschalidis, "Asymptotic network independence in distributed stochastic optimization for machine learning: Examining distributed and centralized stochastic gradient descent," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 114–122, 2020.

[35] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[36] A. Antoniadis, I. Gijbels, and M. Nikolova, "Penalized likelihood regression for generalized linear models with non-quadratic penalties," *Ann. Inst. Statistical Math.*, vol. 63, no. 3, pp. 585–615, 2011.

[37] 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.

[38] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions," *Journal of Machine Learning Research*, vol. 21, no. 58, pp. 1–47, 2020.