OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Optimal Complexity in Non-Convex Decentralized Learning over Time-Varying Networks

Xinmeng Huang XINMENGH@SAS.UPENN.EDU
Graduate Group in Applied Mathematics and Computational Science
University of Pennsylvania

Kun Yuan✉ KUNYUAN@PKU.EDU.CN
Center for Machine Learning Research
Peking University

Abstract

Decentralized optimization with time-varying networks is an emerging paradigm in machine learning. It saves remarkable communication overhead in large-scale deep training and is more robust in wireless scenarios especially when nodes are moving. Federated learning can also be regarded as decentralized optimization with time-varying communication patterns alternating between global averaging and local updates.

While numerous studies exist to clarify its theoretical limits and develop efficient algorithms, it remains unclear what the optimal complexity is for non-convex decentralized stochastic optimization over time-varying networks. The main difficulties lie in how to gauge the effectiveness when transmitting messages between two nodes via time-varying communications, and how to establish the lower bound when the network size is fixed (which is a prerequisite in stochastic optimization). This paper resolves these challenges and establish the first lower bound complexity. We also develop a new decentralized algorithm to nearly attain the lower bound, showing the tightness of the lower bound and the optimality of our algorithm.

1. Introduction

Decentralized optimization. Decentralized optimization is an emerging learning paradigm in which each node only communicates with its immediate neighbors per iteration. By avoiding the central server and maintaining a more balanced communication between each pair of connected nodes, decentralized approaches can significantly speedup the training process of large-scale machine learning models [4, 10, 43]. Although decentralized optimization has been extensively studied in literature, its performance limits with *time-varying* communication patterns has not been fully explored. This paper provides a better understanding in optimal complexity for non-convex decentralized stochastic optimization over time-varying communication networks.

Time-varying communication pattern. Decentralized optimization over time-varying communication networks is ubiquitous in applications. In large-scale deep neural network training, sparse and time-varying network topologies such as one-peer exponential graph [4, 42] and EquiRand [32] endow decentralized learning with a state-of-the-art balance between communication efficiency and convergence rate. In wireless signal processing, time-varying topologies naturally emerge when the nodes (such as cellphones, drones, robots, etc.) are moving [37, 38]. Federated learning [22, 33]

can also be regarded as a special decentralized learning paradigm which admits a time-varying communication pattern alternating between global averaging and local updates.

Prior results in theoretical limits. A series of pioneering works have attempted to establish the optimal complexity in decentralized optimization over *static* communication networks. In the deterministic regime, [30, 31, 34] clarified the theoretical limits and proposed algorithms to (nearly) attain these limits. In the stochastic regime, recent works [21, 47] have established the optimal complexity in the non-convex setting. However, there are few studies on theoretical limits in decentralized optimization over *time-varying* communication networks. A recent useful work [14] establishes the optimal complexity over time-varying networks for deterministic and strongly-convex problems. While this bound is inspiring, its analysis, as well as all existing results in literature to our knowledge, cannot be easily extended to the stochastic setting due to challenges below.

Challenges. When considering a static network topology, it is known that the optimal complexity in decentralized optimization is typically proportional to diameter D of the static topology [30]. Clarifying how the diameter D affects the algorithmic convergence is the key to justifying the influence of the communication network on the optimal convergence rate. However, it is unclear in literature how to gauge, or even define, the graph diameter for a sequence of time-varying networks.

Furthermore, this paper considers decentralized stochastic optimization where *the network size n is a fixed constant*. A fixed n is a prerequisite in distributed stochastic optimization which enables distributed algorithms to achieve the linear speedup in convergence rate $O(\sigma/\sqrt{nT})$ where σ indicates the gradient noise and T is the algorithmic iteration. In decentralized deterministic optimization, however, size n does not appear in the convergence rate. Thus, it does not need to be fixed and can be varied freely to simplify the lower-bound analysis. In fact, references [14, 30, 31] all tune n delicately to derive the optimal complexity for decentralized deterministic optimization over static or time-varying networks. Therefore, the analysis in [14, 30, 31] cannot be extended to decentralized stochastic setting in which the network size n is fixed.

Main results. This paper overcomes the above two challenges and successfully establishes the optimal complexity for decentralized stochastic optimization over time-varying network topologies.

- Inspired by the graph diameter of a static network topology, we introduce a novel *effective graph diameter* to gauge how efficient a message is transmitted between two farthest nodes via a sequence of time-varying decentralized communications.
- We provide the first lower bound complexity for decentralized non-convex stochastic optimization over time-varying networks. The derivation of this lower bound is based on a novel family of *sun-shaped network topologies*. Given any fixed network size n, we can always construct a sequence of time-varying sun-shaped topologies that maintains the optimal relation between the effective graph diameter and the network connectivity.
- We prove that the established lower bound complexity can be nearly attained (up to logarithmic factors) by integrating multiple gossip communications [19, 28, 45] and gradient accumulation [21, 27, 30] to the vanilla stochastic gradient tracking approach [8, 20, 24, 26, 41]. It implies that our complexity bound is tight and the proposed algorithm is nearly optimal.

All established results in this paper as well as those of existing state-of-the-art decentralized learning algorithms over time-varying networks are listed in Table 1.

Table 1: Rate comparison between different decentralized stochastic algorithms over time-varying networks. Parameter n denotes the number of all computing nodes, $\beta \in [0,1)$ denotes the connectivity measure of the weight matrix, σ^2 measures the gradient noise, b^2 denotes data heterogeneity, and T is the number of iterations. Other constants such as the initialization $f(x^{(0)}) - f^\star$ and smoothness constant L are omitted for clarity. Logarithm factors are hidden in the $\tilde{O}(\cdot)$ notation.

Bound type	Reference	Gossip matrix	Convergence rate
Lower	Theorem 4	$\beta \in [0, 1-\frac{1}{n}]$	$\Omega\big(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T(1-\beta)}\big)$
Upper	DSGD [12]	$\beta \in [0,1)$	$O\big(\frac{\sigma}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}} + \frac{b^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{2}{3}}} + \frac{1}{T(1-\beta)}\big)$
	DSGT [40]	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)} + \frac{1}{T(1-\beta)^2}\big)$
	MC-DSGT	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T(1-\beta)}\big)$

Other related works. Decentralized optimization can be tracked back to [36]. Decentralized gradient descent [16, 25, 44], diffusion [7, 29] and dual averaging [9] are early popular decentralized methods. Other advanced variants extend decentralized methods to data-heterogeneous scenarios [1, 13, 20, 35, 40], adaptive momentum settings [18, 23, 46], or asynchronous implementations [17]. When the network topology is time-varying, reference [14] establishes optimal convergence rate under the deterministic and strongly-convex setting. References [14, 15] develop decentralized methods with Nesterov acceleration to nearly achieve such optimal convergence rate. In the stochastic and non-convex setting, the convergence rate of decentralized SGD over general time-varying networks is clarified in [12]. Other references [32, 39, 42] study specific sparse and time-varying network topologies that can further save communication overheads in decentralized SGD. However, none of these works provides the optimal complexity for non-convex decentralized learning over time-varying networks.

2. Problem setup

Problem setup. Consider the following problem with a network of n computing nodes:

$$\min_{x\in\mathbb{R}^d} \; f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{where} \quad f_i(x) := \mathbb{E}_{\xi_i\sim D_i}[F(x;\xi_i)]. \tag{1}$$

Function $f_i(x)$ is local to node i, and random variable ξ_i denotes the local data that follows distribution D_i. Each local data distribution D_i can be different across all nodes.

Assumptions. The optimal convergence rate is established under the following assumptions.

- **Function class.** We let the function class $\mathcal{F}_L^\Delta$ denote the set of all functions satisfying the following assumption for any dimension $d \in \mathbb{N}_+$ and initialization point $x^{(0)} \in \mathbb{R}^d$.

  Assumption 1 (COST FUNCTIONS) *We assume each f_i has L-Lipschitz gradient,* i.e.*,*

  $$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|$$

  for all $i \in [n]$, $x, y \in \mathbb{R}^d$, and $f(x^{(0)}) - \inf_{x\in\mathbb{R}^d} f(x) \le \Delta$ with $f = \frac{1}{n}\sum_{i=1}^n f_i$.

- **Gradient oracle class.** We assume each worker i has access to its local gradient $\nabla f_i(x)$ via a stochastic gradient oracle $O_i(x;\zeta_i)$ subject to independent randomness ζ_i, *e.g.*, the mini-batch sampling $\zeta_i \triangleq \xi_i \sim D_i$. We further assume that the output $O_i(x,\zeta_i)$ is an unbiased estimator of the full-batch gradient $\nabla f_i(x)$ with a bounded variance. Formally, we let the stochastic gradient oracle class $\mathcal{O}_{\sigma^2}$ denote the set of all oracles O_i satisfying Assumption 2.

  Assumption 2 (GRADIENT STOCHASTICITY) *We assume local gradient oracle O_i satisfies*

  $$\mathbb{E}_{\zeta_i}[O_i(x;\zeta_i)] = \nabla f_i(x) \quad \text{and} \quad \mathbb{E}_{\zeta_i}[\|O_i(x;\zeta_i) - \nabla f_i(x)\|^2] \le \sigma^2$$

  for any $x \in \mathbb{R}^d$ and $i \in [n]$.

- **Decentralized communication.** Let $\mathcal{V} = [n]$ denote the set of n computing nodes. For any communication round $t \ge 0$, we assume nodes are connected through a time-varying communication network represented by a graph $G^t = (V, E^t)$, where $E^t \subseteq \{(j,i) \in \mathcal{V} \times \mathcal{V} : i \ne j\}$ is the set of links activated at round t. If a directed link $(j,i) \in E^t$, then node j can transmit information to node i at round t. In decentralized communication protocols, each node i can only receive messages with its immediate neighbors via links in E^t.

- **Weight matrix class.** To characterize the decentralized communication in algorithm development, we associate each time-varying communication graph G^t with a weight matrix W^t (also known as the gossip matrix [25, 44]). As in [14, 21, 47], we consider a sequence of time-varying weight matrices $\{W^t\}_{t=0}^{\infty} \subseteq \mathbb{R}^{n\times n}$ satisfying Assumption 3.

  Assumption 3 (WEIGHT MATRIX) *For any $t \ge 0$, $W^t = [w_{i,j}^t]_{i,j=1}^n$ satisfies*

  1. *if $(j,i) \notin E^t$ and $i \ne j$, then $w_{i,j}^t = 0$;*
  2. *$W^t \mathbb{1}_n = \mathbb{1}_n$ and $\mathbb{1}_n^\top W^t = \mathbb{1}_n^\top$ where $\mathbb{1}_n = [1,\ldots,1]^\top \in \mathbb{R}^n$;*
  3. *there exists a fixed constant $\beta \in [0,1)$ such that $\|W^t - \mathbb{1}_n\mathbb{1}_n^\top/n\|_2 \le \beta$.*

  Note that a weight matrix W^t satisfying Assumption 3 is not necessarily symmetric or positive semi-definite. The constant β is the *connectivity measure* that gauges how well the network topology G^t is connected. Constant $\beta \to 0$ (which implies $W^t \to \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$) indicates a well-connected topology while $\beta \to 1$ (which implies $W^t \to I$) indicates a poor connection. We let $\mathcal{W}_{n,\beta}$ denote the class of all weight matrices $W^t \in \mathbb{R}^{n\times n}$ satisfying Assumption 3.

- **Algorithm class.** We consider an algorithm A in which each node i assesses an unknown local function f_i via the *independent* stochastic gradient oracle $O_i(x;\,\zeta_i) \in \mathcal{O}_{\sigma^2}$. Each node i running algorithm A will maintain a local model copy $x_i^{(t)}$ at round t. We assume A to follow the partial averaging policy, *i.e.*, each node communicates at round t via protocol

  $$z_i = \sum_{j\in\mathcal{N}_i^t} w_{i,j}^t y_j, \quad \forall\, i \in [n]$$

  with some $W^t = [w_{i,j}^t]_{i,j=1}^n \in \mathcal{W}_{n,\beta}$ where y and z are the input and output of the communication protocol. In addition, we assume A to follow the zero-respecting policy [5, 6]. Informally speaking, the zero-respecting policy requires that the number of non-zero entries

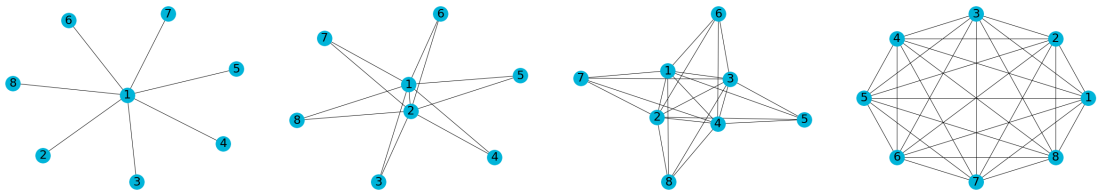

Figure 1: An illustration of the sun-shaped graph with size 8 and center sets $[1], [2], [4], [7]$ (or $[8]$). It is observed that $\mathcal{S}_{8,[1]}$ is a star graph while $\mathcal{S}_{8,[8]}$ is a complete graph.

of local model copy $x_i^{(t)}$ can only be increased by either sampling its own stochastic gradient oracle or interacting with the neighboring nodes. We let $\mathcal{A}_{\{W^t\}_{t=0}^{\infty}}$ be the set of all algorithms following the partial averaging and zero-respecting policies.

With the above classes, this paper will clarify the following question: *Given loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_L^{\Delta}$, stochastic gradient oracles $\{O_i\}_{i=1}^n \subseteq \mathcal{O}_{\sigma^2}$, a sequences of time-varying networks $\{G^t\}_{t=0}^{\infty}$ and its associated weight matrices $\{W^t\}_{t=0}^{\infty} \subseteq \mathcal{W}_{n,\beta}$, what is the optimal complexity to solve problem (1), and what decentralized algorithm $A \in \mathcal{A}_{\{W^t\}_{t=0}^{\infty}}$ can achieve it?*

Notations. We let $[n] := \{1, 2, \cdots, n\}$. For any network $G = ([n], E)$ and node $i \in [n]$, we let $\mathcal{N}_G(i)$ denote $\{j : (j, i) \in E \text{ or } j = i\}$, *i.e.*, the neighborhood set of node i in network G. Similarly, for a subset of nodes $\mathcal{I} \subseteq [n]$, we use $\mathcal{N}_G(\mathcal{I})$ to denote its neighborhood set $\cup_{i\in\mathcal{I}}\mathcal{N}_G(i)$.

3. Sun-shaped graphs and effective distance/diameter

As we have discussed in the **Challenge** paragraph in Section 1, it is unknown in literature (1) how to gauge the graph diameter for a sequence of time-varying network topologies, and (2) how to develop time-varying network topologies that can maintain the optimal relation between graph diameter and the network connectivity when the network size n is fixed. This section will resolve these two challenges by introducing a novel family of sun-shaped time-varying graphs.

Definition 1 (SUN-SHAPED GRAPH) *Given any positive integers $n \geq 2$ and $\mathcal{C} \subseteq [n]$, the sun-shaped graph over nodes $[n]$ with center set $\mathcal{C}$, denoted by $\mathcal{S}_{n,\mathcal{C}}$, is an undirected graph in which the neighborhood $\mathcal{N}_{\mathcal{S}_{n,\mathcal{C}}}(i)$ of node $i \in [n]$ is given by*

$$\mathcal{N}_{\mathcal{S}_{n,\mathcal{C}}}(i) = \begin{cases} [n] & \text{if } i \in \mathcal{C}; \\ \mathcal{C} \cup \{i\} & \text{otherwise.} \end{cases}$$

The center set $\mathcal{C}$ in $\mathcal{S}_{n,\mathcal{C}}$ constitutes a complete subgraph. Nodes in the complete set $[n]\backslash\mathcal{C}$ are connected to each node in $\mathcal{C}$, but there is no connection between any pair of nodes in $[n]\backslash\mathcal{C}$. Note that a sun-shaped graph $\mathcal{S}_{n,\mathcal{C}}$ with $|\mathcal{C}| = 1$ corresponds to a star graph while $|\mathcal{C}| = n$ or $|\mathcal{C}| = n - 1$ corresponds to a complete graph. $\mathcal{S}_{n,\mathcal{C}}$ can be regarded as an intermediate state between the star and complete graphs when $2 \leq |\mathcal{C}| \leq n - 2$, see the illustration in Figure 1.

We next introduce effective graph diameter to gauge how efficient a message is transmitted between two farthest nodes via a sequence of time-varying decentralized communications.

Definition 2 (EFFECTIVE DISTANCE/DIAMETER) *We define the* effective distance $\mathrm{dist}_{\{G^t\}_{t=0}^{\infty}}(i,j)$ *between two nodes $i \neq j$ over a sequence of networks $\{G^t\}_{t=0}^{\infty}$ to be the smallest number of rounds with which a message sent from node i or j at some round t can be received by the other one via decentralized communications (*i.e.*, communicating over $\{G^{t'}\}_{t'=t}^{\infty}$). Formally, we define*

$$\mathrm{dist}_{\{G^t\}_{t=0}^{\infty}}(i,j) := \max\Big\{ \arg\min_R\{R : j \in \mathcal{N}_{G^t}(\mathcal{N}_{G^{t+1}}(\cdots \mathcal{N}_{G^{t+R-1}}(i)\cdots)) \text{ for some } t \geq 0\},$$
$$\arg\min_R\{R : i \in \mathcal{N}_{G^t}(\mathcal{N}_{G^{t+1}}(\cdots \mathcal{N}_{G^{t+R-1}}(j)\cdots)) \text{ for some } t \geq 0\}\Big\}.$$

Similarly, we define the effective distance between two disjoint subsets of nodes $\mathcal{I}_1$, $\mathcal{I}_2 \subsetneq [n]$ as

$$\mathrm{dist}_{\{G^t\}_{t=0}^{\infty}}(\mathcal{I}_1,\mathcal{I}_2) = \min_{i\in\mathcal{I}_1,\, j\in\mathcal{I}_2}\{\mathrm{dist}_{\{G^t\}_{t=0}^{\infty}}(i,j)\}.$$

We define the effective diameter *to be the largest effective distance between any two nodes,* i.e.*,*

$$\mathrm{diam}_{\{G^t\}_{t=0}^{\infty}} := \max_{1\leq i\neq j\leq n}\{\mathrm{dist}_{\{G^t\}_{t=0}^{\infty}}(i,j)\}.$$

The definitions of effective distance and effective diameter are specific to the time-varying networks. We remark that when the networks are static, *i.e.*, $G^t = G$ for any $t \geq 0$, then the effective distance/diameter reduces to the canonical distance/diameter in a static graph.

The following fundamental theorem establishes the relation between the effective distance with respect to a sequence of sun-shaped graphs and the connectivity measure β.

Theorem 3 *Given a fixed $n \geq 2$, two disjoint subsets of nodes $\mathcal{I}_1$, $\mathcal{I}_2 \subsetneq [n]$, and any $\beta \in [0, 1-\frac{1}{n}]$, there exists a sequence of sun-shaped graphs $\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}$ such that*

(1) the graph $\mathcal{S}_{n,\mathcal{C}^t}$ at round t has an associated weight matrix $W^t \in \mathcal{W}_{n,\beta}$, i.e.*, $W^t \in \mathbb{R}^{n\times n}$, $\mathbb{1}_n^\top W^t = \mathbb{1}_n^\top$, $W^t\mathbb{1}_n = \mathbb{1}_n$, and $\|W^t - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top\|_2 \leq \beta$;*

(2) the effective distance between $\mathcal{I}_1$ and $\mathcal{I}_2$ satisfies

$$\mathrm{dist}_{\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}}(\mathcal{I}_1,\mathcal{I}_2) = \Theta\left(\frac{1-(|\mathcal{I}_1|+|\mathcal{I}_2|)/n}{1-\beta}+1\right);$$

In particular, if $1-(|\mathcal{I}_1|+|\mathcal{I}_2|)/n = \Omega(1)$, then $\mathrm{dist}_{\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}}(\mathcal{I}_1,\mathcal{I}_2) = \Theta((1-\beta)^{-1})$.

4. Lower Bound

With the help of Theorem 3, we are ready to establish the lower bound for non-convex decentralized stochastic optimization over time-varying networks. All proof details are in Appendix B.

Theorem 4 *For any $L > 0$, $n \geq 2$, $\beta \in [0, 1-\frac{1}{n}]$, and $\sigma > 0$, there exists a set of loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_L^{\Delta}$, a set of stochastic gradient oracles $\{O_i\}_{i=1}^n \subseteq \mathcal{O}_{\sigma^2}$, and a sequence of weight matrices $\{W^t\}_{t=0}^{\infty} \subseteq \mathcal{W}_{n,\beta}$ resulted from the sun-shaped graphs, such that it holds for the output $\hat{x}$ of any $A \in \mathcal{A}_{\{W^t\}_{t=0}^{\infty}}$ starting form $x^{(0)}$ that*

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] = \Omega\left(\left(\frac{\Delta L\sigma^2}{nT}\right)^{\frac{1}{2}} + \frac{\Delta L}{T(1-\beta)}\right). \tag{2}$$

Algorithm 1 Decentralized Stochastic Gradient Tracking with Multiple Consensus (MC-DSGT)

Input: Initialize $x_i^{(0)} = x^{(0)}$ and $h_i^{(0)} = \frac{1}{nR}\sum_{i=1}^n \sum_{r=0}^{R-1} O_i(x^{(k+1)}; \zeta_i^{(k+1,r)})$ for any $i \in [n]$; initialize $\mathbf{x}^{(0)} = [x_1^{(0)}, \cdots, x_n^{(0)}]^\top$, $\mathbf{h}^{(0)} = [h_1^{(0)}, \cdots, h_n^{(0)}]^\top$, and $\tilde{\mathbf{g}}^{(0)} = \mathbf{h}^{(0)}$; the decentralized gossip communication rounds R
for $k = 0, \cdots, K-1$ **do**
  Update $\mathbf{x}^{(k+1)} =$ **Multi-Consensus**$(\mathbf{x}^{(k)} - \gamma \mathbf{h}^{(k)}, 2kR, (2k+1)R)$
  Query stochastic gradients $\tilde{g}_i^{(k+1)} = \frac{1}{R}\sum_{r=0}^{R-1} O_i(x_i^{(k+1)}; \zeta_i^{(k+1,r)})$ at each node i
  Update $\mathbf{h}^{(k+1)} =$ **Multi-Consensus**$(\mathbf{h}^{(k)} + \tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)}, (2k+1)R, (2k+2)R)$
end for

Algorithm 2 $\mathbf{z}^{(t_2)}$ = Multi-Consensus$(\mathbf{z}^{(t_1)}, t_1, t_2)$

Input: Variable $\mathbf{z}^{(t_1)}$; index t_1 and t_2
for $t = t_1, \cdots, t_2 - 1$ **do**
  Update $\mathbf{z}^{(t+1)} = W^t \mathbf{z}^{(t)}$
end for
return Variable $\mathbf{z}^{(t_2)}$

Remark 5 *While the lower bound is established for $\beta \in [0, 1 - 1/n] \subset [0, 1)$, it approaches to $[0, 1)$ as n goes large. Such interval is broad enough to cover most weight matrices (generated through the Laplacian rule $W = I - L/d_{\max}$) resulted from common topologies such as grid, torus, hypercube, exponential graph, complete graph, Erdos-Renyi graph, geometric random graph, etc. whose β lies in the interval $[0, 1 - 1/n]$ when n is sufficiently large.*

5. Upper Bound

This section presents a decentralized algorithm that achieves the lower bound established in Theorem 4 up to logarithmic factors. The new algorithm is a direct extension of the vanilla decentralized stochastic gradient tracking (DSGT) [20, 40]. Inspired by the algorithm development in [14, 21], we add two additional components to DSGT: gradient accumulation and multiple-consensus communication. The main recursions are listed in Algorithm 1 which utilizes the fast gossip average step [19] in Algorithm 2. We call the new algorithm as MC-DSGT where "MC" indicates "multiple consensus". All proofs are in Appendix C. Since each node takes R gradient queries and R gossip communications at round k, it holds that $T = KR$ when MC-DSGT finishes after K rounds. The following theorems clarify the convergence rate of MC-DSGT where $T = KR$.

Theorem 6 *Given $L > 0$, $n \geq 1$, $\beta \in [0, 1)$, $\sigma > 0$, by choosing the learning rate γ as in (40), the convergence of Algorithm 1 can be bounded for any $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_L^\Delta$ and any $\{W\}_{t=0}^\infty \subseteq \mathcal{W}_{n,\beta}$ that*

$$\frac{1}{K+1}\sum_{k=0}^{K} \mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] = O\left(\left(\frac{\Delta L \sigma^2}{nT}\right)^{\frac{1}{2}} + \frac{R\Delta L}{T} + \left(\frac{\rho^2 \Delta^2 L^2 R \sigma^2}{(1-\rho)^3 T^2}\right)^{\frac{1}{3}} + \frac{\rho^2 R \Delta L}{T(1-\rho)^2}\right)$$

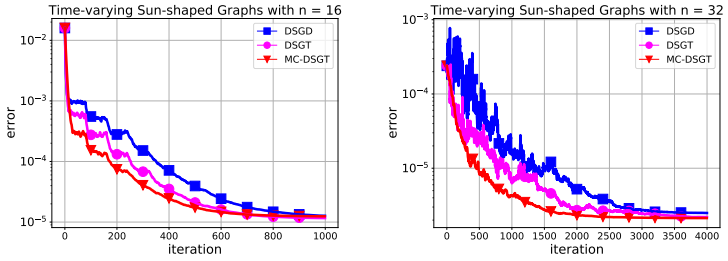

Figure 2: Performance of different stochastic algorithms to solve problem (4). The left plot is with MNIST, and the right plot is with COVTYPE.binary.

where $\rho \triangleq \beta^R \in [0,1)$, $\bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^n x_i^{(k)}$, *and* $T = KR$ *is the total number of gradient queries and gossip communications at each node. If we further set* R *as in* (41)*, then the rate becomes*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] = \tilde{O}\left(\left(\frac{\Delta L\sigma^2}{nT}\right)^{\frac{1}{2}} + \frac{\Delta L}{T(1-\beta)}\right). \tag{3}$$

The rate (3) matches with the lower bound (2) up to logarithm factors. Therefore, our established lower bound is tight and hence optimal. The comparison between MC-DSGT with other state-of-the-art algorithms for non-convex decentralized stochastic optimization is listed in Table 1.

6. Experiments

We consider the logistic regression with a non-convex regularization term [2, 40]. The problem formulation is given by $\min_x \frac{1}{n}\sum_{i=1}^n f_i(x) + \rho r(x)$ where

$$f_i(x) = \frac{1}{m}\sum_{j=1}^{m}\ln(1+\exp(-y_{i,j}\langle h_{i,j}, x\rangle)), \quad r(x) = \sum_{k=1}^{d}\frac{[x]_k^2}{1+[x]_k^2}, \tag{4}$$

$[x]_k$ denotes the k-the entry of $x \in \mathbb{R}^d$, $\{(h_{i,j}, y_{i,j})\}_{j=1}^m$ is the local dataset at node i where $h_{i,j} \in \mathbb{R}^d$, $y_{i,j} \in \{\pm 1\}$ is a feature vector and label, respectively. The regularization $r(x)$ is a smooth but non-convex function and $\rho > 0$ is the regularization weight.

We consider two real datasets: MNIST and COVTYPE.binary. We binarize MNIST dataset by considering datapoints with labels 2 and 4. The regularization weight ρ is chosen as 0.2 (MNIST) and 0.015 (COVTYPE.binary). We partition the two datasets non-uniformly such that a half of the nodes contain 80% positive datapoints while the other half hold 80% negative datapoints. We compare decentralized stochastic gradient descent (DSGD) [12], decentralized stochastic gradient tracking (DSGT) [40] and Algorithm 1 (MC-DSGT) with random time-varying sun-shaped graphs with $(n, |\mathcal{C}|)$ equal to $(16, 1)$ for MNIST and $(32, 4)$ for COVTYPE.binary. We set $R = 2$ and 4 in MC-DSGT for MNIST and COVTYPE.binary, repspectively

The performance of algorithms over MNIST and COVTYPE.binary is illustrated in the left and right plot in Figure 2, respectively. The error metric is taken as $\|\nabla f(\bar{x})\|^2$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i^{(k)}$. In both experiments, we find the convergence rate as well as the robustness to time-varying network topology of MC-DSGT outperforms DSGD and DSGT, which coincides with our theory.

7. Conclusion

This paper provides the first optimal complexity for non-convex decentralized stochastic optimization over time-varying networks. We also generalize DSGT with multiple consensus under time-varying networks to match the optimal bound up to logarithm factors. Future works include establishing the optimal rate for (strongly) convex stochastic scenarios over time-varying networks.

References

[1] Sulaiman A Alghunaim and Kun Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *arXiv preprint arXiv:2110.09993*, 2021.

[2] Anestis Antoniadis, Irène Gijbels, and Mila Nikolova. Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 2011.

[3] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake E. Woodworth. Lower bounds for non-convex stochastic optimization. *ArXiv*, abs/1912.02365, 2019.

[4] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning (ICML)*, pages 344–353, 2019.

[5] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

[6] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021.

[7] Jianshu Chen and Ali H Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.

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

[9] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[10] Shaoduo Gan, Jiawei Jiang, Binhang Yuan, Ce Zhang, Xiangru Lian, Rui Wang, Jianbin Chang, Chengjun Liu, Hongmei Shi, Shengzhuo Zhang, et al. Bagua: scaling up distributed learning with system relaxations. *Proceedings of the VLDB Endowment*, 15(4):804–813, 2021.

[11] Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[12] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, pages 1–12, 2020.

[13] Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[14] Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[15] Huan Li and Zhouchen Lin. Accelerated gradient tracking over time-varying graphs for decentralized optimization. *arXiv preprint arXiv:2104.02596*, 2021.

[16] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5330–5340, 2017.

[17] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 3043–3052, 2018.

[18] Tao Lin, Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *International Conference on Machine Learning*, 2021.

[19] Ji Liu and A Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

[20] Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321. IEEE, 2019.

[21] Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning (ICML)*, pages 7111–7123. PMLR, 2021.

[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[23] Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *arXiv preprint arXiv:1901.09109*, 2019.

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

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

[26] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

[27] Alexander Rogozin, Mikhail Mikhailovich Bochko, Pavel E. Dvurechensky, Alexander V. Gasnikov, and Vladislav Lukoshkin. An accelerated method for decentralized distributed stochastic optimization over time-varying graphs. *IEEE Conference on Decision and Control (CDC)*, 2021.

[28] Alexander Rogozin, Vladislav Lukoshkin, Alexander Gasnikov, Dmitry Kovalev, and Egor Shulgin. Towards accelerated rates for distributed optimization over time-varying networks. In *International Conference on Optimization and Applications*, 2021.

[29] Ali H Sayed. Adaptive networks. *Proceedings of the IEEE*, 102(4):460–497, 2014.

[30] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning (ICML)*, pages 3027–3036, 2017.

[31] Kevin Scaman, Francis Bach, Sébastien Bubeck, Laurent Massoulié, and Yin Tat Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2740–2749, 2018.

[32] Zhuoqing Song, Weijian Li, Kexin Jin, Lei Shi, Ming Yan, Wotao Yin, and Kun Yuan. A simple random consensus method with one-peer communication and o(1) rate. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[33] Sebastian Urban Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

[34] Haoran Sun and Mingyi Hong. Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms. *IEEE Transactions on Signal processing*, 67(22):5912–5928, 2019.

[35] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. d^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856, 2018.

[36] John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9):803–812, 1986.

[37] Sheng-Yuan Tu and Ali H Sayed. Foraging behavior of fish schools via diffusion adaptation. In *2010 2nd International Workshop on Cognitive Information Processing*, pages 63–68. IEEE, 2010.

[38] Sheng-Yuan Tu and Ali H. Sayed. Mobile adaptive networks. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):649–664, 2011. doi: 10.1109/JSTSP.2011.2125943.

[39] Jianyu Wang, Anit Kumar Sahu, Zhouyi Yang, Gauri Joshi, and Soummya Kar. MATCHA: Speeding up decentralized SGD via matching decomposition sampling. *arXiv preprint arXiv:1905.09435*, 2019.

[40] Ran Xin, Usman A Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 2020.

[41] Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *IEEE Conference on Decision and Control (CDC)*, pages 2055–2060, Osaka, Japan, 2015.

[42] Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[43] Bicheng Ying, Kun Yuan, Hanbin Hu, Yiming Chen, and Wotao Yin. Bluefog: Make decentralized algorithms practical for optimization and deep learning. *arXiv preprint arXiv:2111.04287*, 2021.

[44] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[45] Kun Yuan, Sulaiman A Alghunaim, and Xinmeng Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *arXiv preprint arXiv:2105.08023*, 2021.

[46] Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. DecentLaM: Decentralized momentum SGD for large-batch deep training. *International Conference on Computer Vision (ICCV)*, 2021.

[47] Kun Yuan, Xinmeng Huang, Yiming Chen, Xiaohan Zhang, Yingya Zhang, and Pan Pan. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Appendix A. Sun-shaped Graph

Proof [Proof of Theorem 3] It is easy to see that when $|\mathcal{I}_1|+|\mathcal{I}_2| = n$, $\text{dist}_{\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}}(\mathcal{I}_1, \mathcal{I}_2) = 1$ for any graphs $\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}$. Thus in this case, we can simply let $\mathcal{C}^t = [n]$ and $W^t = \beta I_n + (1-\beta)\mathbb{1}_n\mathbb{1}_n^\top$ for any $t \geq 0$. It is easy to see that $W^t \in \mathcal{W}_{n,\beta}$.

Next we consider $|\mathcal{I}_1| + |\mathcal{I}_2| < n$. Let $k = \lceil n(1-\beta) \rceil \in [1, n]$.

Case 1. If $k = n$, *i.e.*, $0 \leq \beta < \frac{1}{n}$, then we again let $\mathcal{C}^t = [n]$ with associate weight matrix $W^t = \beta I_n + (1-\beta)\mathbb{1}_n\mathbb{1}_n^\top$ for all $t \geq 0$. It is easy to see that

$$\text{dist}_{\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}}(\mathcal{I}_1, \mathcal{I}_2) = 1 = \Theta(1) = \Theta\left(\frac{1 - (|\mathcal{I}_1| + |\mathcal{I}_2|)/n}{1-\beta} + 1\right)$$

where the last identity is because $0 \leq 1 - (|\mathcal{I}_1| + |\mathcal{I}_2|)/n \leq 1$ and $(1-\beta)^{-1} = \Theta(1)$.

Case 2. If $1 \le k \le n-1$, then $\frac{1}{n} \le \beta \le 1-\frac{1}{n}$. Let $\mathcal{J}^0, \dots, \mathcal{J}^{p-1}$ with $p = \lfloor (n-|\mathcal{I}_1|-|\mathcal{I}_2|)/k \rfloor$ be disjoint subsets of $[n] \backslash (\mathcal{I}_1 \cup \mathcal{I}_2)$ such that each $\mathcal{J}^q$ $(0 \le q \le p-1)$ exactly contains k nodes. Such $\{\mathcal{J}^q\}_{q=0}^{p-1}$ always exists due to $p \times k \le n - |\mathcal{I}_1| - |\mathcal{I}_2|$. Now let $\mathcal{C}^t = \mathcal{J}^{t \bmod p}$, *i.e.*,

$$\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty} = \{\mathcal{S}_{n,\mathcal{J}^0}, \dots, \mathcal{S}_{n,\mathcal{J}^{p-1}}, \mathcal{S}_{n,\mathcal{J}^0} \dots, \mathcal{S}_{n,\mathcal{J}^{p-1}}, \mathcal{S}_{n,\mathcal{J}^0}, \dots\}.$$

It is easy to see that for any center set $\mathcal{C}$ with $|\mathcal{C}| = k$, the Laplacian $L(\mathcal{S}_{n,\mathcal{C}})$ of graph $\mathcal{S}_{n,\mathcal{C}}$ has eigenvalues:

$$0, \underbrace{k, \dots, k}_{(n-k-1)\text{-folds}}, \underbrace{n, \dots, n}_{k\text{-folds}}.$$

We thus let the associated weight matrices to be $W^t = I_n - \frac{\delta}{n} L(\mathcal{S}_{n,\mathcal{C}^t})$ with $\delta = n(1-\beta)/\lceil n(1-\beta) \rceil \in (0, 1]$ for any $t \ge 0$. Since $\delta < 1$, $\{W^t\}_{t=0}^{\infty}$ are positive semi-definite. Therefore, we have

$$\left\| W^t - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right\| = 1 - \frac{\delta k}{n} = 1 - \frac{n(1-\beta)}{n} = \beta.$$

The rest is to verify $\mathrm{dist}_{\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}}(\mathcal{I}_1, \mathcal{I}_2) = \Theta\left(\frac{1-(|\mathcal{I}_1|+|\mathcal{I}_2|)/n}{1-\beta} + 1\right)$. By the construction of sun-shaped graphs, starting from any round t, the neighborhood of $\mathcal{I}_1$ (or $\mathcal{I}_2$) satisfies

$$\mathcal{N}_{\mathcal{S}_{n,\mathcal{C}^t}}(\mathcal{N}_{\mathcal{S}_{n,\mathcal{C}^{t+1}}}(\cdots \mathcal{N}_{\mathcal{S}_{n,\mathcal{C}^{t+R-1}}}(\mathcal{I}_1) \cdots)) = \begin{cases} \left(\bigcup_{t'=t}^{t+R-1} \mathcal{J}^{t \bmod p}\right) \cup \mathcal{I}_1 & \text{if } R \le p; \\ [n] & \text{if } R > p+1. \end{cases}$$

Therefore, we conclude that

$$\mathrm{dist}_{\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}}(\mathcal{I}_1, \mathcal{I}_2) = p + 1 = \lfloor (n - |\mathcal{I}_1| - |\mathcal{I}_2|)/k \rfloor + 1 = \left\lfloor \frac{n - |\mathcal{I}_1| - |\mathcal{I}_2|}{\lceil n(1-\beta) \rceil} \right\rfloor + 1. \tag{5}$$

On one hand, we easily see

$$\left\lfloor \frac{n - |\mathcal{I}_1| - |\mathcal{I}_2|}{\lceil n(1-\beta) \rceil} \right\rfloor \le \frac{n - |\mathcal{I}_1| - |\mathcal{I}_2|}{n(1-\beta)}. \tag{6}$$

On the other hand, since $n(1-\beta) \ge 1$, we have $\lceil n(1-\beta) \rceil \le 2n(1-\beta)$ and further

$$\left\lfloor \frac{n - |\mathcal{I}_1| - |\mathcal{I}_2|}{\lceil n(1-\beta) \rceil} \right\rfloor + 1 \ge \left\lfloor \frac{n - |\mathcal{I}_1| - |\mathcal{I}_2|}{2n(1-\beta)} \right\rfloor + 1 = \Omega\left(\frac{n - |\mathcal{I}_1| - |\mathcal{I}_2|}{2n(1-\beta)} + 1\right) \tag{7}$$

where the last step is due to $\lfloor x \rfloor + 1 \ge (x+1)/2$ for any $x \ge 0$. Combining (6) and (7) with (5), we reach $\mathrm{dist}_{\{\mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^{\infty}}(\mathcal{I}_1, \mathcal{I}_2) = \Theta\left(\frac{1-(|\mathcal{I}_1|+|\mathcal{I}_2|)/n}{1-\beta} + 1\right)$. ■

Appendix B. Lower Bound

B.1. Proof of Theorem 4

Without loss of generality, we assume algorithms to start from $x^{(0)} = 0$. We denote the j-th coordinate of a vector $x \in \mathbb{R}^d$ by $[x]_j$ for $j = 1, \dots, d$, and let $\mathrm{prog}(x)$ be

$$\mathrm{prog}(x) := \begin{cases} 0 & \text{if } x = 0; \\ \max_{1 \le j \le d}\{j : [x]_j \ne 0\} & \text{otherwise.} \end{cases}$$

Similarly, for a set of points $\mathcal{X} = \{x_1, x_2, \dots\}$, we define $\text{prog}(\mathcal{X}) := \max_{x\in\mathcal{X}} \text{prog}(x)$. As described in [5, 6], a zero chain function f satisfies

$$\text{prog}(\nabla f(x)) \le \text{prog}(x) + 1, \quad \forall\, x \in \mathbb{R}^d,$$

which implies that, starting from $x = 0$, a single gradient evaluation can only make at most one more coordinate for the model parameter x be non-zero.

We prove the two terms of the lower bound in Theorem 4 separately by constructing two hard-to-optimize instances. We first state some key zero-chain functions that will be used to facilitate the analysis.

Lemma 7 (Lemma 2 of [3]) *Let $[x]_j$ denote the j-th coordinate of a vector $x \in \mathbb{R}^d$, and define function*

$$h(x) := -\psi(1)\phi([x]_1) + \sum_{j=1}^{d-1} \Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big)$$

where for $\forall\, z \in \mathbb{R}$,

$$\psi(z) = \begin{cases} 0 & z \le 1/2; \\ \exp\Big(1 - \frac{1}{(2z-1)^2}\Big) & z > 1/2, \end{cases} \quad \phi(z) = \sqrt{e}\int_{-\infty}^{z} e^{\frac{1}{2}t^2}\mathrm{d}t.$$

Then h satisfy the following properties:

1. *$h(x) - \inf_x h(x) \le \delta_0 d$, $\forall\, x \in \mathbb{R}^d$ with $\delta_0 = 12$;*
2. *h is ℓ_0-smooth with $\ell_0 = 152$;*
3. *$\|\nabla h(x)\|_\infty \le g_\infty$, $\forall\, x \in \mathbb{R}^d$ with $g_\infty = 23$;*
4. *$\|\nabla h(x)\|_\infty \ge 1$ for any $x \in \mathbb{R}^d$ with $[x]_d = 0$.*

Lemma 8 (Lemma 4 of [11]) *Let functions*

$$h_1(x) := -2\psi(1)\phi([x]_1) + 2\sum_{j \text{ even},\, 0<j<d} \Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big)$$

and

$$h_2(x) := 2\sum_{j \text{ odd},\, 0<j<d} \Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big).$$

Then h_1 and h_2 satisfy the following properties:

1. *$\frac{1}{2}(h_1 + h_2) = h$, where h is defined in Lemma 7.*
2. *For any $x \in \mathbb{R}^d$, if $\text{prog}(x)$ is odd, then $\text{prog}(\nabla h_1(x)) \le \text{prog}(x)$; if $\text{prog}(x)$ is even, then $\text{prog}(\nabla h_2(x)) \le \text{prog}(x)$.*
3. *h_1 and h_2 are also ℓ_0-smooth with $\ell_0 = 152$.*

Given Lemmas 7 and 8, we now construct two instances that lead to the two terms in lower bound (2), respectively.

Instance 1. The proof of the first term $\Omega((\frac{\Delta L\sigma^2}{nT})^{\frac{1}{2}})$ essentially follows the first example in proving Theorem 1 of [21]. We provide the key steps for the sake of being self-contained.

(Step 1.) Let $f_i = L\lambda^2 h(x/\lambda)/\ell_0, \forall\, i \in [n]$ be homogeneous and hence $f = L\lambda^2 h(x/\lambda)/\ell_0$ where h is defined in Lemma 7 and $\lambda > 0$ is to be specified. Since $\nabla^2 f_i = L\nabla^2 h/\ell_0$ and h is ℓ_0-smooth by Lemma 7, we know f_i is L-smooth for any $\lambda > 0$. By Lemma 7, we have

$$f(0) - \inf_x f(x) = \frac{L\lambda^2}{\ell_0^2}(h(0) - \inf_x h(x)) \leq \frac{L\lambda^2\delta_0 d}{\ell_0}.$$

Therefore, to ensure $f_i \in \mathcal{F}_L^\Delta$, it suffices to let

$$\frac{L\lambda^2\delta_0 d}{\ell_0} \leq \Delta, \quad \text{i.e.,} \quad d\lambda^2 \leq \frac{\ell_0\Delta}{L\delta_0}. \tag{8}$$

(Step 2.) We construct the stochastic gradient oracle O_i on worker $i, \forall\, i \in [n]$ as the follows:

$$[O_i(x;Z)]_j = [\nabla f_i(x)]_j \left(1 + \mathbb{1}\{j > \text{prog}(x)\}\left(\frac{Z}{p} - 1\right)\right), \forall\, x \in \mathbb{R}^d,\, j = 1, \ldots, d$$

with random variable $Z \sim \text{Bernoulii}(p)$ independent of x and f_i, and $p \in (0,1)$ to be specified. It is easy to see O_i is an unbiased stochastic gradient oracle. Moreover, since f_i is zero-chain, we have $\text{prog}(O_i(x;Z)) \leq \text{prog}(\nabla f_i(x)) \leq \text{prog}(x) + 1$ and hence

$$\begin{aligned}
\mathbb{E}[\|[O_i(x;Z)] - \nabla f_i(x)\|^2] &= |[\nabla f_i(x)]_{\text{prog}(x)+1}|^2 \mathbb{E}\left[\left(\frac{Z}{p} - 1\right)^2\right] = |[\nabla f_i(x)]_{\text{prog}(x)+1}|^2 \frac{1-p}{p} \\
&\leq \|\nabla f_i(x)\|_\infty^2 \frac{1-p}{p} \leq \frac{L^2\lambda^2(1-p)}{\ell_0^2 p}\|\nabla h(x)\|_\infty^2 \\
&\overset{\text{Lemma 7}}{\leq} \frac{L^2\lambda^2(1-p)g_\infty^2}{\ell_0^2 p}.
\end{aligned}$$

Therefore, to ensure $O_i \in \mathcal{O}_{\sigma^2}$, it suffices to let

$$p = \min\{\frac{L^2\lambda^2 g_\infty^2}{\ell_0^2\sigma^2}, 1\}. \tag{9}$$

(Step 3.) Let $x_i^{(t)}, \forall\, t \geq 0$ and $i \in [n]$, be the t-th query point of worker i. Since algorithms satisfy the zero-respecting property, as discussed in [5, 6, 21], within T gradient queries on each worker, algorithms can only return model $\hat{x}$ such that

$$\hat{x} \in \text{span}\left(\left\{x^{(0)}, \nabla f_i(x^{(0)}), \left\{\{x_i^{(t)}, \nabla f_i(x_i^{(t)}) : 0 \leq t < T\} : 1 \leq i \leq n\right\}\right\}\right),$$

which implies

$$\text{prog}(\hat{x}) \leq \max_{0\leq t<T} \max_{1\leq i\leq n} \text{prog}(x_i^{(t)}) + 1. \tag{10}$$

By Lemma 2 of [21], we have

$$\mathbb{P}(\text{prog}(\hat{x}) \geq d) \leq \mathbb{P}\left(\max_{0\leq t<T} \max_{1\leq i\leq n} \text{prog}(x_i^{(t)}) \geq d - 1\right) \leq e^{(e-1)npT - d + 1}. \tag{11}$$

On the other hand, when $\mathrm{prog}(\hat{x}) < d$, by Lemma 7, it holds that

$$\min_{\hat{x}\in\mathrm{span}\{\{x_i^{(t)}:1\le i\le n,\,0\le t<T\}\}} \|\nabla f(\hat{x})\| \ge \min_{[\hat{x}]_d=0} \|\nabla f(\hat{x})\| = \frac{L\lambda}{\ell_0} \min_{[\hat{x}]_d=0} \|\nabla h(\hat{x})\| \ge \frac{L\lambda}{\ell_0}. \tag{12}$$

Therefore, by combining (11) and (12), we have

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] \ge \mathbb{P}(\mathrm{prog}^{(T)} < d)\mathbb{E}[\|\nabla f(\hat{x})\|^2 \mid \mathrm{prog}^{(T)} < d] \ge (1 - e^{(e-1)npT-d+1})\frac{L^2\lambda^2}{\ell_0^2}. \tag{13}$$

Let

$$\lambda = \frac{\ell_0}{L}\left(\frac{\Delta L\sigma^2}{3nT\ell_0\delta_0 g_\infty^2}\right)^{\frac{1}{4}} \quad \text{and} \quad d = \left\lfloor \left(\frac{3L\Delta nTg_\infty^2}{\sigma^2\ell_0\delta_0}\right)^{\frac{1}{2}} \right\rfloor.$$

Then (8) naturally holds and $p = \min\{\frac{g_\infty^2}{\sigma^2}\left(\frac{\Delta L\sigma^2}{3nT\ell_0\delta_0 g_\infty^2}\right)^{\frac{1}{2}}, 1\}$ by (9). Without loss of generality, we assume $d \ge 2$, which is guaranteed when $T = \Omega(\frac{\sigma^2}{nL\Delta})$. Then, using the definition of p, we have that

$$\begin{aligned}(e-1)npT - d + 1 &\le (e-1)nT\,\frac{g_\infty^2}{\sigma^2}\left(\frac{\Delta L\sigma^2}{3nT\ell_0\delta_0 g_\infty^2}\right)^{\frac{1}{2}} - d + 1 \\ = \frac{e-1}{3}\left(\frac{3L\Delta nTg_\infty^2}{\sigma^2\ell_0\delta_0}\right)^{\frac{1}{2}} - d + 1 &< \frac{e-1}{3}(d+1) - d + 1 \le 2 - e < 0\end{aligned}$$

which, combined with (13), leads to

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] = \Omega\left(\frac{L^2\lambda^2}{\ell_0^2}\right) = \Omega\left(\left(\frac{\Delta L\sigma^2}{3nT\ell_0\delta_0 g_\infty^2}\right)^{\frac{1}{2}}\right) = \Omega\left(\left(\frac{\Delta L\sigma^2}{nT}\right)^{\frac{1}{2}}\right).$$

Instance 2. The proof for the second term $\Omega(c\Delta LT(1-\beta))$ utilizes weight matrices defined on the sun-shaped graphs described in Theorem 3.

(Step 1.) Let functions

$$\ell_1(x) := -\frac{n}{\lceil n/4\rceil}\psi(1)\phi([x]_1) + \frac{n}{\lceil n/4\rceil}\sum_{j\text{ even},\,0<j<d}\Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big)$$

and

$$\ell_2(x) := \frac{n}{\lceil n/4\rceil}\sum_{j\text{ odd},\,0<j<d}\Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big).$$

By Lemma 8, ℓ_1 and ℓ_2 defined here are $2\ell_0$-smooth. Furthermore, let

$$f_i = \begin{cases} L\lambda^2\ell_1(x/\lambda)/(2\ell_0) & \text{if } i\in\mathcal{I}_1 \triangleq \{j: 1\le j\le \lceil\frac{n}{4}\rceil\}, \\ L\lambda^2\ell_2(x/\lambda)/(2\ell_0) & \text{if } i\in\mathcal{I}_2 \triangleq \{j: n-\lceil\frac{n}{4}\rceil+1\le j\le n\}, \\ 0 & \text{else.}\end{cases}$$

where $\lambda > 0$ is to be specified. To ensure $f_i \in \mathcal{F}_L^\Delta$ for all $1\le i\le n$, it suffices to let

$$\frac{L\lambda^2\Delta_0 d}{2\ell_0} \le \Delta, \quad \textit{i.e.}, \quad d\lambda^2 \le \frac{2\ell_0\Delta}{L\Delta_0}. \tag{14}$$

With the functions defined above, we have $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = L\lambda^2\ell(x/\lambda)/(2\ell_0)$ and

$$\mathrm{prog}(\nabla f_i(x)) \begin{cases} = \mathrm{prog}(x)+1 & \text{if } \{\mathrm{prog}(x) \text{ is even and } i\in\mathcal{I}_1\}\cup\{\mathrm{prog}(x) \text{ is odd and } i\in\mathcal{I}_2\} \\ \le \mathrm{prog}(x) & \text{otherwise.}\end{cases}$$

Therefore, to make progress, *i.e.*, to increase $\mathrm{prog}(x)$, for any gossip algorithm A, it must take the gossip communications to transmit information between $\mathcal{I}_1$ to $\mathcal{I}_2$ alternatively. Namely, it takes at least $\mathrm{dist}_{\{G^t\}_{t=0}^\infty}(\mathcal{I}_1,\mathcal{I}_2)$ rounds of decentralized communications for any possible gossip algorithm A to increase $\mathrm{prog}(\hat{x})$ by 1. Therefore, we have

$$\mathrm{prog}(\hat{x}) \le \max_{1\le i\le n,\, 0\le t<T} \mathrm{prog}(x_i^{(t)}) \le \left\lfloor \frac{T}{\mathrm{dist}_{\{G^t\}_{t=0}^\infty}(\mathcal{I}_1,\mathcal{I}_2)} \right\rfloor + 1, \quad \forall\, T\ge 0. \tag{15}$$

(Step 2.) We consider a gradient oracle that return lossless full-batch gradients, i.e., $O_i(x) = \nabla f_i(x)$, $\forall\, x\in\mathbb{R}^d$, $i\in[n]$. For the construction of graphs and weight matrices, we consider the sequence of sun-shaped graphs $\{G^t := \mathcal{S}_{n,\mathcal{C}^t}\}_{t=0}^\infty$ and their associated weight matrices $\{W^t\}_{t=0}^\infty \in \mathcal{W}_{n,\beta}$ investigated in Theorem 3. Since $1-(|\mathcal{I}_1|+|\mathcal{I}_2|)/n = \Omega(1)$, by Theorem 3, we have $\mathrm{dist}_{\{G^t\}_{t=0}^\infty}(\mathcal{I}_1,\mathcal{I}_2) = \Theta((1-\beta)^{-1})$. Suppose $\mathrm{dist}_{\{G^t\}_{t=0}^\infty}(\mathcal{I}_1,\mathcal{I}_2) \ge 1/(C(1-\beta))$ with some absolute constant C, then by (15), we have

$$\mathrm{prog}(\hat{x}) \le \lfloor C(1-\beta)T\rfloor + 1, \quad \forall\, T\ge 0. \tag{16}$$

(Step 3.) We finally show the error $\mathbb{E}[\|\nabla f(x)\|^2]$ is lower bounded by $\Omega\left(\frac{\Delta L}{(1-\beta)T}\right)$, with any algorithm $A\in\mathcal{A}_{\{W^t\}_{t=0}^\infty}$. For any $T\ge 1/(C(1-\beta)) = \Omega((1-\beta)^{-1})$, let

$$d = \lfloor C(1-\beta)T\rfloor + 2 < 3C(1-\beta)T$$

and

$$\lambda = \frac{L_0}{L}\sqrt{\frac{2\Delta L}{3C(1-\beta)TL_0\Delta_0}}. \tag{17}$$

Then (14) naturally holds. Since $\mathrm{prog}(\hat{x}) < d$ by (16), following (12) and using (17), we have

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] \ge \min_{[\hat{x}]_d=0} \|\nabla f(\hat{x})\|^2 \ge \frac{L^2\lambda^2}{L_0^2} = \Omega\left(\frac{\Delta L}{(1-\beta)T}\right).$$

Appendix C. Upper Bound

C.1. Preliminary

Notation. We first introduce necessary notations as follows.

- $\mathbf{x}^{(k)} = [(x_1^{(k)})^\top;(x_2^{(k)})^\top;\cdots;(x_n^{(k)})^\top]\in\mathbb{R}^{n\times d}$;
- $\tilde{\mathbf{g}}^{(k)} \triangleq \nabla F(\mathbf{x}^{(k)};\boldsymbol{\xi}^{(k,r)}) = [\nabla F_1(x_1^{(k)};\xi_1^{(k,r)})^\top;\cdots;\nabla F_n(x_n^{(k)};\xi_n^{(k,r)})^\top]\in\mathbb{R}^{n\times d}$;
- $\nabla f(\mathbf{x}^{(k)}) = [\nabla f_1(x_1^{(k)})^\top;\nabla f_2(x_2^{(k)})^\top;\cdots;\nabla f_n(x_n^{(k)})^\top]\in\mathbb{R}^{n\times d}$;

- $\bar{\mathbf{x}}^{(k)} = [(\bar{x}^{(k)})^\top; (\bar{x}^{(k)})^\top; \cdots; (\bar{x}^{(k)})^\top] \in \mathbb{R}^{n\times d}$ where $\bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^n x_i^{(k)}$;
- $W^t = [w_{i,j}^t] \in \mathbb{R}^{n\times n}$ is the weight matrix;
- $\mathbb{1}_n = [1, 1, \cdots, 1]^\top \in \mathbb{R}^n$;
- Given two matrices $\mathbf{x}, \mathbf{h} \in \mathbb{R}^{n\times d}$, we define inner product $\langle \mathbf{x}, \mathbf{h}\rangle = \mathrm{tr}(\mathbf{x}^T\mathbf{h})$ and the Frobenius norm $\|\mathbf{x}\|_F^2 = \langle \mathbf{x}, \mathbf{x}\rangle$;
- Given $W \in \mathbb{R}^{n\times n}$, we let $\|W\|_2 = \sigma_{\max}(W)$ where $\sigma_{\max}(\cdot)$ denote the maximum sigular value.

Smoothness. Since each $f_i(x)$ is assumed to be L-smooth, it holds that $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is also L-smooth. As a result, the following inequality holds for any $x, y \in \mathbb{R}^d$:

$$f_i(x) \le f_i(y) + \langle \nabla f_i(y), x - y\rangle + \frac{L}{2}\|x - y\|^2. \tag{18}$$

Gradient noise. For stochastic gradient oracles satisfying Assumption 2, by independence, it holds for any $k \ge 0$ and $R \ge 1$ that

$$\mathbb{E}[\|\tilde{g}_i^{(k)} - \nabla f_i(x_i^{(k)})\|^2] \le \frac{\sigma^2}{R} \quad \text{and} \quad \mathbb{E}\left[\left\|\bar{\tilde{g}}^{(k)} - \frac{1}{n}\sum_{i=1}^n \nabla f(x_i^{(k)})\right\|^2\right] \le \frac{\sigma^2}{nR} \tag{19}$$

where $\bar{\tilde{g}}^{(k)} \triangleq \frac{1}{n}\sum_{i=1}^n \tilde{g}_i^{(k)} = \frac{1}{nR}\sum_{i=1}^n\sum_{r=0}^{R-1} O_i(x_i^{(k)}; \zeta_i^{(k,r)})$.

Network weighting matrix. Since each weight matrix $W^t \in \mathcal{W}_{n,\beta}$, it holds that

$$\left\|W^t - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top\right\|_2 \le \beta. \tag{20}$$

Following (20), it holds for a sequence of weight matrices $W^{t_1}, \ldots, W^{t_2-1}$ that

$$\left\|\prod_{t=t_1}^{t_2-1} W^t - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top\right\|_2 \le \beta^{t_2-t_1}. \tag{21}$$

Therefore, when $t_2 - t_1$ grows, $\prod_{t=t_1}^{t_2-1} W^t$ exponentially converges to $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$.

Submultiplicativity of the Frobenius norm. For any matrix $W \in \mathbb{R}^{n\times n}$ and $\mathbf{z} \in \mathbb{R}^{n\times d}$, it holds that

$$\|W\mathbf{z}\|_F \le \|W\|_2\|\mathbf{z}\|_F. \tag{22}$$

To verify it, by letting z_j be the j-th row of $\mathbf{z}$, we have $\|W\mathbf{z}\|_F^2 = \sum_{j=1}^d \|Wz_j\|_2^2 \le \sum_{j=1}^d \|W\|_2^2\|z_j\|_2^2 = \|W\|_2^2\|\mathbf{z}\|_F^2$.

C.2. Proof of Theorem 6

Our proof is adapted from the proof of [40, Theorem 1], which presents the convergence rate of stochastic decentralized gradient tracking with single consensus operation and a static weight matrix. We generalize the proof to suit multiple consensus and time-varying weight matrices.

We use the matrix-form notations of the algorithm mostly for convenience. At the beginning of phase k, the three quantities of interests are $\mathbf{x}^{(k)}$, $\mathbf{h}^{(k)}$ and $\tilde{\mathbf{g}}^{(k)} \triangleq \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k,r)})$, and the update rule for any $k \geq 0$ is

$$\mathbf{x}^{(k+1)} = \mathbf{W}_R^{(2k)}(\mathbf{x}^{(k+1)} - \gamma \mathbf{h}^{(k)}), \tag{23}$$

$$\mathbf{h}^{(k+1)} = \mathbf{W}_R^{(2k+1)}(\mathbf{h}^{(k)} + \tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)}) \tag{24}$$

where $\mathbf{W}_R^{(k)} \triangleq \prod_{t=kR}^{(k+1)R-1} W^t$ for any $k \geq 0$ and $R \geq 1$. By (21), we have $\|\mathbf{W}_R^{(k)} - \mathbb{1}\mathbb{1}^\top/n\|_2 \leq \beta^R$ for any $k \geq 0$. By multiplying $\mathbb{1}_n\mathbb{1}_n^\top/n$ to the left-side of (23) and (24), we have

$$\begin{aligned} \bar{x}^{(k+1)} &= \bar{x}^{(k+1)} - \gamma \bar{h}^{(k)}), \\ \bar{h}^{(k+1)} &= \bar{h}^{(k)} + \bar{\tilde{g}}^{(k+1)} - \bar{\tilde{g}}^{(k)}. \end{aligned} \tag{25}$$

Since $\bar{h}^{(0)} = \bar{\tilde{g}}^{(0)}$, by iterating (25) over $0, \ldots, k-1$, it holds that $\bar{h}^{(k)} = \bar{\tilde{g}}^{(k)}$ for any $k \geq 0$. We will use the following descent lemma, which is adapted from [40, Lemma 3].

Lemma 9 (DESCENT LEMMA) *Under Assumption 1, 2, 3, if $0 < \gamma \leq \frac{1}{2L}$, then we have for any $k \geq 0$,*

$$\mathbb{E}[f(\bar{x}^{(k+1)})] \leq \mathbb{E}[f(\bar{x}^{(k)})] - \frac{\gamma}{2}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] - \frac{\gamma}{4}\mathbb{E}[\|\bar{g}^{(k)}\|^2] + \frac{\gamma L^2}{2n}\mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2] + \frac{\gamma^2 L \sigma^2}{2nR}.$$

where $\Pi \triangleq I - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$.

By iterating Lemma 9 over $k = 0, \ldots, K$, we obtain

$$\begin{aligned} &\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] \\ \leq &\frac{2\Delta}{\gamma(K+1)} + \frac{\gamma L \sigma^2}{nR} - \frac{1}{2(K+1)}\sum_{k=0}^{K}\mathbb{E}[\|\bar{g}^{(k)}\|^2] + \frac{L^2}{n(K+1)}\sum_{k=0}^{K}\mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2] \end{aligned} \tag{26}$$

where $\Delta \geq f(x^{(0)}) - \min_x f(x)$.

We next turn to bound the consensus error $\mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2]$, which relies on the following recursion bound of consensus errors.

Lemma 10 (RECURSION OF CONSENSUS ERROR) *Under Assumption 1, 2, 3, denoting $\rho \triangleq \beta^R$, it holds for $0 < \gamma \leq \frac{1-\rho^2}{24(1+\rho^2)L}$ that*

$$\mathbb{E}[\|\Pi \mathbf{x}^{(k+1)}\|_F^2] \leq \frac{2\rho^2}{1+\rho^2}\mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2] + \frac{2\gamma^2\rho^2}{1-\rho^2}\mathbb{E}[\|\Pi \mathbf{h}^{(k)}\|_F^2]$$

$$\mathbb{E}[\|\Pi \mathbf{h}^{(k+1)}\|_F^2] \leq \frac{36\rho^2 L^2}{1-\rho^2}\mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2] + \frac{2\rho^2}{1+\rho^2}\mathbb{E}[\|\Pi \mathbf{h}^{(k)}\|_F^2] + \frac{12n\gamma^2\rho^2 L^2}{1-\rho^2}\mathbb{E}[\|\bar{g}^{(k)}\|^2] + 6n\frac{\sigma^2}{R}.$$

Proof Multiplying Π to the left side of (23) and (24), we have

$$\Pi\mathbf{x}^{(k+1)} = \Pi\mathbf{W}_R^{(2k)}(\mathbf{x}^{(k+1)} - \gamma\mathbf{h}^{(k)}), \tag{27}$$
$$\Pi\mathbf{h}^{(k+1)} = \Pi\mathbf{W}_R^{(2k+1)}(\mathbf{h}^{(k)} + \tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)}). \tag{28}$$

Therefore, following (27), by using $\|\Pi\mathbf{W}_R^{(2k)}\mathbf{a}\|_F \leq \rho\|\Pi\mathbf{a}\|_F$ for any $\mathbf{a} \in \mathbb{R}^{n\times n}$ and $-\langle\mathbf{a},\mathbf{b}\rangle \leq \frac{1-\rho^2}{1+\rho^2}\|\mathbf{a}\|_F^2 + \frac{1+\rho^2}{1-\rho^2}\|\mathbf{b}\|_F^2$ for any $\mathbf{a},\mathbf{b} \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned}
\|\Pi\mathbf{x}^{(k+1)}\|_F^2 =& \|\Pi\mathbf{W}_R^{(2k)}\mathbf{x}^{(k)}\|_F^2 - 2\gamma\langle\Pi\mathbf{W}_R^{(2k)}\mathbf{x}^{(k)}, \Pi\mathbf{W}_R^{(2k)}\mathbf{h}^{(k)}\rangle_F + \gamma^2\|\Pi\mathbf{W}_R^{(2k)}\mathbf{h}^{(k)}\|_F^2 \\
\leq& \rho^2\|\Pi\mathbf{x}^{(k)}\|_F^2 + \frac{\rho^2(1-\rho^2)}{1+\rho^2}\|\Pi\mathbf{x}^{(k)}\|_F^2 + \frac{\gamma^2\rho^2(1+\rho^2)}{1-\rho^2}\|\Pi\mathbf{h}^{(k)}\|_F^2 + \gamma^2\rho^2\|\Pi\mathbf{h}^{(k)}\|_F^2 \\
=& \frac{2\rho^2}{1+\rho^2}\|\Pi\mathbf{x}^{(k)}\|_F^2 + \frac{2\gamma^2\rho^2}{1-\rho^2}\|\Pi\mathbf{h}^{(k)}\|_F^2.
\end{aligned}$$

Following (28), we can bound $\|\Pi\mathbf{h}^{(k+1)}\|_F^2$ as follows:

$$\begin{aligned}
\mathbb{E}[\|\Pi\mathbf{h}^{(k+1)}\|_F^2] =& \mathbb{E}[\|\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}\|_F^2] + 2\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)})\rangle_F] \\
&+ \mathbb{E}[\|\Pi\mathbf{W}_R^{(2k+1)}(\tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)})\|_F^2] \\
\leq& \rho^2\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|_F^2] + 2\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k+1)}) - \tilde{\mathbf{g}}^{(k)})\rangle_F] \\
&+ \rho^2\mathbb{E}[\|\Pi(\tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)})\|_F^2] \\
=& \rho^2\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|_F^2] + 2\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)})\rangle_F] \\
&+ 2\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)}))\rangle_F] \\
&+ \rho^2\mathbb{E}[\|\tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)}\|_F^2]
\end{aligned} \tag{29}$$

where the inequality follows $\|\Pi\mathbf{W}_R^{(2k)}\mathbf{a}\|_F \leq \rho\|\Pi\mathbf{a}\|_F \leq \rho\|\mathbf{a}\|_F$ and $\mathbb{E}[\tilde{\mathbf{g}}^{(k+1)} \mid \mathbf{h}^{(k)}, \tilde{\mathbf{g}}^{(k)}] = \nabla f(\mathbf{x}^{(k+1)})$. We next bound the terms in (29) one by one. By using the similar derivation to [40, Lemma 5], we can easily reach

$$\begin{aligned}
\mathbb{E}[\|\tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)}\|_F^2] =& \mathbb{E}[\|\nabla f(\mathbf{x}^{(k+1)}) - \tilde{\mathbf{g}}^{(k)}\|_F^2] + \mathbb{E}[\|\tilde{\mathbf{g}}^{(k+1)} - \nabla f(\mathbf{x}^{(k+1)})\|_F^2] \\
\leq& 2\mathbb{E}[\|\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)})\|_F^2] + 2\mathbb{E}[\|\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)}\|_F^2] + \frac{n\sigma^2}{R} \\
\leq& 2L^2\mathbb{E}[\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_F^2] + \frac{3n\sigma^2}{R}
\end{aligned} \tag{30}$$

and

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_F^2] \leq& 3\mathbb{E}[\|\Pi\mathbf{x}^{(k+1)}\|_F^2] + 3\mathbb{E}[\|\Pi\mathbf{x}^{(k)}\|_F^2] + 3\mathbb{E}[\|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|_F^2] \\
\leq& 3\mathbb{E}[\|\Pi\mathbf{x}^{(k+1)}\|_F^2] + 3\mathbb{E}[\|\Pi\mathbf{x}^{(k)}\|_F^2] + 3\gamma^2\mathbb{E}[\|\bar{\overline{\mathbf{g}}}^{(k)}\|_F^2] \\
\leq& 9\mathbb{E}[\|\Pi\mathbf{x}^{(k)}\|_F^2] + 6\gamma^2\rho^2\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|_F^2] + 3n\gamma^2\mathbb{E}[\|\bar{g}^{(k)}\|^2] + \frac{3\gamma^2\sigma^2}{R}
\end{aligned} \tag{31}$$

where we use $\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)} = -\gamma\bar{\tilde{\mathbf{g}}}^{(k)}$ and $\mathbb{E}[\|\bar{\tilde{\mathbf{g}}}^{(k)}\|_F^2] \leq \mathbb{E}[\|\bar{\mathbf{g}}^{(k)}\|_F^2] + \sigma^2/R = n\mathbb{E}[\|\bar{g}^{(k)}\|^2] + \sigma^2/R$. Combining (30) and (31) together, we reach

$$\begin{aligned}
&\mathbb{E}[\|\tilde{\mathbf{g}}^{(k+1)} - \tilde{\mathbf{g}}^{(k)}\|_F^2] \\
\leq& 18L^2\mathbb{E}[\|\Pi\mathbf{x}^{(k)}\|_F^2] + 12\gamma^2\rho^2L^2\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|_F^2] + 6n\gamma^2L^2\mathbb{E}[\|\bar{g}^{(k)}\|^2] + (3n + 6n\gamma^2L^2)\frac{\sigma^2}{R}. \quad (32)
\end{aligned}$$

We next turn to bound $\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)})\rangle_F]$ in (29). For any $k \geq 1$, since $\mathbf{h}^{(k)} = \mathbf{W}_R^{(2k-1)R}(\mathbf{h}^{(k-1)} + \tilde{\mathbf{g}}^{(k)} - \tilde{\mathbf{g}}^{(k-1)})$, $\mathbb{E}[\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)} \mid \mathbf{h}^{(k-1)}, \tilde{\mathbf{g}}^{(k-1)}] = 0$, we reach

$$\begin{aligned}
&\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)})\rangle_F] \\
=&\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{W}_R^{(2k-1)}\tilde{\mathbf{g}}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)})\rangle_F] \\
=&\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{W}_R^{(2k-1)}(\tilde{\mathbf{g}}^{(k)} - \nabla f(\mathbf{x}^{(k)})), \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)})\rangle_F].
\end{aligned}$$

Since

$$\left\|\left(\Pi\mathbf{W}_R^{(2k+1)}\mathbf{W}_R^{(2k-1)}\right)^\top\Pi\mathbf{W}_R^{(2k+1)}\right\|_2 = \left\|\left(\mathbf{W}_R^{(2k+1)}\mathbf{W}_R^{(2k-1)}\right)^\top\mathbf{W}_R^{(2k+1)} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top\right\|_2 \leq \rho^3,$$

we further have

$$\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)})\rangle_F] \leq \mathbb{E}[\|\nabla f(\mathbf{x}^{(k)}) - \tilde{\mathbf{g}}^{(k)}\|_F^2] \leq \frac{n\rho^2\sigma^2}{R}. \quad (33)$$

It is easy to see that (33) also holds for $k = 0$. We finally bound the last term $\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)}))\rangle_F]$ in (29). Since $\|\Pi\mathbf{W}_R^{(2k+1)}\mathbf{a}\|_F \leq \rho\|\mathbf{a}\|_F$ for any $\mathbf{a} \in \mathbb{R}^{n\times d}$, we have

$$\begin{aligned}
&\mathbb{E}[\langle\Pi\mathbf{W}_R^{(2k+1)}\mathbf{h}^{(k)}, \Pi\mathbf{W}_R^{(2k+1)}(\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)}))\rangle_F] \leq \rho^2L\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|_F\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_F] \\
\leq&\rho^2L\mathbb{E}\left[\|\Pi\mathbf{h}^{(k)}\|_F\left(\|\Pi\mathbf{x}^{(k+1)}\|_F + \|\Pi\mathbf{x}^{(k)}\|_F + \|\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)}\|_F\right)\right] \\
\leq&\rho^2L\mathbb{E}\left[\|\Pi\mathbf{h}^{(k)}\|_F\left(2\|\Pi\mathbf{x}^{(k)}\|_F + \gamma\rho\|\Pi\mathbf{h}^k\|_F + \gamma\|\bar{\tilde{\mathbf{g}}}^{(k)}\|_F\right)\right] \quad (34)
\end{aligned}$$

where we us $\|\Pi\mathbf{x}^{(k+1)}\|_F \leq \rho\|\Pi\mathbf{x}^{(k)}\|_F + \gamma\rho\|\Pi\mathbf{h}^{(k)}\|_F$ and $\bar{\mathbf{x}}^{(k+1)} - \bar{\mathbf{x}}^{(k)} = -\gamma\bar{\tilde{\mathbf{g}}}^{(k)}$ in the last inequality. By Young's inequality, we have for any $\eta_1, \eta_2 > 0$ that

$$\begin{aligned}
&\mathbb{E}[\rho\|\Pi\mathbf{h}^{(k)}\|_F\gamma\rho L\|\bar{\tilde{\mathbf{g}}}^{(k)}\|_F] \\
\leq&0.5\eta_1\rho^2\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|_F^2] + 0.5\eta_1^{-1}\gamma^2\rho^2L^2\mathbb{E}[\|\bar{\tilde{\mathbf{g}}}^{(k)}\|_F^2] \\
\leq&0.5\eta_1\rho^2\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|_F^2] + 0.5\eta_1^{-1}\gamma^2\rho^2L^2n\mathbb{E}[\|\bar{g}^{(k)}\|^2] + 0.5\eta_1^{-1}\gamma^2\rho^2L^2\frac{\sigma^2}{R} \quad (35)
\end{aligned}$$

and

$$2\mathbb{E}[\rho\|\Pi\mathbf{h}^{(k)}\|_F\rho L\|\Pi\mathbf{x}^{(k)}\|_F] \leq \eta_2\rho^2\mathbb{E}[\|\Pi\mathbf{h}^{(k)}\|] + \eta_2^{-1}\rho^2L^2\mathbb{E}[\|\Pi\mathbf{x}^{(k)}\|_F]. \quad (36)$$

Plugging (35) and (36) into (34), we have

$$\begin{aligned}&\mathbb{E}[\langle \Pi \mathbf{W}_R^{(2k+1)} \mathbf{h}^{(k)}, \Pi \mathbf{W}_R^{(2k+1)} (\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)})) \rangle_F] \\ \leq & \rho^2(\gamma\rho L + 0.5\eta_1 + \eta_2)\mathbb{E}[\|\Pi \mathbf{h}^{(k)}\|_F^2] + \eta_2^{-1}\rho^2 L^2 \mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2] \\ & + 0.5\eta_1^{-1}\gamma^2\rho^2 L^2 n \mathbb{E}[\|\bar{g}^{(k)}\|^2] + 0.5\eta_1^{-1}\gamma^2\rho^2 L^2 \frac{\sigma^2}{R}. \end{aligned} \tag{37}$$

Plugging (32), (33), and (37) into (29), we reach

$$\begin{aligned}\mathbb{E}[\|\Pi \mathbf{h}^{(k+1)}\|_F^2] \leq & \rho^2(1 + 12\gamma^2\rho^2 L^2 + 2\gamma\rho L + \eta_1 + 2\eta_2)\mathbb{E}[\|\Pi \mathbf{h}^{(k)}\|_F^2] \\ & + \rho^2 L^2 (18 + 2\eta_2^{-1}) \mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2 \\ & + n\gamma^2\rho^2 L^2 (6 + \eta_1^{-1}) \mathbb{E}[\|\bar{g}^{(k)}\|^2] + (5\rho^2 n + 2n\gamma^2\rho^2 L^2 + \eta_1^{-1}\gamma^2\rho^2 L^2) \frac{\sigma^2}{R}. \end{aligned} \tag{38}$$

Letting $\eta_1 = \frac{2(1-\rho^2)}{9(1+\rho^2)}$ and $\eta_2 = \frac{1-\rho^2}{9(1+\rho^2)}$, then it holds for any $0 \leq \gamma \leq \frac{1-\rho^2}{24(1+\rho^2)L}$ that

$$\begin{aligned}\rho^2(1 + 12\gamma^2\rho^2 L^2 + 2\gamma\rho L + \eta_1 + 2\eta_2) &\leq \frac{2\rho^2}{1+\rho^2} \\ \rho^2 L^2 (18 + 2\eta_2^{-1}) &\leq \frac{36\rho^2 L^2}{1-\rho^2} \\ n\gamma^2\rho^2 L^2 (6 + \eta_1^{-1}) &\leq \frac{12n\gamma^2\rho^2 L^2}{1-\rho^2} \\ 5\rho^2 n + 2n\gamma^2\rho^2 L^2 + \eta_1^{-1}\gamma^2\rho^2 L^2 &\leq 6n, \end{aligned}$$

which, combined with (38), leads to the conclusion. ■

Letting $a_k \triangleq [\frac{1}{n}\mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2], \frac{1}{nL^2}\mathbb{E}[\|\Pi \mathbf{h}^{(k)}\|_F^2]]^\top \in \mathbb{R}^2$, $b_k \triangleq [0, \frac{12\gamma^2\rho^2}{1-\rho^2}\mathbb{E}[\|\bar{g}^{(k)}\|_F^2] + \frac{6\sigma^2}{RL^2}]^\top \in \mathbb{R}^2$ for any $k \geq 0$, and

$$M \triangleq \begin{bmatrix} \frac{2\rho^2}{1+\rho^2} & \frac{2\gamma^2\rho^2 L^2}{1-\rho^2} \\ \frac{36\rho^2}{1-\rho^2} & \frac{2\rho^2}{1+\rho^2} \end{bmatrix},$$

by Lemma 10, it holds that

$$a_{k+1} \preceq M a_k + b_k$$

where $\preceq$ indicates entry-wise inequality. Since $\gamma \leq \frac{(1-\rho^2)^2}{9\rho^2(1+\rho^2)L}$, one can check that there exists $v_1, v_2 \geq 0$ such that $M[v_1, v_2]^\top \prec [v_1, v_2]^\top$. Therefore, by [40, Lemma 9], we have for any $k \geq 0$ that

$$\sum_{\ell=0}^{k} M^\ell \preceq (I_{2\times 2} - M)^{-1} \preceq \begin{bmatrix} \frac{9(1+\rho^2)}{1-\rho^2} & \frac{18\gamma^2\rho^2(1+\rho^2)^2 L^2}{(1-\rho^2)^3} \\ \frac{324\rho^2(1+\rho^2)^2}{(1-\rho^2)^3} & \frac{9(1+\rho^2)}{1-\rho^2} \end{bmatrix}.$$

Therefore, we reach

$$\begin{aligned}\sum_{k=0}^{K} a_k \preceq & \sum_{k=0}^{K}\left(M^k a_0+\sum_{\ell=0}^{k-1} M^\ell b_{k-1-\ell}\right)\\ \preceq & \sum_{k=0}^{\infty} M^k\left(a_0+\sum_{k=0}^{K-1} b_k\right)\\ \preceq & (I_{2\times 2}-M)^{-1}\left(a_0+\sum_{k=0}^{K-1} b_k\right).\end{aligned}$$

Since $a^{(0)}=0$ by our initialization, considering the first entry of the above, we have

$$\frac{L^2}{n(K+1)}\sum_{k=0}^{K}\mathbb{E}[\|\Pi \mathbf{x}^{(k)}\|_F^2] \le \frac{216\gamma^4\rho^4(1+\rho^2)^2L^4}{(1-\rho^2)^4(K+1)}\sum_{k=0}^{K}\mathbb{E}[\|\bar{g}^{(k)}\|_F^2]+\frac{108\gamma^2\rho^2(1+\rho^2)^2L^2\sigma^2}{(1-\rho^2)^3R}. \tag{39}$$

When $\gamma \le \frac{1-\rho^2}{5\rho\sqrt{1+\rho^2}L}$,

$$\frac{216\gamma^4\rho^4(1+\rho^2)^2L^4}{(1-\rho^2)^4(K+1)} \le \frac{1}{2(K+1)}.$$

Hence, plugging (39) into (26) yields

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] \le \frac{2\Delta}{\gamma(K+1)}+\frac{\gamma L\sigma^2}{nR}+\frac{108\gamma^2\rho^2(1+\rho^2)^2L^2\sigma^2}{(1-\rho^2)^3R}.$$

Plugging

$$\gamma = \min\left\{\frac{1}{2L}, \frac{1-\rho^2}{24(1+\rho^2)L}, \frac{(1-\rho^2)^2}{9\rho^2(1+\rho^2)L}, \frac{1-\rho^2}{5\rho\sqrt{1+\rho^2}L}, \left(\frac{(1-\rho^2)^3R\Delta}{108\rho^2(1+\rho^2)^2L^2\sigma^2(K+1)}\right)^{\frac{1}{3}}\right\} \tag{40}$$

$$= \Theta\left(\min\left\{\frac{1-\rho}{L}, \frac{(1-\rho)^2}{\rho^2 L}, \left(\frac{(1-\rho)^3R^2\Delta}{\rho^2L^2\sigma^2T}\right)^{\frac{1}{3}}\right\}\right)$$

and $T=KR$ into the above, we reach

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] = O\left(\left(\frac{\Delta L\sigma^2}{nT}\right)^{\frac{1}{2}}+\frac{R\Delta L}{T}+\left(\frac{\rho^2\Delta^2L^2R\sigma^2}{(1-\rho)^3T^2}\right)^{\frac{1}{3}}+\frac{\rho^2R\Delta L}{T(1-\rho)^2}\right).$$

Furthermore, if one set

$$R=\frac{1}{1-\beta}\max\left\{\ln(2), \ln\left(\frac{n^{\frac{3}{4}}L^{\frac{1}{4}}\Delta^{\frac{1}{4}}}{T^{\frac{1}{4}}(1-\beta)^{\frac{1}{2}}\sigma^{\frac{1}{2}}}\right)\right\} = \tilde{O}\left(\frac{1}{1-\beta}\right), \tag{41}$$

so that

$$\rho=\beta^R \le e^{-(1-\beta)R} \le \min\left\{\frac{1}{2}, \frac{T^{\frac{1}{4}}(1-\beta)^{\frac{1}{2}}\sigma^{\frac{1}{2}}}{n^{\frac{3}{4}}L^{\frac{1}{4}}\Delta^{\frac{1}{4}}}\right\},$$

then we obtain

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] = \tilde{O}\left(\left(\frac{\Delta L\sigma^2}{nT}\right)^{\frac{1}{2}} + \frac{\Delta L}{T(1-\beta)}\right).$$