Revisiting Optimal Convergence Rate for Smooth and Non-convex Stochastic Decentralized Optimization

Kun Yuan[1,3,*], **Xinmeng Huang**[2*], **Yiming Chen**[1,4*], **Xiaohan Zhang**[2], **Yingya Zhang**[1], **Pan Pan**[1]

[1]DAMO Academy, Alibaba Group [2]University of Pennsylvania
[3]Peking University [4]MetaCarbon

{kun.yuan, yingya.zyy, panpan.pp}@alibaba-inc.com
xinmengh@sas.upenn.edu yiming@metacarbon.vip zxiaohan@seas.upenn.edu

Abstract

Decentralized optimization is effective to save communication in large-scale machine learning. Although numerous algorithms have been proposed with theoretical guarantees and empirical successes, the performance limits in decentralized optimization, especially the influence of network topology and its associated weight matrix on the optimal convergence rate, have not been fully understood. While Lu and Sa [42] have recently provided an optimal rate for non-convex stochastic decentralized optimization with weight matrices defined over linear graphs, the optimal rate with *general* weight matrices remains unclear.

This paper revisits non-convex stochastic decentralized optimization and establishes an optimal convergence rate with general weight matrices. In addition, we also establish the optimal rate when non-convex loss functions further satisfy the Polyak-Lojasiewicz (PL) condition. Following existing lines of analysis in literature cannot achieve these results. Instead, we leverage the Ring-Lattice graph to admit general weight matrices while maintaining the optimal relation between the graph diameter and weight matrix connectivity. Lastly, we develop a new decentralized algorithm to nearly attain the above two optimal rates under additional mild conditions.

1 Introduction

1.1 Motivation. Decentralized optimization is an emerging paradigm for large-scale machine learning. By letting each node average with its neighbors, decentralized algorithms save great communication overhead compared to traditional approaches with a central server. While numerous effective decentralized algorithms (see, e.g., [49, 11, 60, 37, 5, 39, 76]) have been proposed in literature showing both theoretical guarantees and empirical successes, the performance limits in decentralized optimization have not been fully clarified. This paper provides further understandings in optimal convergence rate in decentralized optimization and develops algorithms to achieve these rates.

Weight matrix and connectivity measure. Assume n computing nodes are connected by some network topology (graph). We associate the topology with a weight matrix $W = [w_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ in which $w_{ij} \in (0,1)$ if node j is connected to node i otherwise $w_{ij} = 0$. We introduce $\beta := \|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\| \in (0,1)$, where $\mathbb{1}_n = (1,\dots,1)^T \in \mathbb{R}^n$ has all entries being 1, as the *connectivity measure* to gauge how well the network topology is connected. Quantity $\beta \to 0$ (which implies $W \to \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$) indicates a well-connected topology while $\beta \to 1$ (which implies $W \to I$) indicates a badly-connected topology. It is the weight matrix W and its connectivity measure β that bring major challenges to convergence analysis in decentralized optimization.

*Equal Contribution. Corresponding Author: Kun Yuan

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: Rate comparison between different algorithms in smooth and non-convex stochastic decentralized optimization. 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. Detailed complexities with constant L are listed in Table 6 in Appendix B. The definition of transient iteration complexity can be found in Remark 6 (the smaller the better). "LB" is lower bound while "UB" is upper bound. Notation $\tilde{O}(\cdot)$ hides all logarithm factors.

	References	**Gossip matrix**	**Convergence rate**	**Tran. iters.**
LB	[42]	$\beta = \cos(\pi/n)$	$\Omega\big(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T(1-\beta)^{\frac{1}{2}}}\big)$	$O\big(\frac{n}{(1-\beta)\sigma^2}\big)$
	Theorem 2	$\beta \in [0, \cos(\pi/n)]$	$\Omega\big(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T(1-\beta)^{\frac{1}{2}}}\big)$	$O\big(\frac{n}{(1-\beta)\sigma^2}\big)$
UB	DSGD [28]	$\beta \in [0,1)$	$O\big(\frac{\sigma}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}} + \frac{b^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{2}{3}}}\big)$	$O\big(\frac{n^3}{(1-\beta)^2\sigma^2}\big)^\dagger$
	D^2/ED [64]	$\beta \in [0,1)$	$O\big(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T(1-\beta)^3}\big)$	$O\big(\frac{n}{(1-\beta)^6\sigma^2}\big)$
	DSGT [27]	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}}\big)$	$\tilde{O}\big(\frac{n^3}{(1-\beta)^2\sigma^2}\big)$
	DeTAG [42]	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T(1-\beta)^{\frac{1}{2}}}\big)$	$\tilde{O}\big(\frac{n}{(1-\beta)\sigma^2}\big)$
	MG-DSGD	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T(1-\beta)^{\frac{1}{2}}}\big)$	$\tilde{O}\big(\frac{n}{(1-\beta)\sigma^2}\big)$

† The complete complexity is $O\big(\frac{n^3}{(1-\beta)^2 \min\{\sigma^2, (1-\beta)^2 b^2\}}\big)$, which reduces to $O\big(\frac{n^3}{(1-\beta)^2\sigma^2}\big)$ for small σ^2.

Decentralized optimization. Decentralized approaches are built upon the partial averaging $x_i^+ = \sum_{j\in\mathcal{N}_i} w_{ij}x_j$ (or $x^+ = Wx$ in a more compact manner) where $\mathcal{N}_i$ is the set of neighbors of node i (including node i itself). The sparsity and connectivity of W significantly influence the efficiency and effectiveness of the partial averaging. Decentralized gradient descent [49, 77], diffusion [11, 57] and dual averaging [18] are well-known decentralized methods. Other advanced algorithms, which can correct the bias caused by heterogeneous data distributions, include decentralized ADMM [43, 61], explicit bias-correction [60, 78, 36], gradient tracking [48, 16, 52, 72], and dual acceleration [58, 66].

In the stochastic regime, decentralize SGD [11, 57, 37, 28] and its momentum and adaptive variants [39, 76, 45] have attracted a lot of attentions. After being extended to directed topologies [5, 73], time-varying topologies [28, 46, 67], asynchronous settings [38], and data-heterogeneous scenarios [64, 71, 41, 2, 27], decentralize SGD has much more applicable scenarios with significantly improved performance. Decentralized algorithms are also closely related to federated learning methods [44, 62, 74, 26, 35] when they admit weight matrices alternating between $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$ and the identity matrix.

Prior understandings in theoretical limits. A series of pioneering works have shed lights on the theoretical limits in decentralized optimization. These works establish the optimal convergence rate for convex or non-stochastic decentralized optimization [58, 59, 63, 30], and propose decentralized algorithms that (nearly) match these optimal bounds [58, 59, 66, 63, 31, 30]. However, there are few studies on the theoretical limits in non-convex stochastic decentralized optimization, which is quite common in large-scale deep learning, even when the costs are smooth.

A recent novel work [42] establishes the optimal complexity for smooth and non-convex stochastic decentralized optimization. While this optimal bound is inspiring, it is valid for a restrictive family of weight matrix W associated with the linear graph whose connectivity measure $\beta = \cos(\pi/n)$[2]. However, the linear graph is just one of the commonly used topologies. A more difficult but fundamental question still remains open:

What is the optimal convergence rate of smooth and non-convex stochastic decentralized optimization with a general weight matrix (that does not necessarily satisfy $\beta = \cos(\pi/n)$)?

It is not known whether the lower bound in [42] still holds for other commonly-used weight matrices resulted from grids, toruses, hypercubes, etc. whose connectivity measure $\beta \neq \cos(\pi/n)$.

Challenges. Following existing lines of analysis, to our knowledge, cannot provide an assuring answer. It is known that the communication complexity in decentralized optimization is typically

[2] See Corollary 1 in the latest arXiv version of [42] uploaded in Jan. 2022: arXiv:2006.08085v4.

Table 2: Rate comparison between different algorithms in smooth and non-convex stochastic decentralized optimization under the PL condition. Constants such as the initialization $f(x^{(0)}) - f^\star$ and smoothness constant L are omitted for clarity. Detailed complexities with constants L and μ are listed in Table 6 in Appendix B.

	References	**Gossip matrix**	**Convergence rate**	**Tran. iters. complexity**
LB	Theorem 3	$\beta \in [0, \cos(\pi/n)]$	$\Omega\big(\frac{\sigma^2}{nT} + e^{-T(1-\beta)^{\frac{1}{2}}}\big)$	$\tilde{O}\big(\frac{1}{(1-\beta)^{1/2}}\big)$
UB	DSGD [28]†	$\beta \in [0, 1)$	$\tilde{O}\big(\frac{\sigma^2}{nT} + \frac{\sigma^2}{T^2(1-\beta)} + \frac{b^2}{T^2(1-\beta)^2}\big)$	$\tilde{O}\big(\frac{n}{(1-\beta)\min\{1,(1-\beta)b^2\}}\big)$
	DAGD[55]†‡	$\beta \in [0, 1)$	$\tilde{O}\big(\frac{\sigma^2}{nT} + e^{-T(1-\beta)^{\frac{1}{2}}}\big)$	$\tilde{O}\big(\frac{1}{(1-\beta)^{1/2}}\big)$
	D^2/ED [2, 75]	$\beta \in [0, 1)$	$\tilde{O}\big(\frac{\sigma^2}{nT} + \frac{\sigma^2}{T^2(1-\beta)} + e^{-T(1-\beta)}\big)$	$\tilde{O}\big(\frac{n}{1-\beta}\big)$
	DSGT [2]	$\beta \in [0, 1)$	$\tilde{O}\big(\frac{\sigma^2}{nT} + \frac{\sigma^2}{T^2(1-\beta)} + e^{-T(1-\beta)}\big)$	$\tilde{O}\big(\frac{n}{1-\beta}\big)$
	DSGT [71]	$\beta \in [0, 1)$	$\tilde{O}\big(\frac{\sigma^2}{nT} + \frac{\sigma^2}{T^2(1-\beta)^3} + e^{-T(1-\beta)}\big)$	$\tilde{O}\big(\frac{n}{(1-\beta)^3}\big)$
	MG-DSGD	$\beta \in [0, 1)$	$\tilde{O}\big(\frac{\sigma^2}{nT} + e^{-T(1-\beta)^{\frac{1}{2}}}\big)$	$\tilde{O}\big(\frac{1}{(1-\beta)^{1/2}}\big)$

† This rate is derived under the strongly-convex assumption.

‡ This rate is achieved by utilizing increasing (non-constant) mini-batch sizes.

proportional to the diameter D of the network topology [58]. Establishing the relation between D and β is key to clarifying the influence of the weight matrix on optimal convergence rate. Existing works [58, 59, 42] utilize the linear graph to establish the optimal relation as $D = \Theta(1/\sqrt{1-\beta})$. However, weight matrices associated with linear graph are quite *limited*. Given a fixed network size n (which is a prerequisite when discussing complexities in distributed stochastic optimization [42, 23, 29, 20]), they have to satisfy $\beta = \cos(\pi/n)$. Weight matrices with other β cannot be implied from the linear graph. To examine the lower bounds with a more general weight matrix W, [63] proposes a wider class of networks named linear-star graph, which, however, leads to *suboptimal* relation between D and β. As a result, a network topology that admits general weight matrices W while maintaining the optimal relation between D and β is critical to clarify the above fundamental question.

1.2 Main results. This paper successfully identifies the desired network topology which facilitates the theoretical limits over the general weight matrices. Our contributions are summarized as follows.

- We discover that ring-lattice graph maintains the optimal relation between D and β for a much broader class of weight matrices. Given any fixed n and $\beta \in [0, \cos(\pi/n)]$, we can always construct a weight matrix $W \in \mathbb{R}^{n\times n}$ associated with some ring-lattice graph, whose connectivity measure is β, that satisfies $D = \Theta(1/\sqrt{1-\beta})$. Since $\cos(\pi/n)$ approaches to 1 when n is large, the ring-lattice graph admits far broader weight matrices than linear graph used in [58, 59, 42].
- With the above results, we provide the optimal convergence rate for smooth and non-convex stochastic decentralized optimization that holds for any weight matrix with $\beta \in [0, \cos(\pi/n)]$. In contrast, the optimal rate in [42] only applies to weight matrix satisfying $\beta = \cos(\pi/n)$. We also provide the optimal convergence rate when the non-convex cost functions further satisfy the Polyak-Lojasiewicz (PL) condition [25] for any weight matrix with $\beta \in [0, \cos(\pi/n)]$.
- We prove that the above two optimal complexities can be nearly-attained (under mild assumptions and up to the condition number L/μ as well as other logarithmic factors) by simply integrating multiple gossip communication [40, 56] and gradient accumulation [58, 55, 42] to the vanilla decentralized SGD (D-SGD) algorithm. Compared to DeTAG [42], the extended D-SGD avoids the gradient tracking steps and saves half of the communication overheads. Compared to DAGD [55], the extended D-SGD achieves the near-optimality using constant mini-batch sizes and has theoretical guarantees in the non-convex (as well as the PL) scenario.

All established results in this paper as well as those of existing state-of-the-art decentralized algorithms are listed in Tables 1 and 2. The smoothness constant L and PL constant μ are omitted in these tables for clarity. The detailed complexities of these algorithms with constant L and μ are listed in Tables 6 and 7 in Appendix B.

1.3 Other related works. Some other related works are discussed as follows.

Lower bounds and optimal algorithms. The lower bounds and optimal algorithms for stochastic and convex optimization have been intensively studied in [33, 34, 53, 1, 17]. References [9, 10] established the lower bounds to find stationary solutions for non-convex optimization, and [20, 3] proposed algorithms with near-optimal convergence rate for non-convex finite-sum minimization. In deterministic decentralized optimization, the optimal convergence rate was established for strongly convex optimization [58], non-smooth convex optimization [59], and non-convex optimization [63]. Furthermore, the lower bounds and optimal algorithms over time-varying topologies are studied in [31, 55, 56]. In stochastic decentralized optimization, works [75] and [28] established the lower bounds for strongly-convex decentralized SGD. [42] proves the optimal complexities for non-convex problems when weight matrices satisfy $\beta = \cos(\pi/n)$.

Ring-lattice Graph. The ring-lattice graph was used in [68, 21, 12] as a basis model to generate the small world graph. However, its relation between graph diameter and connectivity measure is not clarified to our knowledge. A related work [70] examines the robustness of the ring-lattice graph, but does not discuss its connectivity measure. Furthermore, ring-lattice has never been utilized to establish the influence of the weight matrix on the optimal rate in decentralized optimization before.

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 assume each local loss function is L-smooth, i.e.,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|, \quad \forall x, y \in \mathbb{R}^n, \tag{2}$$

  for some constant $L > 0$. We let the class $\mathcal{F}_L$ denote the set of all functions satisfying (2). We may also assume the global loss $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ satisfies the μ-PL condition:

$$2\mu(f(x) - f^\star) \le \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^n, \tag{3}$$

  where $f^\star$ is the optimal function value in problem (1) and $\mu > 0$ is some constant. We let the function class $\mathcal{F}_{L,\mu}$ denote the set of all functions satisfying both (2) and (3).

- **Oracle class.** We assume each node i interacts with $f_i(x)$ via the stochastic gradient oracle $\tilde{g}_i(x,\xi)$ where ξ is a random variable. In addition, we assume oracle $\tilde{g}_i(x,\xi)$ satisfies

$$\mathbb{E}_\xi[\tilde{g}_i(x,\xi)] = \nabla f_i(x) \quad \text{and} \quad \mathbb{E}_\xi\|\tilde{g}_i(x,\xi) - \nabla f_i(x)\|^2 \le \sigma^2. \tag{4}$$

  We let $\mathcal{O}_{\sigma^2}$ denote the class of all such stochastic gradient oracles.

- **Weight matrix class.** We let $\mathcal{W}_{n,\beta}$ denote the class of all gossip matrices $W \in \mathbb{R}^{n\times n}$ satisfying

$$W\mathbb{1}_n = \mathbb{1}_n, \quad \mathbb{1}_n^T W = \mathbb{1}_n^T, \quad \|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\| = \beta \in [0,1), \tag{5}$$

  where $\mathbb{1}_n = (1,\dots,1)^T \in \mathbb{R}^n$ with all entries being 1. If W is positive semi-definite, then connectivity measure β is its second largest eigenvalue.

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

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

  for some $W = [w_{ij}]_{i,j=1}^n \in \mathcal{W}_{n,\beta}$ where $[n] := \{1,\cdots,n\}$, and y and z are the input and output variables of the communication protocol. In addition, we assume A to follow the zero-respecting

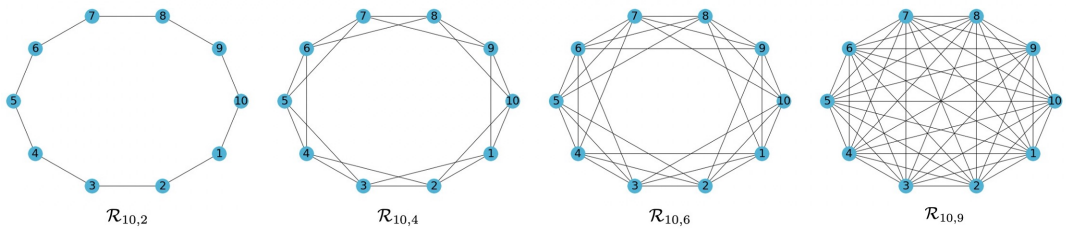

Figure 1: An illustration of the ring-lattice graph with size 10 and degrees $2, 4, 6, 9$. It is observed that $\mathcal{R}_{10,2}$ is a ring graph while $\mathcal{R}_{10,9}$ is a complete graph.

policy [9, 10, 24]. Informally speaking, the zero-respecting policy requires that the number of non-zero entries of local model copy $x_i^{(k)}$ can only be increased by sampling its own stochastic gradient oracle or interacting with the neighboring nodes. The zero-respecting policy is widely followed by all methods in Tables 1 and 2 and their momentum and adaptive variants. We let $\mathcal{A}_W$ be the set of all algorithms following the partial averaging and zero-respecting policies.

Lower bound metrics. With the above classes, we now introduce the measures for convergence rate's lower bound. Given loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_L$ (or $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}$), stochastic gradient oracles $\{\tilde{g}_i\}_{i=1}^n \subseteq \mathcal{O}_{\sigma^2}$, a weight matrix $W \in \mathcal{W}_{n,\beta}$, an algorithm $A \in \mathcal{A}_W$, the number of gradient queries (or communications rounds) T, we let $\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T}$ (which will be denoted as $\hat{x}$ when there is no ambiguity) be the output of algorithm A with $\hat{x} \in \mathrm{Span}\big(\{\{x_j^{(t)}\}_{j=1}^n\}_{t=1}^T\big)$. The lower bound for smooth and non-convex stochastic decentralized optimization is defined as

$$\inf_{A\in\mathcal{A}_W} \sup_{W\in\mathcal{W}_{n,\beta}} \sup_{\{\tilde{g}_i\}_{i=1}^n\in\mathcal{O}_{\sigma^2}} \sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_L} \mathbb{E}\|\nabla f(\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T})\|^2. \tag{6}$$

The lower bound under the PL condition is defined as

$$\inf_{A\in\mathcal{A}_W} \sup_{W\in\mathcal{W}_{n,\beta}} \sup_{\{\tilde{g}_i\}_{i=1}^n\in\mathcal{O}_{\sigma^2}} \sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_{L,\mu}} \mathbb{E}[f(\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T}) - f^\star]. \tag{7}$$

Fixed network size n. This paper considers *stochastic* decentralized 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 $O(\sigma/\sqrt{nT})$, see the algorithms listed in Tables 1 and 2. In decentralized deterministic optimization, however, size n does not appear in the convergence rate. Thus, it does not need to be specified and can be varied freely when establishing the lower bounds [58, 59].

3 Ring-Lattice graph

The main challenge in analyzing theoretical limits in decentralized optimization lies in clarifying the influence of network topology and its associated weight matrix. It is intuitive that the graph diameter D affects the convergence rate. In the worst case, one node needs to take at least D communication steps to transmit its own gradient information to the other. If the relation between D and β is established, the influence of the weight matrix on lower bounds will become clear. However, existing results either establishes suboptimal relation between D and β [63] or admits restrictive weight matrices [58, 59, 42] given fixed n. Now we show that the ring-lattice graph maintains the optimal relation between D and β for a broad class of weight matrices. All proof details are in Appendix C.

Definition 1 (Ring-Lattice Graph [69]). *Given any positive integers n, k such that k is even and $2 \le k < n-1$, the k-regular ring-lattice graph over nodes $\{1, \cdots, n\}$, denoted by $\mathcal{R}_{n,k}$, is an undirected graph in which each node $i \in [n]$ is connected with the set of nodes $\{(i+\ell) \bmod n : 1 \le |\ell| \le k/2\}$ where $(i+\ell) \bmod n$ is defined as*

$$(i+\ell) \bmod n = \begin{cases} i+\ell & \text{if } 1 \le i+\ell \le n; \\ i+\ell+n & \text{if } i+\ell < 1; \\ i+\ell-n & \text{if } i+\ell > n. \end{cases}$$

Apparently, each node in $\mathcal{R}_{n,k}$ has degree k.

Ring-lattice is a k-regular graph. When $k = 2$, ring-lattice will reduce to the ring graph. While Definition 1 excludes the case in which $k = n-1$, the complete graph can be exceptionally viewed as the ring lattice with degree $k = n-1$. Therefore, for general $2 \leq k \leq n-1$, $\mathcal{R}_{n,k}$ can be regarded as an intermediate state between the ring and complete graphs, see the illustration in Fig. 1. The following lemma clarifies the diameter of $\mathcal{R}_{n,k}$.

Lemma 1. *The diameter of the ring-lattice graph $\mathcal{R}_{n,k}$ is $D = \Theta(n/k)$.*

It is derived from the above lemma that the diameters of $\mathcal{R}_{n,2}$ and $\mathcal{R}_{n,n}$ are $\Theta(n)$ and $\Theta(1)$, respectively, which are consistent with the ring and complete graphs.

The following fundamental theorem establishes the relation between D and connectivity measure β.

Theorem 1. *Given a fixed n and any $\beta \in [0, \cos(\pi/n)]$, we can always construct a ring-lattice graph $\mathcal{R}_{n,k}$ so that*

(1) it has an associated weight matrix $W \in \mathcal{W}_{n,\beta}$, i.e., $W \in \mathbb{R}^{n\times n}$ and $\|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\| = \beta$,

(2) its diameter D satisfies $D = \Theta(1/\sqrt{1-\beta})$.

Remark 1. *The works [58, 59, 42] utilize the linear graph to establish the same relation $D = \Theta(1/\sqrt{1-\beta})$. However, given fixed network size n, the linear graph admits limited weight matrices with $\beta = \cos(\pi/n)$.*

Remark 2. *Sun and Hong [63] propose a slightly wider class of networks named linear-star graph, which, however, leads to a suboptimal relation $D = \Omega(1/[n(1-\beta)])$ [63, Proof of Corollary 3.1].*

Remark 3. *It is worth noting that for deterministic decentralized optimization where n may vary, one can tune n to make $\beta = \cos(\pi/n)$ lie in the full interval $(0, 1)$ as shown in [58, 59]. In other words, [58, 59] proved that $D = \Theta(1/\sqrt{1-\beta})$ for any $\beta \in (0, 1)$ if n is a varying parameter. However, this result is not valid in stochastic scenarios where n is required to be fixed.*

4 Lower bounds with general weight matrices

This section will establish the lower bounds for smooth and non-convex decentralized stochastic optimization. All proof details are in Appendix D. When the cost function does not satisfy the PL condition (3), we will derive that the convergence rate is lower bounded by $\Omega(\frac{\sigma}{\sqrt{nT}} + \frac{D}{T})$. This result together with Theorem 1 leads to:

Theorem 2. *For any $L > 0$, $n \geq 2$, $\beta \in [0, \cos(\pi/n)]$, and $\sigma > 0$, there exists a set of loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_L$, a set of stochastic gradient oracles $\{\tilde{g}_i\}_{i=1}^n \subseteq \mathcal{O}_{\sigma^2}$, and a weight matrix $W \in \mathcal{W}_{n,\beta}$, such that it holds for any $A \in \mathcal{A}_W$ starting form $x^{(0)}$ that*

$$\mathbb{E}\|\nabla f(\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T})\|^2 = \Omega\left(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{L}{T\sqrt{1-\beta}}\right). \tag{8}$$

Remark 4 (Comparison with existing bounds). *Comparing to the lower bound established in [42], Theorem 2 admits a much broader class of weight matrices with $\beta \in [0, \cos(\pi/n)]$. While the interval $[0, \cos(\pi/n)] \subset [0, 1)$, 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 line, ring, grid, torus, hypercube, exponential graph, complete graph, Erdos-Renyi graph, geometric random graph, etc. whose β lies in the interval $[0, \cos(\pi/n)]$ when n is sufficiently large, see details in Appendix D.*

Remark 5 (Linear speedup). *The first term $1/\sqrt{nT}$ dominates (8) when T is sufficiently large, which will require $T = O(1/(n\epsilon^2))$ (it is inversely proportional to n) iterations to reach accuracy ϵ. Therefore, if the term involving nT is dominating the rate for some T, we say the algorithm is in its linear-speedup stage. Linear speedup is a fundamental property that distributed algorithms enjoy.*

Remark 6 (Transient iteration complexity). *Due to the topology-incurred overhead in convergence rate, a decentralized algorithm has to experience a transient stage to achieve linear-speedup. Transient iterations are referred to those iterations before an algorithm reaches linear-speedup stage. To reach linear speedup in the lower bound (8), T has to satisfy $\frac{1}{T\sqrt{1-\beta}} \leq \frac{\sigma}{\sqrt{nT}}$ and hence $T = \mathcal{O}(\frac{n}{(1-\beta)\sigma^2})$, which is the transient iteration complexity shown in the first line in Table 1. The smaller the transient iterations are, the faster the algorithm can achieve linear speedup. Transient iteration complexity is an important metric to evaluate how sensitive the algorithm is to β.*

Algorithm 1: Decentralized SGD with multiple gossip steps (MG-DSGD)

Require: Initialize $x_i^{(0)} = 0$, the weight matrix W, and the rounds of gossip steps R.

for $k = 0, 1, 2, ..., K$, every node i **do**

Sample $\{\xi_i^{(k,r)}\}_{r=1}^R$ independently and let $g_i^{(k)} = \frac{1}{R}\sum_{r=1}^{R} \nabla F(x_i^{(k)}; \xi_i^{(k,r)})$; ▷ grad. accumulation

Update $\phi_i^{(k+1)} = x_i^{(k)} - \gamma g_i^{(k)}$; ▷ local stochastic gradient descent

Update $x_i^{(k+1)} =$ **FastGossipAverage**$(\{\phi_i^{(k+1)}\}_{i=1}^n, W, R)$; ▷ multiple gossips

Algorithm 2: $x_i =$ **FastGossipAverage**$(\{\phi_i\}_{i=1}^n, W, R)$

Require: $\{\phi_i\}_{i=1}^n$, W, R, and β; let $z_i^{(0)} = z_i^{(-1)} = \phi_i$ and $\eta = \frac{1-\sqrt{1-\beta^2}}{1+\sqrt{1+\beta^2}}$.

for $r = 0, 1, 2, ..., R-1$, every node i **do**

Update $z_i^{(r+1)} = (1+\eta)\sum_{j\in\mathcal{N}_i} w_{ij} z_i^{(r)} - \eta z_i^{(r-1)}$; ▷ fast gossip averaging

Output: $x_i = z_i^{(R)}$;

When the cost function satisfies PL condition (3), we prove the convergence rate is lower bounded by $\Omega(\frac{\sigma}{\sqrt{nT}} + \exp(-CT/D))$ where C is some constant. This result together with Theorem 1 leads to:

Theorem 3. *For any $L > 0$, $n \geq 2$, $\beta \in [0, \cos(\pi/n)]$, and $\sigma > 0$, there exists a set of loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}$, a set of stochastic gradient oracles $\{\tilde{g}_i\}_{i=1}^n \subseteq \mathcal{O}_{\sigma^2}$, and a weight matrix $W \in \mathcal{W}_{n,\beta}$, such that it holds for any $A \in \mathcal{A}_W$ starting from $x^{(0)}$ that*

$$\mathbb{E}[f(\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T}) - f^\star] = \Omega\left(\frac{\sigma^2}{\mu n T} + \frac{\mu}{L}\exp(-T\sqrt{\mu/L}\sqrt{1-\beta})\right). \tag{9}$$

Rate (9) *results in an $\tilde{\mathcal{O}}((1-\beta)^{-\frac{1}{2}})$ transient complexity.*

Expression (9) is the first lower bound for smooth and non-convex stochastic decentralized optimization under the PL condition to our knowledge. Please note that the above lower bound also applies to strongly-convex stochastic decentralized optimization, see the construction details in Appendix D.

5 DSGD with multiple gossip communications

This section will present a decentralized algorithm that achieve the lower bounds established in Sec. 4 up to constants L and μ. The new algorithm is a direct extension of the vanilla decentralized SGD (DSGD) [11, 28]. Inspired by the algorithm development in [55, 42], we add two additional components to DSGD: gradient accumulation and multiple-gossip communication. The main recursions are listed in Algorithm 1 which utilizes the fast gossip average step [40] in Algorithm 2. We call the new algorithm as MG-DSGD where "MG" indicates "multiple gossips". All proofs are in Appendix E.

The DeTAG algorithm proposed in [42] applies the gradient accumulation and multiple-gossip communication techniques to gradient tracking, which achieves the optimal rate in non-convex (without PL condition) stochastic decentralized optimization. This paper finds in MG-DSGD that the multiple gossip technique can significantly reduce the influence of data heterogeneity, and the gradient tracking step can be saved to achieve the convergence optimality. Similar observations also appeared in [55] albeit in the strongly-convex scenario. Removing gradient tracking from algorithm updates brings two benefits. First, it saves half of the communication overhead and memory storage per iteration. Second, it enables numerous effective acceleration techniques designed exclusively for decentralized SGD such as Quasi-global momentum [39], DecentLaM [76], adaptive momentum [45], etc. which cannot be easily extended to gradient tracking.

Let T be the toal number of gradient queries (or decentralized communications) at each node. Since each node takes R gradient queries and R communications at round k, it holds that $T = KR$ when MG-DSGD finishes after K rounds. The following theorems clarify the convergence rate of MG-DSGD where $T = KR$. We omit the initialization constants below. Our analysis techniques relate to the classic SGD analysis but with inaccurate oracles in the nonconvex case [19].

Theorem 4. *Given $L > 0$, $n \geq 2$, $\beta \in [0, 1)$, $\sigma > 0$, and let A denote Algorithm 1. Assuming that $\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq b^2$ for any $x \in \mathbb{R}^n$ and the number of gossip rounds R is set as in Appendix E, the convergence of A can be bounded for any $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_L$ and any $W \in \mathcal{W}_{n,\beta}$ that*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = \tilde{O}\left(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{L}{T\sqrt{1-\beta}}\right) \quad \textit{where} \quad \bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(k)}. \tag{10}$$

Remark 7. *The rate* (10) *matches with the lower bound* (8) *up to logarithm factors. Furthermore, we remark that the gradient similarity assumption $\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq b^2$ is not required to obtain the lower bound in Theorem 2. Due to these two restrictions, MG-DSGD is a nearly-optimal (not fully-optimal) algorithm. However, while the analysis in MG-DSGD needs to assume bounded gradient similarity b^2, it only affects the convergence rate as a logarithm term hidden in the $\tilde{\mathcal{O}}(\cdot)$ notation. If one wants to completely removes the gradient similarity assumption from Theorem 4, he/she can refer to the techniques in gradient tracking [42, 71, 41].*

The rate (10) matches with the lower bound (8) up to logarithm factors. The comparison between MG-DSGD with other state-of-the-art algorithms is listed in Table 1. MG-DSGD matches DeTAG in convergence rate and saves half of the communications, see experiments in Fig. 2. While the analysis in MG-DSGD needs to assume bounded data heterogeneity b^2, it only affects the convergence rate as a logarithm term hidden in the $\tilde{\mathcal{O}}(\cdot)$ notation.

Theorem 5. *Under the same assumptions as in Theorem 4 and setting R as in Appendix E, the convergence of A can be bounded for any loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}$, and any $W \in \mathcal{W}_{n,\beta}$ by*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[f(\bar{x}^{(k)}) - f^\star] = \tilde{O}\left(\frac{L\sigma^2}{\mu^2 nT} + \exp(-\mu T\sqrt{1-\beta}/L)\right). \tag{11}$$

Remark 8. *Comparing* (11) *and lower bound* (9)*, we find its dependence on σ^2, n, T, and $1-\beta$ is optimal while that on L and μ is not. One reason is that our algorithm does not involve any Nesterov acceleration [50] mechanism. The established upper bound in a recent work [55] is closer to the lower bound* (9) *by utilizing Nesterov acceleration, see Table 7 in Appendix B. However, it requires an increasingly large mini-batch of data per iteration, and only works for the strongly-convex scenario. It is not known whether the acceleration mechanisms in [55] can be extended to the non-convex and PL scenario. The bound* (11) *is state-of-the-art under the general PL condition to our knowledge.*

6 Experiments

This section will validate our theoretical results by empirically comparing different decentralized algorithms DSGD [37], D^2 [65], DSGT [80], DeTAG [42] and MG-DSGD in deep learning.

Implementation setup. We implement all decentralized algorithms with PyTorch [51] 1.6.0 using NCCL 2.8.3 (CUDA 10.1) as the communication backend, in which the partial averaging primitive is implemented by NCCL send/receive primitives. For parallel SGD, we used PyTorch's native Distributed Data Parallel (DDP) module. All the models and training scripts in this section run on servers with 8 NVIDIA V100 GPUs with each GPU treated as one node. Similar to existing works [42], we use Ring graph throughout all experiments.

Tasks and training details. A series of experiments are carried out with CIFAR-10 [32] and ImageNet [15] to compare the aforementioned methods. For CIFAR-10 dataset, it consists of 50,000 training images and 10,000 validation images in 10 classes. For ImageNet dataset, it consists of 1,281,167 training images and 50,000 validation images in 1000 classes. We utilize ResNet variants [22] on CIFAR-10 (ResNet-20 with 0.27M parameters and ResNet-18 with 11.17M parameters) and ResNet-50 on ImageNet. For training protocol, we train total 300 epochs and the learning rate is warmed up in the first 5 epochs and is decayed by a factor of 10 at 150 and 250-th epoch. For learning rate, we tuned a strong baseline in the PSGD setting (5e-3 for single node) and used the same setting in all decentralized methods. The batch size is set to 128 on each node. All CIFAR-10 experiments are repeated three times with different seeds. For fair comparison, we maintain the total number of gradient queries the same for algorithms in experiments, i.e., $T = KR$.

Performance with homogeneous data. Table 3 compares the proposed MG-DSGD with centralized parallel SGD and other baselines with homogeneous data distributions. For DeTAG and MG-DSGD, the communication round R is set as 2. We also performed an ablation study on the effect of

Table 3: Accuracy comparison on CIFAR-10 with homogeneous data (8 nodes).

METHODS	WITH MOMENTUM RESNET18	WITH MOMENTUM RESNET20	WITHOUT MOMENTUM RESNET18	WITHOUT MOMENTUM RESNET20
PSGD	93.99 ± 0.52	91.62 ± 0.13	93.22 ± 0.19	89.06 ± 0.08
DSGD	94.54 ± 0.05	91.46 ± 0.06	93.67 ± 0.04	89.14 ± 0.08
DSGT	94.34 ± 0.04	91.64 ± 0.01	93.57 ± 0.02	88.85 ± 0.11
D^2	92.51 ± 0.07	89.83 ± 0.08	92.18 ± 0.08	88.67 ± 0.11
DETAG	94.40 ± 0.13	**91.77 ± 0.25**	93.17 ± 0.18	88.97 ± 0.11
MG-DSGD	**94.57 ± 0.05**	**91.77 ± 0.09**	**93.75 ± 0.12**	**89.18 ± 0.08**

Table 4: Accuracy comparison on CIFAR-10 with heterogeneous data (ResNet-20).

METHODS	DSGD	DSGT	D^2	DETAG	MG-DSGD
$\alpha = 0.1$	58.14 ± 1.76	55.16 ± 0.42	53.85 ± 0.40	58.77 ± 1.30	**59.42 ± 3.36**
$\alpha = 1$	84.74 ± 0.97	84.74 ± 0.32	84.1 ± 0.64	84.82 ± 0.19	**84.91 ± 0.74**
$\alpha = 10$	90.19 ± 0.47	89.83 ± 0.12	89.80 ± 0.25	90.27 ± 0.08	**90.43 ± 0.04**

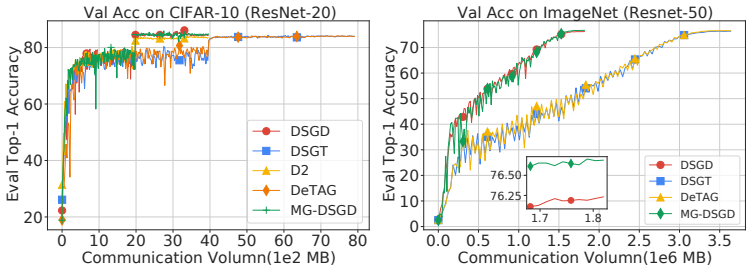

Figure 2: Accuracy in terms of the communication volumes. Left: CIFAR-10 with heterogeneous data ($\alpha = 1$); Right: ImageNet with homogeneous data.

momentum acceleration. Except for the ResNet-20 scenario with momentum in which MG-DSGD and DeTAG achieve the same performance, MG-DSGD consistently outperforms baseline methods in other scenarios. Moreover, MG-DSGD is more communication-efficient than DSGT and DeTAG.

Performance with heteogeneous data. Following previous works [79, 39], we simulate data heterogeneity among nodes with a Dirichlet distribution-based partitioning. A quantity α controls the degree of the data heterogeneity. The training data for a particular class tends to concentrate in a single node as $\alpha \to 0$, i.e. becoming more heterogeneous, while the homogeneous data distribution is achieved as $\alpha \to \infty$. We tested $\alpha = 0.1/1/10$ for all compared methods as corresponding to a setting with high/mild/low heterogeneity. As shown in Table 4, DeTAG and MG-DSGD outperform the other baselines by visible margins, and MG-DSGD is with slightly better accuracies. The left plot in Fig. 2 illustrates that MG-DSGD achieve a competitive accuracy to DeTAG using much less communication budgets on CIFAR-10 with $\alpha = 1$.

Large-batch training in ImageNet. To train ResNet-50 on ImageNet, utilization of the large-batch is necessary to speedup the training process. Since MG-DSGD is based on DSGD, we can extend the large-batch acceleration techniques introduced in [76, 39] to MG-DSGD to achieve better performance. These techniques are developed exclusively for DSGD and cannot be integrated to DeTAG easily. Following [76], we train total 120 epochs. The learning rate is warmed up in the first 20 epochs and is decayed in a cosine annealing scheduler. The results are listed in Table 5, showing that MG-DSGD has better accuracy than other approaches. Fig. 2(right) illustrates how different algorithms perform in terms of communication budgets used.

Table 5: Performance of different methods (ImageNet).

Methods	Acc.
DSGD	76.23 ± 0.06
DSGT	76.37 ± 0.07
DeTAG	76.57 ± 0.02
MG-DSGD	**76.64 ± 0.03**

More results. In Appendix F, we depict the accuracy curves in terms of epochs (rather than communication volumes). The ablation on the influence of inner loop R is also listed there.

7 Conclusion

This paper revisits the theoretical limits in smooth and non-convex stochastic decentralized optimization. Existing lower bound proved in [42] is only valid for special weight matrices with connectivity

measure $\beta = \cos(\pi/n)$. It is not known what the lower bound is when using weight matrices with other β. To address this issue, this paper establishes the lower bound for any weight matrix with $\beta \in [0, \cos(\pi/n)]$. We also clarify the lower bound when loss functions satisfy the PL condition. A new algorithm is proposed to match these lower bounds up to logarithm factors.

Acknowledgements

The authors are grateful to all anonymous reviewers in NeurIPS 2022 for the very helpful discussions during the rebuttal process as well as useful references on accelerated decentralized gradient descent.

References

[1] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 22, 2009.

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

[3] Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[4] 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.

[5] 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.

[6] Itai Benjamini, Gady Kozma, and Nicholas Wormald. The mixing time of the giant component of a random graph. *Random Structures & Algorithms*, 45(3):383–407, 2014.

[7] Stephen P Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Mixing times for random walks on geometric random graphs. In *ALENEX/ANALCO*, pages 240–249, 2005.

[8] Lucien Le Cam. Asymptotic methods in statistical decision theory. 1986.

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

[10] 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.

[11] 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.

[12] Yu Chen, Sanjeev Khanna, and Huani Li. On weighted graph sparsification by linear sketching. *ArXiv*, 2022.

[13] Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Accelerating gossip SGD with periodic global averaging. In *International Conference on Machine Learning (ICML)*, 2021.

[14] Imre Csiszár and János Körner. *Information Theory - Coding Theorems for Discrete Memoryless Systems, Second Edition*. Cambridge University Press, 2011.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[16] 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.

[17] Jelena Diakonikolas and Cristóbal Guzmán. Lower bounds for parallel and randomized convex optimization. In *Conference on Learning Theory*, pages 1132–1157. PMLR, 2019.

[18] 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.

[19] Pavel Dvurechensky. Gradient method with inexact oracle for composite non-convex optimization, 2017.

[20] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

[21] Brian Hayes. Computing science: Graph theory in practice: Part ii. *American Scientist*, 88(2):104–109, 2000.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[23] Samuel Horvath, D. Kovalev, Konstantin Mishchenko, Sebastian U. Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv*, 2019.

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

[25] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[26] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, pages 5132–5143. PMLR, 2020.

[27] 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.

[28] 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.

[29] 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*, 2020.

[30] 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.

[31] Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:18342–18352, 2020.

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[33] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[34] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1):167–215, 2018.

[35] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

[36] Z. Li, W. Shi, and M. Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, July 2019. early acces. Also available on arXiv:1704.07807.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

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

[43] Gonzalo Mateos, Juan Andrés Bazerque, and Georgios B Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276, 2010.

[44] 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.

[45] 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.

[46] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

[47] Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[48] 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.

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

[50] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

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

[53] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning (ICML)*, pages 1571–1578, 2012.

[54] Aaditya Ramdas and Aarti Singh. Optimal rates for stochastic convex optimization under tsybakov noise condition. In *International Conference on Machine Learning*, pages 365–373, 2013.

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

[60] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[61] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

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

[63] 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.

[64] 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.

[65] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, pages 6155–6165. PMLR, 2019.

[66] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. *Optimization Methods and Software*, pages 1–40, 2020.

[67] 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.

[68] Duncan J Watts. Small worlds: the dynamics of networks between order and randomness, 1999.

[69] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[70] Jun Wu, Mauricio Barahona, Yue-jin Tan, and Hong-zhong Deng. Robustness of random graphs based on graph spectra. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4):043101, 2012.

[71] 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.

[72] 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.

[73] 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.

[74] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

[75] 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.

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

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

[78] Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H. Sayed. Exact dffusion for distributed optimization and learning – Part I: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708 – 723, 2019.

[79] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR, 2019.

[80] Jiaqi Zhang and Keyou You. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Preliminary

Notation. We first introduce necessary notations as follows.

- $\mathbf{x}^{(k)} = [(x_1^{(k)})^T; (x_2^{(k)})^T; \cdots; (x_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $\nabla F(\mathbf{x}^{(k)}; \xi^{(k)}) = [\nabla F_1(x_1^{(k)}; \xi_1^{(k)})^T; \cdots; \nabla F_n(x_n^{(k)}; \xi_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $\nabla F(\mathbf{x}^{(k)}; \xi^{(k,r)}) = [\nabla F_1(x_1^{(k)}; \xi_1^{(k,r)})^T; \cdots; \nabla F_n(x_n^{(k)}; \xi_n^{(k,r)})^T] \in \mathbb{R}^{n\times d}$
- $\nabla f(\mathbf{x}^{(k)}) = [\nabla f_1(x_1^{(k)})^T; \nabla f_2(x_2^{(k)})^T; \cdots; \nabla f_n(x_n^{(k)})^T] \in \mathbb{R}^{n\times d}$
- $\bar{\mathbf{x}}^{(k)} = [(\bar{x}^{(k)})^T; (\bar{x}^{(k)})^T; \cdots; (\bar{x}^{(k)})^T] \in \mathbb{R}^{n\times d}$ where $\bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^n x_i^{(k)}$
- $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ is the weight matrix.
- $\mathbb{1}_n = \mathrm{col}\{1, 1, \cdots, 1\} \in \mathbb{R}^n$.
- Given two matrices $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n\times d}$, we define inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \mathrm{tr}(\mathbf{x}^T\mathbf{y})$ 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.

DSGD in matrix notation. The recursion of DSGD can be written in matrix notation:

$$\mathbf{x}^{(k+1)} = W\big(\mathbf{x}^{(k)} - \gamma\nabla F(\mathbf{x}^{(k)}; \xi^{(k)})\big) \tag{12}$$

MG-DSGD in matrix notation. The main recursion of MG-DSGD can be written in matrix notation:

$$\mathbf{g}^{(k)} = \frac{1}{R}\sum_{r=1}^{R} \nabla F(\mathbf{x}^{(k)}; \xi^{(k,r)}) \tag{13}$$

$$\mathbf{x}^{(k+1)} = \bar{M}(\mathbf{x}^{(k)} - \gamma\mathbf{g}^{(k)}) \tag{14}$$

where the weight matrix $\bar{M} = M^{(R)}$ and $M^{(R)}$ is achieved via the fast gossip averaging loop:

$$M^{(-1)} = M^{(0)} = I \tag{15}$$

$$M^{(r+1)} = (1+\eta)WM^{(r)} - \eta M^{(r-1)}, \quad \forall\, r = 0, 1, \cdots, R-1. \tag{16}$$

Smoothness. Since each $f_i(x) \in \mathcal{F}_L$ is 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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$:

$$f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}) - \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \le \langle \nabla f_i(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \tag{17}$$

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

$$\|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\|_2 = \beta. \tag{18}$$

Furthermore, one can establish the following upper result for the multi-gossip matrix $\bar{M}$.

Proposition 1 (Proposition 3 of [40])**.** *Let $M^{(r)}$ defined by iterations* (15) *and* (16)*, then it holds that for any $r = 0, \ldots, R$,*

$$M^{(r)}\mathbb{1}_n = [M^{(r)}]^T\mathbb{1}_n = \mathbb{1}_n, \quad \textit{and} \quad \|M^{(r)} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\|_2 \le \sqrt{2}\Big(1 - \sqrt{1-\beta}\Big)^r. \tag{19}$$

Therefore, when r grows, $M^{(r)}$ exponentially converges to $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$.

Submultiplicativity of the Frobenius norm. Given matrices $W \in \mathbb{R}^{n\times n}$ and $\mathbf{y} \in \mathbb{R}^{n\times d}$, it holds that

$$\|W\mathbf{y}\|_F \le \|W\|_2\|\mathbf{y}\|_F. \tag{20}$$

To verify it, by letting y_j be the j-th column of $\mathbf{y}$, we have $\|W\mathbf{y}\|_F^2 = \sum_{j=1}^d \|Wy_j\|_2^2 \le \sum_{j=1}^d \|W\|_2^2\|y_j\|_2^2 = \|W\|_2^2\|\mathbf{y}\|_F^2$.

B More detailed comparison

Table 6: Rate comparison in smooth and non-convex stochastic decentralized optimization. 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, L denotes smoothness constant, and T is the number of iterations. The definition of transient iteration complexity can be found in Remark 6 (the smaller the better). "LB" is lower bound while "UB" is upper bound. Notation $\tilde{O}(\cdot)$ hides all logarithm factors.

	References	Gossip matrix	Convergence rate	Tran. iters.
LB	[42]	$\beta = \cos(\pi/n)$	$\Omega\big(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{L}{T(1-\beta)^{\frac{1}{2}}}\big)$	$O\big(\frac{nL}{(1-\beta)\sigma^2}\big)$
	Theorem 2	$\boldsymbol{\beta \in [0, \cos(\pi/n)]}$	$\Omega\big(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{L}{T(1-\beta)^{\frac{1}{2}}}\big)$	$O\big(\frac{nL}{(1-\beta)\sigma^2}\big)$
UB	DSGD [28]	$\beta \in [0,1)$	$O\big(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}L^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}} + \frac{b^{\frac{2}{3}}L^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{2}{3}}}\big)$	$O\big(\frac{n^3L}{(1-\beta)^2\sigma^2}\big)^{\dagger}$
	D^2/ED [64]	$\beta \in [0,1)$	$O\big(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{L}{T(1-\beta)^3}\big)$	$O\big(\frac{nL}{(1-\beta)^6\sigma^2}\big)$
	DSGT [27]	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}L^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}}\big)$	$\tilde{O}\big(\frac{n^3L}{(1-\beta)^2\sigma^2}\big)$
	DeTAG [42]	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{L}{T(1-\beta)^{\frac{1}{2}}}\big)$	$\tilde{O}\big(\frac{nL}{(1-\beta)\sigma^2}\big)$
	MG-DSGD	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma\sqrt{L}}{\sqrt{nT}} + \frac{L}{T(1-\beta)^{\frac{1}{2}}}\big)$	$\tilde{O}\big(\frac{nL}{(1-\beta)\sigma^2}\big)$

† The complete complexity is $O\big(\frac{n^3L}{(1-\beta)^2\min\{\sigma^2,(1-\beta)^2b^2\}}\big)$, which reduces to $O\big(\frac{n^3L}{(1-\beta)^2\sigma^2}\big)$ for small σ^2.

Table 7: Rate comparison between different algorithms in smooth and non-convex stochastic decentralized optimization under the PL condition. Quantity μ is the PL constant.

	References	Gossip matrix	Convergence rate	Tran. iters.
LB	Theorem 3	$\beta \in [0, \cos(\pi/n)]$	$\Omega\big(\frac{\sigma^2}{\mu nT} + \frac{\mu}{L}e^{-T\sqrt{\mu(1-\beta)/L}}\big)$	$\tilde{O}\big(\frac{(L/\mu)^{1/2}}{(1-\beta)^{1/2}}\big)$
UB	DSGD [28]†	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma^2}{\mu nT} + \frac{L\sigma^2}{\mu^2T^2(1-\beta)} + \frac{Lb^2}{\mu^2T^2(1-\beta)^2}\big)$	$\tilde{O}\big(\frac{nL/\mu}{(1-\beta)\min\{1,(1-\beta)b^2\}}\big)$
	DAGD[55]†‡	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma^2}{\mu nT} + e^{-T\sqrt{\mu(1-\beta)/L}}\big)$	$\tilde{O}\big(\frac{(L/\mu)^{1/2}}{(1-\beta)^{1/2}}\big)$
	D^2/ED [75]†	$\beta \in [0,1)$	$\tilde{O}\big(\frac{\sigma^2}{\mu nT} + \frac{L\sigma^2}{\mu^2T^2(1-\beta)} + e^{-\mu T(1-\beta)/L}\big)$	$\tilde{O}\big(\frac{nL/\mu}{1-\beta}\big)$
	DSGT [2]	$\beta \in [0,1)$	$\tilde{O}\big(\frac{L\sigma^2}{\mu^2nT} + \frac{L^2\sigma^2}{\mu^3T^2(1-\beta)} + e^{-\mu T(1-\beta)/L}\big)$	$\tilde{O}\big(\frac{nL/\mu}{1-\beta}\big)$
	DSGT [71]	$\beta \in [0,1)$	$\tilde{O}\big(\frac{L\sigma^2}{\mu^2nT} + \frac{L^2\sigma^2}{\mu^3T^2(1-\beta)^3} + e^{-\mu T(1-\beta)/L}\big)$	$\tilde{O}\big(\frac{nL/\mu}{(1-\beta)^3}\big)$
	MG-DSGD	$\beta \in [0,1)$	$\tilde{O}\big(\frac{L\sigma^2}{\mu^2nT} + e^{-T\mu(1-\beta)^{\frac{1}{2}}/L}\big)$	$\tilde{O}\big(\frac{L/\mu}{(1-\beta)^{1/2}}\big)$

† These rates are derived under the strongly-convex assumption, not the general PL condition.
‡ This rate is achieved by utilizing increasing (non-constant) mini-batch sizes.

C Ring-Lattice graph

For any $2 \leq k < n-1$ and k is even, the ring-lattice graph $\mathcal{R}_{n,k}$ has several preferable properties. The following lemma estimates the order of the diameter of $\mathcal{R}_{n,k}$.

Lemma 1 (Formal version) *For any two nodes $1 \leq i < j \leq n$, the distance between i and j is $\lceil \frac{2\min\{j-i, i+n-j\}}{k} \rceil$. In particularly, the diameter of $\mathcal{R}_{n,k}$ is $D_{n,k} = \lceil \frac{2\lfloor n/2 \rfloor}{k} \rceil = \Theta(\frac{n}{k})$.*

Proof. By symmetry of $\mathcal{R}_{n,k}$, it suffices to consider the case that $j - i \leq i + n - j$. On the one side, the path $\{i, k/2+i, k+i, \ldots, (\lceil 2(j-i)/k \rceil - 1)k/2 + i, j\}$ connects the pair (i,j) with length $\lceil 2(j-i)/k \rceil$, which leads to $\mathrm{dist}(i,j) \leq \lceil 2(j-i)/k \rfloor$. On the other side, let $\{i_0 \triangleq$

$i, i_1, \ldots, i_{m-1}, i_m \triangleq j\}$ be one of the shortest paths connecting (i, j). Without loss of generality, we assume $i < i_1 < \cdots < i_{m-1} < j$. Since $|i_r - i_{r-1}| \leq k/2$ for any $r = 1, \ldots, m$ by the definition of $\mathcal{R}_{n,k}$, it holds that

$$\mathrm{dist}(i, j) = m \geq \frac{|i_m - i_0|}{k/2} = \frac{2(j-i)}{k}.$$

Since $\mathrm{dist}(i, j)$ is a integer, we reach $\mathrm{dist}(i, j) \geq \lceil \frac{2(j-i)}{k} \rceil$. The diameter is readily obtained by maximizing the expression of $\mathrm{dist}(i, j)$ with respect i and j. □

The following lemma clarifies the Laplacian matrix associated with $\mathcal{R}_{n,k}$ and its eigenvalues.

Lemma 2. *The Laplacian matrix of $\mathcal{R}_{n,k}$ is given by $L_{n,k} = kI - A_{n,k}$ with adjacency matrix $A_{n,k} = \sum_{\ell=1}^{k/2}(J^\ell + (J^T)^\ell)$ where*

$$J = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n} \quad \textit{and} \quad J^T = J^{-1} \in \mathbb{R}^{n\times n}.$$

Moreover, the eigenvalues of the Laplacian matrix $L_{n,k}$ are given by

$$\mu_j \triangleq k + 1 - \sum_{\ell=-k/2}^{k/2} \cos(\frac{2\pi(j-1)\ell}{n}) \in [0, 2k], \quad \forall\, j \in \{1 \ldots, n\} \tag{21}$$

where μ_j is the j-th eigenvalue of $L_{n,k}$. In addition, it holds that

$$(1 - \pi^2/12)\frac{\pi^2 k(k+1)(k+2)}{6n^2} \leq \min_{2\leq j\leq n} \mu_j \leq \frac{\pi^2 k(k+1)(k+2)}{6n^2}. \tag{22}$$

Proof. (*Laplacian matrix.*) Since every node in $\mathcal{R}_{n,k}$ is of degree k, the degree matrix of $\mathcal{R}_{n,k}$ is kI. The adjacency matrix $A_{n,k}$ can be easily achieved by following the construction of $\mathcal{R}_{n,k}$.

(*Eigenvalues.*) Let $\omega := e^{2\pi i/n}$ (where i is the imaginary number), then J can be decomposed as $J = U\mathrm{diag}(\omega^0 = 1, \omega, \ldots, \omega^{n-1})U^H$ with a unitary matrix $U = \frac{1}{\sqrt{n}}[\omega^{(p-1)(q-1)}]_{p,q=1}^n$. Since $J^\ell = U\mathrm{diag}(1, \omega^\ell, \ldots, \omega^{\ell(n-1)})U^H$ for all $-k/2 \leq \ell \leq k/2$, and $J^T = J^H = J^{-1} = U\mathrm{diag}(1, \bar{\omega} \ldots, \bar{\omega}^{n-1})U^H$ where $\bar{\omega} = e^{-2\pi i/n} = \omega^{-1}$, we have that

$$\begin{aligned} I + A_{n,k} &= \sum_{\ell=-k/2}^{k/2} J^\ell = U\left(\sum_{\ell=-k/2}^{k/2} \mathrm{diag}(1, \omega^\ell, \ldots, \omega^{\ell(n-1)})\right)U^H \\ &= U\left(\sum_{\ell=-k/2}^{k/2} \mathrm{diag}(1, \cos(\frac{2\pi\ell}{n}), \ldots, \cos(\frac{2\pi\ell(n-1)}{n})\right)U^H \\ &= U\mathrm{diag}\left(k+1, \sum_{\ell=-k/2}^{k/2} \cos(\frac{2\pi\ell}{n}), \ldots, \sum_{\ell=-k/2}^{k/2} \cos(\frac{2\pi\ell(n-1)}{n})\right)U^H. \end{aligned}$$

Therefore, the eigenvalues of $L_{n,k}$ are real numbers $\{\mu_j \triangleq k + 1 - \sum_{\ell=-k/2}^{k/2} \cos(\frac{2\pi(j-1)\ell}{n}) : j = 1, \ldots, n\}$. It is easy to verify that $\mu_j \geq 0$ and

$$\mu_j = k - \sum_{1\leq|\ell|\leq k/2,\, k\neq 0} \cos(\frac{2\pi(j-1)\ell}{n}) \leq 2k \quad \forall\, j = 1, \ldots, n.$$

Next we establish the bounds for $\min_{2\leq j\leq n} \mu_j$. For the upper bound, by the inequality $\cos(\theta) \geq 1 - \theta^2/2$ for any $\theta \in \mathbb{R}$, we have that

$$\min_{2\leq j\leq n} \mu_j \leq \mu_2 \leq (k+1) - \sum_{\ell=-k/2}^{k/2} (1 - \frac{2\pi^2\ell^2}{n^2}) = 2\pi^2 \sum_{\ell=-k/2}^{k/2} \frac{\ell^2}{n^2} = \frac{\pi^2 k(k+1)(k+2)}{6n^2}. \tag{23}$$

For the lower bound, by applying Lemma 3 (see below) with $h_k(\theta)$ defined as $\sum_{\ell=-k/2}^{k/2}\cos(2\ell\theta)$ and $\alpha=\frac{\pi}{n}$, we reach

$$\begin{aligned}\min_{2\le j\le n}\mu_j \ge& \inf_{\theta\in[\alpha,\pi-\alpha]}(k+1-h_k(\theta))\\ \ge& k+1-\max\left\{\underbrace{(k+1)\left(1-\frac{k(k+2)\pi^2/n^2}{1+(k+1)^2\pi^2/n^2}\right)^{\frac{1}{2}}}_{\text{I}},\underbrace{h_k(\pi/n)}_{\text{II}}\right\}\end{aligned}\tag{24}$$

For term I, using the inequality $(1-z)^{\frac{1}{2}}\le 1-\frac{1}{2}z$ for any $z\in[0,1]$, and $k+1\le n$, we have

$$\begin{aligned}\text{I}=&(k+1)\left(1-\frac{k(k+2)\pi^2/n^2}{1+(k+1)^2\pi^2/n^2}\right)^{\frac{1}{2}}\\ \le&(k+1)\left(1-\frac{k(k+2)\pi^2/n^2}{2(1+(k+1)^2\pi^2/n^2)}\right)\le(k+1)\left(1-\frac{\pi^2k(k+2)}{2(1+\pi^2)}\right).\end{aligned}\tag{25}$$

For term II, using the inequality $\cos(\theta)\le 1-\theta^2/2+\theta^4/24\le 1-\frac{12-\pi^2}{24}\theta^2$ for all $\theta\in[0,\pi]$, we have

$$\begin{aligned}\text{II}=&h_k(\pi/n)=\sum_{\ell=-k/2}^{k/2}\cos(\frac{2\pi\ell}{n})\le\sum_{\ell=-k/2}^{k/2}\left(1-\frac{(12-\pi^2)\pi^2}{6}\frac{\ell^2}{n^2}\right)\\ =&k+1-\frac{(12-\pi^2)\pi^2}{6}\sum_{\ell=-k/2}^{k/2}\frac{\ell^2}{n^2}=k+1-(1-\pi^2/12)\frac{\pi^2k(k+1)(k+2)}{6n^2}.\end{aligned}\tag{26}$$

Plugging (25) and (26) into (24), we reach (22). □

We also establish an auxiliary lemma to facilitate the results in Lemma 2.

Lemma 3. *For any $k\ge 1$ and k is even, let $h_k(\theta)=\sum_{\ell=-k/2}^{k/2}\cos(2\ell\theta)$ for any $\theta\in[\alpha,\pi-\alpha]$ with some $\alpha\in(0,\frac{\pi}{k+1}]$. It holds that*

$$\max_{\theta\in[\alpha,\pi-\alpha]}h_k(\theta)\le\max\left\{(k+1)\sqrt{\frac{1+\alpha^2}{1+(k+1)^2\alpha^2}},h_k(\alpha)\right\}\tag{27}$$

Proof. Since $|h_k(\pi/2)|=1$ while $(k+1)\sqrt{\frac{1+\alpha^2}{1+(k+1)^2\alpha^2}}>1$, we know (27) holds when $\theta=\pi/2$. Since $h_k(\pi-\theta)=h_k(\theta)$ by the definition of $h_k(\theta)$, it suffices to consider $\max_{\theta\in(\alpha,\pi/2)}|h_k(\theta)|$.

For any $\theta\in(\alpha,\pi/2)$, it holds that

$$\begin{aligned}h_k(\theta)=&\sum_{\ell=-k/2}^{k/2}\cos(2\ell\theta)=\frac{1}{2\sin(\theta)}\sum_{\ell=-k/2}^{k/2}2\cos(2\ell\theta)\sin(\theta)\\ =&\frac{1}{2\sin(\theta)}\sum_{\ell=-k/2}^{k/2}(\sin((2\ell+1)\theta)-\sin((2\ell-1)\theta))\\ =&\frac{\sin((k+1)\theta)}{\sin(\theta)}.\end{aligned}$$

Therefore, we have that

$$\begin{aligned}h_k'(\theta)=&\frac{(k+1)\cos((k+1)\theta)\sin(\theta)-\sin((k+1)\theta)\cos(\theta)}{\sin^2(\theta)}\\ =&\frac{\cos((k+1)\theta)\cos(\theta)}{\sin^2(\theta)}\left((k+1)\tan(\theta)-\tan((k+1)\theta)\right).\quad(\text{if }\cos((k+1)\theta)\cos(\theta)\ne 0)\end{aligned}$$

The special case that $\cos((k+1)\theta)\cos(\theta) = 0$ and $h_k'(\theta) = 0$ leads to that $\cos(\theta) = 0$, which is impossible for $z \in (\alpha, \pi/2)$. Therefore, all the extrema of h_k on $(\alpha, \pi/2)$ must satisfy

$$\tan((k+1)\theta) = (k+1)\tan(\theta). \tag{28}$$

By combining the cases of endpoints, i.e., $\theta = \alpha$ and $\theta = \pi/2$, with (28) for the interior extrema, we have that

$$\begin{aligned}
&\max_{\theta\in[\alpha,\pi/2]} h_k(\theta) = \max\{\sup_{\substack{\theta\in(\alpha,\pi/2)\\ \text{(28) holds}}} h_k(\theta), h_k(\alpha), h_k(\pi/2)\} \\
\leq &\max\{\sup_{\substack{\theta\in(\alpha,\pi/2)\\ \text{(28) holds}}} \sqrt{h_k(\theta)^2}, h_k(\alpha), 1\} \\
= &\max\{\sup_{\substack{\theta\in(\alpha,\pi/2)\\ \text{(28) holds}}} \left(\frac{\frac{\tan((k+1)\theta)^2}{1+\tan((k+1)\theta)^2}}{\frac{\tan(\theta)^2}{1+\tan(\theta)^2}}\right)^{\frac{1}{2}}, h_k(\alpha)^2, 1\} \\
\leq &\max\{\sup_{\substack{\theta\in(\alpha,\pi/2)\\ \text{(28) holds}}} \left(\frac{\frac{(k+1)^2\tan(\theta)^2}{1+(k+1)^2\tan(\theta)^2}}{\frac{\tan(\theta)^2}{1+\tan(\theta)^2}}\right)^{\frac{1}{2}}, h_k(\alpha), 1\} \\
\leq &\max\{(k+1)\sup_{\theta\in(\alpha,\pi/2)} \left(\frac{1+\tan(\theta)^2}{1+(k+1)^2\tan(\theta)^2}\right)^{\frac{1}{2}}, h_k(\alpha)\} \\
\leq &\max\{(k+1)\left(\frac{1+\alpha^2}{1+(k+1)^2\alpha^2}\right)^{\frac{1}{2}}, h_k(\alpha)\},
\end{aligned}$$

where the last inequality is because the function $\frac{1+z}{1+(k+1)^2 z}$ is decreasing with respect to $z \in [0, +\infty)$ and $\tan(\theta) \geq \tan(\alpha) \geq \alpha$. □

Theorem 1 *Given a fixed n and any $\beta \in [0, \cos(\pi/n)]$, there exists a ring-lattice graph with an associated weight matrix W such that*

(i) $W \in \mathcal{W}_{n,\beta}$, i.e., $W \in \mathbb{R}^{n\times n}$ and $\|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\| = \beta$;

(ii) its diameter D satisfies $D = \Theta(1/\sqrt{1-\beta})$.

Proof. We prove this theorem in two cases:

(Case 1: $0 \leq \beta \leq \min\{\cos(\pi/9), \cos(\pi/n)\}$.) In this case, we consider the complete graph whose diameter $D = 1$, and let $W = \frac{1-\beta}{n}\mathbb{1}_n\mathbb{1}_n^T + \beta I$, then it holds that $\|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\|_2 = \beta$. Furthermore, since $1 \geq 1-\beta \geq 1-\cos(\pi/9) = \Omega(1)$, we naturally have $1 = D = \Omega((1-\beta)^{-\frac{1}{2}})$. Note that the complete graph can also be regarded as a special ring-lattice graph $\mathcal{R}_{n,n-1}$.

(Case 2: $\cos(\pi/9) < \beta \leq \cos(\pi/n)$.) Note that this case requires $n \geq 10$. Letting

$$k = 2\left\lceil n\sqrt{\frac{3(1-\beta)}{\pi^2(1-\pi^2/12)}}\right\rceil \geq \max\left\{2n\sqrt{\frac{3(1-\beta)}{\pi^2(1-\pi^2/12)}}, 2\right\}, \tag{29}$$

we consider the ring-lattice graph $\mathcal{R}_{n,k}$ with the weight matrix

$$W = I - \frac{1-\beta}{\min_{2\leq j\leq n}\mu_j} L_{n,k} \tag{30}$$

where k is defined as in (29) and $\{\mu_j : 2 \leq j \leq n\}$ are given in Lemma 2. For such ring-lattice graph and its associated weight matrix defined in (30), we prove $D = \Omega((1-\beta)^{-\frac{1}{2}})$ in following steps:

- We first check whether $\mathcal{R}_{n,k}$ is well-defined, i.e., $2 \leq k < n-1$ and k is even. It is easy to observe that k is even by (29). Furthermore, since $\cos(\pi/9) < \beta$ and $n \geq 10$, we have

$$k < 2\left(n\sqrt{\frac{3(1-\beta)}{\pi^2(1-\pi^2/12)}}+1\right) < 2\left(n\sqrt{\frac{3(1-\cos(\pi/9))}{\pi^2(1-\pi^2/12)}}+1\right) < 0.7n+2 \leq n-1. \tag{31}$$

- We next check whether $W_{n,k} \in \mathcal{W}_{n,\beta}$, i.e., $\|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\|_2 = \beta$. By Lemma 2, we have

$$\frac{1-\beta}{\min_{2\leq j\leq n}\mu_j} \leq \frac{6(1-\beta)n^2}{(1-\pi^2/12)\pi^2 k(k+1)(k+2)} < \frac{1}{2k}\frac{12(1-\beta)n^2}{(1-\pi^2/12)\pi^2k^2} \leq \frac{1}{2k}.$$

  Since all eigenvalues of $L_{n,k}$ lie in $[0, 2k]$ (see (21)), it holds that W is positive semi-definite (see definition (30)). Moreover,

$$\|W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\|_2 = \max_{2\leq\ell\leq n}\left\{1 - \frac{1-\beta}{\min\{\mu_j : 2\leq j\leq n\}}\mu_\ell\right\} = \beta \quad i.e., \quad W \in \mathcal{W}_{n,\beta}.$$

- Finally, we check whether $\Theta(\frac{n}{k}) = D = \Theta((1-\beta)^{-\frac{1}{2}})$. By using the inequality $\cos(\theta) \leq 1 - \theta^2/2 + \theta^4/24$ with $\theta = \pi/n$, we have

$$\beta \leq \cos(\pi/n) \leq 1 - \pi^2/2n^2 + \pi^4/24n^4 \leq 1 - \pi^2/3n^2.$$

  Therefore, we have that $n\sqrt{\frac{3(1-\beta)}{\pi^2(1-\pi^2/12)}} \geq n\sqrt{\frac{\pi^2}{n^2(1-\pi^2/12)}} \geq 2$ and hence

$$k < 2\left(n\sqrt{\frac{3(1-\beta)}{\pi^2(1-\pi^2/12)}}+1\right) \leq 3n\sqrt{\frac{3(1-\beta)}{\pi^2(1-\pi^2/12)}}. \tag{32}$$

  Combining (29) and (32), we achieve

$$\frac{n}{k} = \Theta((1-\beta)^{-\frac{1}{2}}).$$

  This fact together with Lemma 1 leads to $D = \Theta(1/\sqrt{1-\beta})$.

□

D Lower bounds

D.1 Non-convex Case: Proof of Theorem 2

In this subsection, we establish the lower bound of the smooth and non-convex decentralized stochastic optimization. Our analysis builds upon [42] but utilizes the proposed ring-lattice graphs for the construction of worst-case instances, which significantly broadens the scope of weight matrices that the lower bound can apply to, i.e., from $\beta = \cos(\pi/n)$ in [42] to any $\beta \in [0, \cos(\pi/n)]$. Specifically, we will prove for any $\beta \in [0, \cos(\pi/n)]$ and $T = \Omega(1/\sqrt{1-\beta})$,

$$\begin{aligned}&\inf_{A\in\mathcal{A}_W}\sup_{W\in\mathcal{W}_{n,\beta}}\sup_{\{\tilde{g}_i\}_{i=1}^n\subseteq\mathcal{O}_\sigma^2}\sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_L}\mathbb{E}\|\nabla f(\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T})\|^2\\ =&\Omega\left(\frac{\sqrt{\Delta L}\sigma}{\sqrt{nT}} + \frac{\Delta L}{T\sqrt{1-\beta}}\right).\end{aligned} \tag{33}$$

where $\Delta := \mathbb{E}[f(x^{(0)})] - \min_x f(x)$.

We establish the two terms in (33) separately as in [42] by constructing two hard-to-optimize instances. We denote the j-th coordinate of a vector $x \in \mathbb{R}^d$ by $[x]_j$ for $j = 1, \ldots, d$, and let $\mathrm{prog}(x)$ be

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

Similarly, for a set of multiple points $\mathcal{X} = \{x_1, x_2, \dots\}$, we define $\mathrm{prog}(\mathcal{X}) := \max_{x\in\mathcal{X}} \mathrm{prog}(x)$. As described in [9, 4], a function f is called zero-chain if it satisfies

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

which implies that, starting from $x = 0$, a single gradient evaluation can only earn at most one more non-zero coordinate for the model parameters. We next introduce a key zero-chain function to facilitate the analysis.

Lemma 4 (Lemma 2 of [4])**.** *Let function*

$$\ell(x) := -\Psi(1)\Phi([x]_1) + \sum_{j=1}^{d-1}\Big(\Psi(-[x]_j)\Phi(-[x]_{j+1}) - \Psi([x]_j)\Phi([x]_{j+1})\Big)$$

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

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

Then ℓ *satisfy several properties as below:*

1. $\ell(x) - \inf_x \ell(x) \le \Delta_0 d$, $\forall\, x \in \mathbb{R}^d$ *with* $\Delta_0 = 12$.
2. ℓ *is* L_0*-smooth with* $L_0 = 152$.
3. $\|\nabla \ell(x)\|_\infty \le G_0$, $\forall\, x \in \mathbb{R}^d$ *with* $G_0 = 23$.
4. $\|\nabla \ell(x)\|_\infty \ge 1$ *for any* $x \in \mathbb{R}^d$ *with* $[x]_d = 0$.

We next establish the two terms in (33) by constructing two hard-to-optimize instances.

Instance 1. The proof for the first term $\Omega((\frac{\Delta L\sigma^2}{nT})^{\frac{1}{2}})$ essentially follows [42]. We provide the proof for the sake of being self-contained.

(Step 1.) Let all f_i be $L\lambda^2\ell(x/\lambda)/L_0$ where ℓ is defined in Lemma 4 and λ is to be specified, and thus $f = L\lambda^2\ell(x/\lambda)/L_0$. Since $\nabla^2 f_i = L\nabla^2\ell/L_0$ and h is L_0-smooth (Lemma 4), we know f_i is L-smooth for any $\lambda > 0$. By Lemma 4, we have

$$f(0) - \inf_x f(x) = \frac{L\lambda^2}{L_0}(\ell(0) - \inf_x \ell(x)) \le \frac{L\lambda^2\Delta_0 d}{L_0}.$$

Therefore, to ensure $f_i \in \mathcal{F}_L$ for all $1 \le i \le n$ and $f(0) - \inf_x f(x) \le \Delta$, it suffices to let

$$\frac{L\lambda^2\Delta_0 d}{L_0} \le \Delta, \quad \text{i.e.,} \quad d\lambda^2 \le \frac{L_0\Delta}{L\Delta_0}. \tag{34}$$

(Step 2.) We construct the stochastic gradient oracle $\tilde{g}_i$ $\forall\, i = 1, \dots, n$ as the follows:

$$[\tilde{g}_i(x)]_j = [\nabla f_i(x)]_j\left(1 + \mathbb{1}\{j > \mathrm{prog}(x)\}\left(\frac{Z}{p} - 1\right)\right), \forall\, x \in \mathbb{R}^d,\ j = 1, \dots, d$$

with $Z \sim \text{Bernoulli}(p)$, and $p \in (0, 1)$ to be specified. The oracle $\tilde{g}_i(x)$ has probability p to zero $[\nabla f_i(x)]_{\mathrm{prog}(x)+1}$. It is easy to see $\tilde{g}_i$ is unbiased, i.e., $\mathbb{E}[\tilde{g}_i(x)] = \nabla f_i(x)$ for all $x \in \mathbb{R}^d$. Moreover, since f_is are zero-chain, we have $\mathrm{prog}(\tilde{g}_i(x)) \le \mathrm{prog}(\nabla f_i(x)) \le \mathrm{prog}(x) + 1$ and hence

$$\begin{aligned}
\mathbb{E}[\|\tilde{g}_i(x) - \nabla f_i(x)\|^2] &= |[\nabla f_i(x)]_{\mathrm{prog}(x)+1}|^2\mathbb{E}\left[\left(\frac{Z}{p} - 1\right)^2\right] = |[\nabla f_i(x)]_{\mathrm{prog}(x)+1}|^2\frac{1-p}{p} \\
&\le \|\nabla f_i(x)\|_\infty^2\frac{1-p}{p} \le \frac{L^2\lambda^2(1-p)}{L_0^2 p}\|\nabla \ell(x)\|_\infty^2 \\
&\overset{\text{Lemma 4}}{\le} \frac{L^2\lambda^2(1-p)G_0^2}{L_0^2 p}.
\end{aligned}$$

Therefore, to ensure $\tilde{g}_i \in \mathcal{O}_{\sigma^2}$ for all $1 \le i \le n$, it suffices to let

$$p = \min\{\frac{L^2\lambda^2 G_0^2}{L_0^2\sigma^2}, 1\}. \tag{35}$$

We let $W = (1-\beta)\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T + \beta I$, then obviously $W \in \mathcal{W}_{n,\beta}$.

(Step 3.) Next we show the error $\mathbb{E}[\|\nabla f(x)\|^2]$ is lower bounded by $\Omega((\frac{\Delta L\sigma^2}{nT})^{\frac{1}{2}})$, with any algorithm $A \in \mathcal{A}_W$. Let $x_i^{(t)}, \forall\, t = 0, \dots$ and $1 \le i \le n$, be the t-th query point of node i. Let $\mathrm{prog}^{(t)} = \max_{1\le i\le n,\, 0\le s<t} \mathrm{prog}(x_i^{(s)})$. By Lemma 2 of [42], we have

$$\mathbb{P}(\mathrm{prog}^{(T)} \ge d) \le e^{(e-1)npT-d}. \tag{36}$$

On the other hand, when $\mathrm{prog}^{(T)} < d$, by the fourth point in Lemma 4, it holds that

$$\min_{\hat{x}\in\mathrm{span}\{\{x_i^{(t)}\}_{1\le i\le n, 0\le t<T}\}} \|\nabla f(\hat{x})\| \ge \min_{[\hat{x}]_d=0} \|\nabla f(\hat{x})\| = \frac{L\lambda}{L_0}\min_{[\hat{x}]_d=0}\|\nabla \ell(\hat{x})\| \ge \frac{L\lambda}{L_0}. \tag{37}$$

Therefore, by combining (36) and (37), we have

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] \ge (1 - e^{(e-1)npT-d})\frac{L^2\lambda^2}{L_0^2}. \tag{38}$$

Let

$$\lambda = \frac{L_0}{L}\left(\frac{\Delta L\sigma^2}{3nTL_0\Delta_0 G_0^2}\right)^{\frac{1}{4}} \quad \text{and} \quad d = \left\lfloor \left(\frac{3L\Delta nTG_0^2}{\sigma^2 L_0\Delta_0}\right)^{\frac{1}{2}} \right\rfloor. \tag{39}$$

Then (34) naturally holds and $p = \min\{\frac{G_0^2}{\sigma^2}\left(\frac{\Delta L\sigma^2}{3nTL_0\Delta_0G_0^2}\right)^{\frac{1}{2}}, 1\}$ by plugging (39) into (35). Without loss of generality, we assume T is sufficiently large such that $d \ge 2$. Then, using the definition of p, we have that

$$\begin{aligned}(e-1)npT - d &\le (e-1)nT\,\frac{G_0^2}{\sigma^2}\left(\frac{\Delta L\sigma^2}{3nTL_0\Delta_0G_0^2}\right)^{\frac{1}{2}} - d \\ =&\frac{e-1}{3}\left(\frac{3L\Delta nTG_0^2}{\sigma^2L_0\Delta_0}\right)^{\frac{1}{2}} - d < \frac{e-1}{3}(d+1) - d \le 1 - e < 0\end{aligned}$$

which, combined with (38), further implies

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] = \Omega\left(\frac{L^2\lambda^2}{L_0^2}\right) = \Omega\left(\left(\frac{\Delta L\sigma^2}{3nTL_0\Delta_0G_0^2}\right)^{\frac{1}{2}}\right) = \Omega\left(\left(\frac{\Delta L\sigma^2}{nT}\right)^{\frac{1}{2}}\right).$$

Instance 2. The proof for the second term $\Omega(\frac{\Delta L}{T\sqrt{1-\beta}})$ utilizes weight matrices defined on the ring-lattice graphs described in Theorem 1.

(Step 1.) Let functions

$$\ell_1(x) := -\frac{n}{\lceil n/3\rceil}\Psi(1)\Phi([x]_1) + \frac{n}{\lceil n/3\rceil}\sum_{j \text{ even},\, 0<j<d}\Big(\Psi(-[x]_j)\Phi(-[x]_{j+1}) - \Psi([x]_j)\Phi([x]_{j+1})\Big)$$

and

$$\ell_2(x) := \frac{n}{\lceil n/3\rceil}\sum_{j \text{ odd},\, 0<j<d}\Big(\Psi(-[x]_j)\Phi(-[x]_{j+1}) - \Psi([x]_j)\Phi([x]_{j+1})\Big).$$

Compared to the ℓ function in instance 1, the ℓ_1 and ℓ_2 defined here are $3L_0$-smooth. Furthermore, let

$$f_i = \begin{cases} L\lambda^2\ell_1(x/\lambda)/(3L_0) & \text{if } i \in E_1 \triangleq \{j : 1 \le j \le \lceil \frac{n}{3}\rceil\}, \\ L\lambda^2\ell_2(x/\lambda)/(3L_0) & \text{if } i \in E_2 \triangleq \{j : \lfloor \frac{n}{2}\rfloor + 1 \le j \le \lfloor \frac{n}{2}\rfloor + \lceil \frac{n}{3}\rceil\}, \\ 0 & \text{else.}\end{cases}$$

where $\lambda > 0$ is to be specified. To ensure $f_i \in \mathcal{F}_L$ for all $1 \le i \le n$ and $f(0) - \inf_x f(x) \le \Delta$, it suffices to let

$$\frac{L\lambda^2 \Delta_0 d}{3L_0} \le \Delta, \quad \text{i.e.,} \quad d\lambda^2 \le \frac{3L_0\Delta}{L\Delta_0}. \tag{40}$$

With the functions defined above, we have $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = L\lambda^2 \ell(x/\lambda)/(3L_0)$ and

$$\mathrm{prog}(\nabla f_i(x)) \begin{cases} = \mathrm{prog}(x) + 1 & \text{if } \{\mathrm{prog}(x) \text{ is even and } i \in E_1\} \cup \{\mathrm{prog}(x) \text{ is odd and } i \in E_2\} \\ \le \mathrm{prog}(x) & \text{otherwise.} \end{cases}$$

Therefore, to make progress (i.e., to increase $\mathrm{prog}(x)$), for any gossip algorithm $A \in \mathcal{A}_W$, one must take the gossip communication protocol to transmit information between E_1 to E_2 alternatively. Namely, it takes at least $\mathrm{dist}(E_1, E_2)$ rounds of gossip communications for any possible gossip algorithm A to increase $\mathrm{prog}(\hat{x})$ by 1. Therefore, we have

$$\mathrm{prog}^{(T)} = \max_{1\le i\le n,\, 0\le t<T} \mathrm{prog}(x_i^{(t)}) \le \left\lfloor \frac{T}{\mathrm{dist}(E_1, E_2)} \right\rfloor + 1, \quad \forall\, T \ge 0. \tag{41}$$

(Step 2.) We consider a gradient oracle that return lossless full-batch gradients, i.e., $\tilde{g}_i = \nabla f_i(x)$, $\forall\, x \in \mathbb{R}^d$, $1 \le i \le n$. For the construction of weight matrix, we consider the ring-lattice graph with diameter D in Theorem 1 and its associated weight matrix W such that $W \in \mathcal{W}_{n,\beta}$. Then by Lemma 1 and Theorem 1, we have $\mathrm{dist}(E_1, E_2) = \mathrm{dist}(\lceil n/3 \rceil, \lfloor n/2 \rfloor + 1) = \Theta(D) = \Theta(1/\sqrt{1-\beta})$. Suppose $\mathrm{dist}(E_1, E_2) \ge 1/(C\sqrt{1-\beta})$ with some absolute constant C, then by (41), we have

$$\mathrm{prog}^{(T)} = \max_{1\le i\le n,\, 0\le s\le T} \mathrm{prog}(x_i^{(s)}) \le \left\lfloor C\sqrt{1-\beta}T \right\rfloor + 1, \quad \forall\, T \ge 0. \tag{42}$$

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

$$d = \left\lfloor C\sqrt{1-\beta}T \right\rfloor + 2 < 4C\sqrt{1-\beta}T$$

and

$$\lambda = \left(\frac{3L_0\Delta}{4L\Delta_0 C\sqrt{1-\beta}T} \right)^{\frac{1}{2}}. \tag{43}$$

Then (40) naturally holds. Since $\mathrm{prog}^{(T)} < d$ by (42), following (37) and using (43), we have

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] \ge \min_{[\hat{x}]_d = 0} \|\nabla f(\hat{x})\|^2 \ge \frac{L^2\lambda^2}{9L_0^2} = \Omega\left(\frac{\Delta L}{\sqrt{1-\beta}T} \right).$$

D.2 Non-convex Case with PL Condition: Proof of Theorem 3

In this subsection, we provide the proof for smooth and non-convex decentralized stochastic optimization under the PL condition: for any $\beta \in [0, \cos(\pi/n)]$ and $T = \Omega(1/\sqrt{1-\beta})$,

$$\begin{aligned} &\inf_{A\in\mathcal{A}_W} \sup_{W\in\mathcal{W}_{n,\beta}} \sup_{\{\tilde{g}_i\}_{i=1}^n \subseteq \mathcal{O}_\sigma^2} \sup_{\{f_i\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}} \mathbb{E}[f(\hat{x}_{A,\{f_i\}_{i=1}^n, \{\tilde{g}_i\}_{i=1}^n, W, T}) - f^\star] \\ =&\Omega\left(\frac{\sigma^2}{\mu n T} + \frac{\mu\Delta}{L} \exp(-\sqrt{\mu/L}\sqrt{1-\beta}T) \right). \end{aligned} \tag{44}$$

where $\Delta := \mathbb{E}[f(x^{(0)})] - \min_x f(x)$. We still prove the two terms in (44) separately.

Instance 1. Our proof for first term $\Omega(\frac{\sigma^2}{\mu n T})$ is inspired by [54], who study the lower bounds in the stochastic but single-node regime.

We consider all functions $f_i = f$ are homogeneous and $W = \frac{1-\beta}{n}\mathbb{1}_n\mathbb{1}_n^T + \beta I \in \mathcal{W}_{n\beta}$. We then choose two functions $f^1,\ f^{-1} \in \mathcal{F}_{L,\mu}$ which are close enough to each other so that f^1 and f^{-1} are hard to distinguish within T gradient queries on each node. The indistinguishability between f^1 and f^{-1} follows the standard Le Cam's method in hypothesis testing. Next, we carefully show that

the indistinguishability between f^1 and f^{-1} can be properly translated into the lower bound of the algorithmic performance.

Recall that $[x]_j$ denotes the j-th coordinate of vector $x \in \mathbb{R}^d$. Let

$$f^v = \frac{1}{2}\left(\mu([x]_1 - v\lambda)^2 + L\sum_{j=2}^{d}[x]_j^2\right),$$

where $v \in \{\pm 1\}$, d can be any integer greater than 2, and $\lambda > 0$ is a quantity to be determined. Clearly, f^v is L-smooth and μ-strongly convex, and thus satisfies the μ-PL condition. The optimum $x^{v,\star}$ of f^v is $v\lambda e_1$ with function value $\min_{x\in\mathbb{R}^d} f^v(x) = 0$, where e_1 is the first canonical vector.

We construct the gradient oracle with $\mathcal{N}(0, \sigma^2)$ noise. Specifically, on query at point x on node i, the oracle returns $\tilde{g}_i(x) = \nabla f_i(x) + s_i$ with independent noise $s_i \sim \mathcal{N}(0, \sigma^2)$. Obviously $\tilde{g}_i \in \mathcal{O}_{\sigma^2}$ for all $1 \le i \le n$.

Let $\mathbf{S}^{(T)} := \{(\mathbf{x}^{(t)} \triangleq (x_1^{(t)}, \ldots, x_n^{(t)}), \tilde{\mathbf{g}}(\mathbf{x}^{(t)}) \triangleq (\tilde{g}_1(x_1^{(t)}), \ldots, \tilde{g}_n(x_n^{(t)}))\}_{t=0}^{T-1}$ be set of variables corresponding to the sequence of T queries on all nodes. Let $P^{v,T} := P(\mathbf{S}^{(T)} \mid f^v)$ be the joint distribution of $\mathbf{S}^{(T)}$ if the underlying function was f^v, i.e., $f_1 = \cdots = f_n = f^v$.

Lemma 5. *Let $V \sim \mathrm{Unif}(\{\pm 1\})$, then for any optimization procedure $\hat{x}$ based on the observed gradients, it holds that*

$$\mathbb{E}_V[f^V(\hat{x})] \ge \mu\lambda^2/2 \inf_{\hat{V}} \mathbb{P}(\hat{V} \ne V) \tag{45}$$

where the randomness is over the random index V and the observed data $\mathbf{S}^{(T)}$ and the infimum is taken over all testing procedures $\hat{V}$ based on the observed queries.

Proof. Since $P^{v,T}$ is probability of the observed queries conditioned on $\{V = v\}$, by Markov's inequality, we have

$$\mathbb{E}_V[f^V(\hat{x})] \ge \mu\lambda^2/2\, \mathbb{E}_V[P^{V,T}(f^V(\hat{x}) \ge \mu\lambda^2/2)]. \tag{46}$$

Now, we define the test $\hat{V}_{\mathrm{op}}$ based on the optimization procedure $\hat{x}$ as follows:

$$\hat{V}_{\mathrm{op}} = \begin{cases} v & \text{if } f^v(\hat{x}) < \mu\lambda^2/2 \\ \text{randomly pick one from } \{\pm 1\} & \text{otherwise.} \end{cases},$$

By the definition of f^v, at most one of $\{f^1(\hat{x}), f^{-1}(\hat{x})\}$ is strictly below $\mu\lambda^2/2$, so $\hat{V}_{\mathrm{op}}$ is well-defined. By the definition of $\hat{V}_{\mathrm{op}}$, $\{\hat{V}_{\mathrm{op}} \ne v\}$ implies $\{f^v(\hat{x}) \ge \mu\lambda^2/2\}$. Thus we have

$$P^{v,T}(f^v(\hat{x}) \ge \mu\lambda^2/2) \ge P^{V,T}(\hat{V}_{\mathrm{op}} \ne v) \ge \inf_{\hat{V}} \mathbb{P}(\hat{V} \ne v) \quad \text{for any } v \in \{\pm 1\} \tag{47}$$

Plugging (47) into (46), we reach the conclusion. □

Based on Lemma 5, we know that the worst-case optimization performance in the above scenario can be lower bounded by the indistinguishabiligty between to hypothesises f^1 and f^{-1}. Recall the following standard result of Le Cam [8]

$$\inf_{\hat{V}}\left(P^{1,T}(\hat{V} \ne 1) + P^{-1,T}(\hat{V} \ne -1)\right) = 1 - \|P^{1,T} - P^{-1,T}\|_{TV} \tag{48}$$

where $\|P^{1,T} - P^{-1,T}\|_{TV}$ is the total variation distance between the two distributions. We can precisely measure the indistinguishabiligty between f^1 and f^{-1} by $\|P^{1,T} - P^{-1,T}\|_{TV}$. Since the total variation distance is hard to compute, we turn to compute the KL-divergence, which is a relaxation of the total variation distance due to Pinsker's inequality [14]:

$$\|P^{1,T} - P^{-1,T}\|_{TV}^2 \le \frac{1}{2}\mathrm{D}_{\mathrm{KL}}(P^{1,T}||P^{-1,T}). \tag{49}$$

Lemma 6. *The KL-divergence between $P^{1,T}$ and $P^{-1,T}$ is upper bounded by:*

$$\mathrm{D}_{\mathrm{KL}}(P^{1,T}||P^{-1,T}) \le 2nT\mu^2\lambda^2/\sigma^2. \tag{50}$$

Proof. Since the observed data obey Markov's property, i.e., $\tilde{g}_i(x_i^{(t)}) \perp \mathbf{S}^{(t)} \mid x_i^{(t)}$ for any $1 \le i \le n$ and $0 \le t < T$, we have

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(P^{1,T}||P^{-1,T}) = \mathbb{E}_{P^{1,T}}\left[\log\left(P^{1,T}/P^{-1,T}\right)\right] \\
=&\mathbb{E}_{P^{1,T}}\left[\log\left(\prod_{t=0}^{T-1}\mathbb{P}(\tilde{\mathbf{g}}(\mathbf{x}^{(t)}) \mid \mathbf{x}^{(t)}, f^1)/\prod_{t=0}^{T-1}\mathbb{P}(\tilde{\mathbf{g}}(\mathbf{x}^{(t)}) \mid \mathbf{x}^{(t)}, f^{-1})\right)\right] \\
=&\sum_{t=0}^{T-1}\mathbb{E}_{\mathbf{S}^{(t)}|f^1}\left[\mathbb{E}_{\tilde{\mathbf{g}}(\mathbf{x}^{(t)})\sim\mathcal{N}(\nabla f^1(\boldsymbol{x}),\sigma^2)}\left[\log\left(\mathbb{P}(\mathbf{g}(\mathbf{x}^{(t)}) \mid \mathbf{x}^{(t)}, f^1)/\mathbb{P}(\mathbf{g}(\mathbf{x}^{(t)}) \mid \mathbf{x}^{(t)}, f^{-1})\right)\right]\right] \\
=&\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{S}^{(t)}|f^1}\left[\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}(\tilde{g}(x_i^{(t)}) \mid x_i^{(t)}, f^1)||\mathbb{P}(g(x_i^{(t)}) \mid x_i^{(t)}, f^{-1})\right)\right].
\end{aligned}
$$

Then we apply the uniform upper bound over x as follows:

$$
\mathrm{D}_{\mathrm{KL}}(P^{1,T}||P^{-1,T}) \le nT\sup_{x,i}\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}(\tilde{g}_i(x) \mid x, f^1)||\mathbb{P}(\tilde{g}_i(x) \mid x, f^{-1})\right). \tag{51}
$$

Note the KL-divergence between two Gaussians can be computed explicitly from their means and variances as follows: for any $x \in \mathbb{R}^d$

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}(\tilde{g}_i(x) \mid x, f^1)||\mathbb{P}(\tilde{g}_i(x) \mid x, f^{-1})\right) \\
=&\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(\nabla f^1(x),\sigma^2))||\mathcal{N}(\nabla f^{-1}(x),\sigma^2))\right) = \frac{1}{2\sigma^2}\|\nabla f^1(x) - \nabla f^{-1}(x)\|^2 = \frac{2\mu^2\lambda^2}{\sigma^2}.
\end{aligned} \tag{52}
$$

Plugging (52) into (51), we reach the result (50). □

Therefore, by Lemma 5 and Lemma 6, we can obtain the lower bound based (44). Specifically, for $f^1,\ f^{-1} \in \mathcal{F}_{L,\mu}$ constructed above, assume that the algorithm A receives stochastic gradients with $\mathcal{N}(0,\sigma^2)$ noise, then we have

$$
\begin{aligned}
&\inf_{A\in\mathcal{A}_W}\sup_{W\in\mathcal{W}_{n,\beta}}\sup_{\{\tilde{g}_i\}_{i=1}^n\subseteq\mathcal{O}_\sigma^2}\sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_{L,\mu}}\mathbb{E}[f(\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T}) - f^\star] \\
\ge&\inf_{A\in\mathcal{A}_W}\max_{v\in\{\pm1\}}\mathbb{E}[f^v(\hat{x}_{A,\{f^v\}_{i=1}^n,\mathcal{N}(\nabla f^v(x),\sigma^2),\frac{1-\beta}{n}\mathbb{1}_n\mathbb{1}_n^T+\beta I,T})] \\
\ge&\inf_{A\in\mathcal{A}_W}\mathbb{E}_{V\sim\mathrm{Unif}(\{\pm1\})}[\mathbb{E}[f^V(\hat{x}_{A,\{f^V\}_{i=1}^n,\mathcal{N}(\nabla f^V(x),\sigma^2),\frac{1-\beta}{n}\mathbb{1}_n\mathbb{1}_n^T+\beta I,T})]].
\end{aligned} \tag{53}
$$

Plugging Lemma 5 and Lemma 6 into (53) and using Pinsker's inequality, we immediately have

$$
\begin{aligned}
&\inf_{A\in\mathcal{A}_W}\sup_{W\in\mathcal{W}_{n,\beta}}\sup_{\{\tilde{g}_i\}_{i=1}^n\subseteq\mathcal{O}_\sigma^2}\sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_{L,\mu}}\mathbb{E}[f(\hat{x}_{A,\{f_i\}_{i=1}^n,\{\tilde{g}_i\}_{i=1}^n,W,T}) - f^\star] \\
\ge&\frac{\mu\lambda^2}{2}\left(1-\sqrt{\mathrm{D}_{\mathrm{KL}}(P^{1,T}||P^{-1,T})/2}\right) \\
=&\frac{\mu\lambda^2}{2}\left(1-\frac{\mu\lambda}{\sigma}\sqrt{nT}\right).
\end{aligned} \tag{54}
$$

Choosing $\lambda = \frac{2\sigma}{3\mu\sqrt{nT}}$ in (54), we reach the lower bound $\Omega(\frac{\sigma^2}{\mu nT})$.

Instance 2. Our proof for the term $\Omega(\mu\Delta/L\exp(-\sqrt{\mu/L}\sqrt{1-\beta}T))$ builds on the similar idea to [58]: splitting the function used by Nesterov to prove the lower bound for strongly convex and smooth optimization [50]. However, the number of nodes n is fixed in our analysis while [58] needs n to be varying when establishing the lower bound. Besides, our construction allows the objective functions to be suitable to an arbitrary initialization scope $\Delta = f(x^{(0)}) - \min_x f(x)$ which, however, is not fully addressed and discussed in [58]. These differences make our analysis novel and stronger.

We construct deterministic scenarios where no gradient noise is employed in the oracle. We assume the variable $x \in \ell_2 \triangleq \{([x]_1,[x]_2,\ldots,) : \sum_{r=1}^\infty [x]_r^2 < \infty\}$ to be infinitely dimensional and square-summable for simplicity. It is easy to adapt the argument for finitely dimensional variables as long

as the dimension is proportionally larger than T. Without loss of generality, we assume all the algorithms start from $x^{(0)} = 0 \in \ell_2$. Let M be

$$M = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \in \mathbb{R}^{\infty \times \infty},$$

then it is easy to see $0 \preceq M \preceq 4I$. Let $E_1 \triangleq \{j : 1 \le j \le \lceil \frac{n}{3} \rceil\}$ and $E_2 \triangleq \{j : \lfloor \frac{n}{2} \rfloor + 1 \le j \le \lfloor \frac{n}{2} \rfloor + \lceil \frac{n}{3} \rceil\}$ as in Appendix D.1, and let

$$f_i(x) = \begin{cases} \frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{12}\frac{n}{\lceil n/3 \rceil}\left([x]_1^2 + \sum_{r\ge 1}([x]_{2r} - [x]_{2r+1})^2 - 2\lambda[x]_1\right) & \text{if } i \in E_1, \\ \frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{12}\frac{n}{\lceil n/3 \rceil}\sum_{r\ge 1}([x]_{2r-1} - [x]_{2r})^2 & \text{if } i \in E_2, \\ \frac{\mu}{2}\|x\|^2 & \text{otherwise.} \end{cases} \tag{55}$$

where $\lambda \in \mathbb{R}$ is to be specified. It is easy to see that $[x]_1^2 + \sum_{r\ge 1}([x]_{2r} - [x]_{2r+1})^2 - 2\lambda[x]_1$ and $\sum_{r\ge 1}([x]_{2r-1} - [x]_{2r})^2$ are convex and 4-smooth. We thus have all f_i are L-smooth and μ-strongly convex, which implies $f_i \in \mathcal{F}_{L,\mu}$ for all $1 \le i \le n$. We further have $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{\mu}{2}\|x\|^2 + \frac{L-\mu}{12}\left(x^T M x - 2\lambda[x]_1\right)$.

For the functions defined above, we establish that

Lemma 7. *Denote $\kappa := L/\mu > 1$, then it holds that for any x and $r \ge 1$ satisfying $\mathrm{prog}(x) \le r$,*

$$f(x) - \min_x f(x) \ge \frac{\mu}{2L}\left(1 - \frac{6}{\sqrt{3+6\kappa}+3}\right)^{2r}(f(x^{(0)}) - \min_x f(x)).$$

Proof. The minimum $x^\star$ of function f satisfies $\left(\frac{L-\mu}{6}M + \mu\right)x - \lambda\frac{L-\mu}{6}e_1 = 0$, which is equivalent to

$$\begin{aligned} \left(2 + \frac{6}{\kappa - 1}\right)[x]_1 - [x]_2 &= \lambda, \\ -[x]_{j-1} + \left(2 + \frac{6}{\kappa-1}\right)[x]_j - [x]_{j+1} &= 0, \quad \forall\, j \ge 2. \end{aligned} \tag{56}$$

Let q be the smallest root of the equation $q^2 - \left(2 + \frac{6}{\kappa-1}\right)q + 1 = 0$, then the variable

$$x^\star = \left([x^\star]_j = \lambda q^j\right)_{j\ge 1}$$

satisfies (56). By the strong convexity of f, $x^\star$ is the unique solution. Therefore, when $\mathrm{prog}(x) \le r$, it holds that

$$\|x - x^\star\|^2 \ge \sum_{j=r+1}^{\infty} \lambda^2 q^{2j} = \lambda^2 \frac{q^{2(r+1)}}{1-q^2} = q^{2r}\|x^{(0)} - x^\star\|^2.$$

Finally, noting that

$$q = 1 + \frac{3 - \sqrt{3+6\kappa}}{\kappa - 1} = 1 - \frac{6}{\sqrt{3+6\kappa}+3},$$

and using the strong convexity of f, we reach the conclusion. □

Moreover, it is easy to see that the optimal value of $f(x)$ is

$$\min_x f(x) = f(x^\star) = -\frac{L-\mu}{12}\lambda[x^\star]_1 = -\frac{L-\mu}{12}\lambda^2 q.$$

Therefore, for any given $\Delta > 0$, we can chose $\lambda = \sqrt{\frac{12\Delta}{(L-\mu)q}}$ such that $f(x^{(0)}) - \min_x f(x) = \Delta$.

Similar to the argument in Appendix D.1, we have

$$\mathrm{prog}(\nabla f_i(x)) \begin{cases} = \mathrm{prog}(x) + 1 & \text{if } \{\mathrm{prog}(x) \text{ is even and } i \in E_1\} \cup \{\mathrm{prog}(x) \text{ is odd and } i \in E_2\} \\ \leq \mathrm{prog}(x) & \text{otherwise.} \end{cases}$$

We further have

$$\mathrm{prog}^{(T)} = \max_{1 \leq i \leq n,\, 0 \leq s \leq T} \mathrm{prog}(x_i^{(t)}) \leq \left\lfloor \frac{T}{\mathrm{dist}(E_1, E_2)} \right\rfloor + 1, \quad \forall T \geq 0. \tag{57}$$

Therefore, with T rounds of gossip communications budget in total, starting from $x_i^{(0)} = 0$ for any $i = 1, \ldots, n$, any gossip algorithm A can only achieves at most $\mathrm{prog}(\hat{x}) = 1 + \lfloor \frac{T}{\mathrm{dist}(E_1, E_2)} \rfloor$, which, combined with Lemma 7, leads to lower bound that

$$\begin{aligned} f(\hat{x}) - \min_x f(x) \geq & \frac{\mu\Delta}{2L} \left(1 - \frac{6}{\sqrt{3 + 6\kappa} + 3}\right)^{2 + 2\lfloor \frac{T}{\mathrm{dist}(E_1, E_2)} \rfloor} \\ = & \Omega\left(\frac{\mu}{L} \exp\left(-\frac{\sqrt{6}\lfloor \frac{T}{\mathrm{dist}(E_1, E_2)} \rfloor}{\sqrt{\kappa}}\right) \Delta\right). \end{aligned} \tag{58}$$

where Δ measures the initialization $\mathbb{E}[f(x^{(0)})] - \min_x f(x)$. The rest follows using the ring-lattice associated weight matrix $W \in \mathcal{W}_{n,\beta}$ such that

$$\mathrm{dist}(E_1, E_2) = \Omega((1 - \beta)^{-\frac{1}{2}}).$$

D.3 Connectivity measures in common weight matrices

Table D.3 lists the order of the connectivity measure β for weight matrices, if generated through the Laplacian rule $W = I - L/d_{\max}$ in which L is the Laplacian matrix and $d_{\max}$ is the maximum degree, associated with commonly-used topologies. Since $\cos(\frac{\pi}{n}) \geq 1 - \frac{\pi^2}{2n^2}$, we have $[0, 1 - \frac{\pi^2}{2n^2}] \subseteq [0, \cos(\pi/n)]$. Since β listed in Table D.3 lies in $[0, 1 - \frac{\pi^2}{2n^2}]$ for sufficiently large n, we conclude that our lower bound established in Theorems 2 and 3 applies to weight matrices associated with topologies listed in Table D.3.

Table 8: Order of connectivity measure β. "E.-R. Rand": Erdos-Renyi random graph $G(n, p)$ with probability $p = (1 + a)\ln(n)/n$ for some $a > 0$; "Geo. Rand": geometric random graph $G(n, r)$ with radius $r^2 = (1 + a)\ln(n)/n$ for some $a > 0$.

Topology	Order of β
Grid [47]	1-$\Theta([n\ln(n)]^{-1})$
Torus [47]	1-$\Theta(n^{-1})$
Hypercube [47]	1-$\Theta([\ln(n)]^{-1})$
Exponential [73]	1-$\Theta([\ln(n)]^{-1})$
Complete	0
E.-R. Rand. [6]	1-$\Theta([\ln(n)]^{-1})$
Geo. Rand. [7]	1-$\Theta(\ln(n)n^{-1})$

E Convergence in MG-DSGD

E.1 Smooth and non-convex setting

This section examines the convergence rate of MG-DSGD under the smooth and non-convex setting. Since MG-DSGD is a direct variant based on the vanilla DSGD, we first establish the convergence of DSGD to facilitate the convergence of MG-DSGD.

E.1.1 Convergence rate of DSGD

The convergence rate of DSGD has been established in literatures such as [28, 13]. We adjust the analysis therein to achieve a slightly different result, which lays the foundation for the later convergence analysis in MG-DSGD.

The following lemma established in [28, Lemma 8] (also in [13, Lemma 6]) shows how $\mathbb{E}[f(\bar{x}^{(k)})]$ evolves with iterations.

Lemma 8 (DESCENT LEMMA [28]). *If $\{f_i(x)\}_{i=1}^n \subseteq \mathcal{F}_L$, $\{\tilde{g}_i\} \subseteq O_{\sigma^2}$, and learning rate $\gamma < \frac{1}{4L}$, it holds for $k = 0, 1, 2, \cdots$ that*

$$\mathbb{E}[f(\bar{x}^{(k+1)})] \leq \mathbb{E}[f(\bar{x}^{(k)})] - \frac{\gamma}{4}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 + \frac{3\gamma L^2}{4n}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + \frac{\gamma^2\sigma^2 L}{2n}, \tag{59}$$

where $\bar{\mathbf{x}}^{(k)} = [(\bar{x}^{(k)})^T; \cdots; (\bar{x}^{(k)})^T] \in \mathbb{R}^{n\times d}$, and $\bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^n x_i^{(k)}$.

The following lemma established in [13, Lemma 8] shows how the ergodic consensus term evolves with iterations.

Lemma 9 (CONSENSUS LEMMA [13]). *If $\{f_i(x)\}_{i=1}^n \subseteq \mathcal{F}_L$, $\{\tilde{g}_i\} \subseteq O_{\sigma^2}$, $W \in \mathcal{W}_{n,\beta}$, $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \leq b^2$ for any $x \in \mathbb{R}^d$, and learning rate $\gamma \leq \frac{1-\beta}{4\beta L}$, it holds that*

$$\frac{1}{K+1}\sum_{k=0}^K \mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 \leq \frac{12n\beta^2\gamma^2}{(1-\beta)^2(K+1)}\sum_{k=0}^K \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 + \frac{2n\gamma^2\beta^2}{1-\beta}\left(\frac{3b^2}{1-\beta} + \sigma^2\right). \tag{60}$$

Proof. The Lemma 8 in [13] covers both DSGD and DSGD with periodic global average. To recover the result for DSGD, we let $C_\beta = 1/(1-\beta)$, $D_\beta = 1/(1-\beta)$ in [13, Lemma 8] to achieve (60). □

The following lemma establishes the convergence rate of DSGD under the smooth and non-convex setting.

Lemma 10 (DSGD CONVERGENCE). *If $\{f_i(x)\}_{i=1}^n \subseteq \mathcal{F}_L$, $\{\tilde{g}_i\} \subseteq O_{\sigma^2}$, $W \in \mathcal{W}_{n,\beta}$, $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \leq b^2$ for any $x \in \mathbb{R}^d$, and learning rate is set as*

$$\gamma = \min\left\{\frac{\sqrt{2n\Delta}}{\sigma\sqrt{L(K+1)}}, \frac{1-\beta}{6\beta L}, \frac{1}{4L}\right\}, \tag{61}$$

where $\Delta = \mathbb{E}[f(x^{(0)})] - f^\star$, it holds that

$$\frac{1}{K+1}\sum_{k=0}^K \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = \mathcal{O}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nK}} + \frac{\beta^2 Ln\Delta}{(1-\beta)K} + \frac{\beta^2 Lnb^2\Delta}{(1-\beta)^2 K\sigma^2} + \frac{\beta L\Delta}{(1-\beta)K} + \frac{L\Delta}{K}\right) \tag{62}$$

Remark 9. *This result is slightly different from Theorem 2 in [28] where β does not appear in the numerator of any terms in the upper bound therein. However, as we will show in the analysis of MG-DSGD, the β appearing in the numerator of the third term in* (62) *can significantly reduce the influence of data heterogeneity b^2 if $\beta \to 0$.*

Proof. When $\gamma \leq 1/(4L)$, we know from Lemma 8 that

$$\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 \leq \frac{4}{\gamma}\mathbb{E}[f(\bar{x}^{(k)})] - \frac{4}{\gamma}\mathbb{E}[f(\bar{x}^{(k+1)})] + \frac{3L^2}{n}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + \frac{2\gamma\sigma^2 L}{n}, \tag{63}$$

Taking running average over the above inequality, we obtain

$$\begin{aligned}\frac{1}{K+1}\sum_{k=0}^K \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 \leq \frac{4}{\gamma(K+1)}\Big(\mathbb{E}[f(\bar{x}^{(0)})] - \mathbb{E}[f(\bar{x}^{(K+1)})]\Big) + \frac{2\gamma\sigma^2 L}{n}, \\ + \frac{3L^2}{n(K+1)}\sum_{k=0}^K \mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2\end{aligned} \tag{64}$$

With Lemma (60) and the fact that $\mathbb{E}[f(\bar{x}^{(0)})] - \mathbb{E}[f(\bar{x}^{(K+1)})] \leq \mathbb{E}[f(\bar{x}^{(0)})] - f^\star$, (64) becomes

$$\begin{aligned}&\Big(1 - \frac{36\beta^2 L^2\gamma^2}{(1-\beta)^2}\Big)\frac{1}{K+1}\sum_{k=0}^K \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 \\ &\leq \frac{4\Big(\mathbb{E}[f(\bar{x}^{(0)})] - f^\star\Big)}{\gamma(K+1)} + \frac{2\gamma\sigma^2 L}{n} + \frac{6\gamma^2\beta^2 L^2\sigma^2}{1-\beta} + \frac{18\gamma^2\beta^2 L^2 b^2}{(1-\beta)^2}\end{aligned} \tag{65}$$

If learning rate $\gamma \leq \frac{1-\beta}{12\beta L}$, it holds that $1 - \frac{36\beta^2 L^2 \gamma^2}{1-\beta} \geq 1/2$ and hence

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 \leq \frac{8\Delta}{\gamma(K+1)} + \frac{4\gamma\sigma^2 L}{n} + \frac{12\gamma^2\beta^2 L^2\sigma^2}{1-\beta} + \frac{36\gamma^2\beta^2 L^2 b^2}{(1-\beta)^2} \tag{66}$$

where we introduce constant $\Delta = \mathbb{E}[f(\bar{x}^{(0)})] - f^\star$. Next we set

$$\gamma_1 = \frac{\sqrt{2\Delta n}}{\sigma\sqrt{L(K+1)}}, \quad \gamma_2 = \frac{1-\beta}{6\beta L}, \quad \gamma_3 = \frac{1}{4L}, \quad \gamma = \min\{\gamma_1, \gamma_2, \gamma_3\}. \tag{67}$$

Apparently, it holds that

$$\begin{aligned}\frac{8\Delta}{\gamma(K+1)} &\leq \frac{8\Delta}{\gamma_1(K+1)} + \frac{8\Delta}{\gamma_2(K+1)} + \frac{8\Delta}{\gamma_3(K+1)} \\ &= \mathcal{O}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nK}} + \frac{\beta\Delta L}{(1-\beta)K} + \frac{\Delta L}{K}\right)\end{aligned} \tag{68}$$

and

$$\begin{aligned}\frac{4\gamma\sigma^2 L}{n} + \frac{12\gamma^2\beta^2 L^2\sigma^2}{1-\beta} + \frac{36\gamma^2\beta^2 L^2 b^2}{(1-\beta)^2} &\leq \frac{4\gamma_1\sigma^2 L}{n} + \frac{12\gamma_1^2\beta^2 L^2\sigma^2}{1-\beta} + \frac{36\gamma_1^2\beta^2 L^2 b^2}{(1-\beta)^2} \\ &= \mathcal{O}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nK}} + \frac{\beta^2 L\Delta n}{(1-\beta)K} + \frac{\beta^2 L\Delta n b^2}{(1-\beta)^2 K\sigma^2}\right)\end{aligned} \tag{69}$$

Combining (66), (68) and (69), we have (62).

□

E.1.2 Convergence rate of MG-DSGD

Recall the recursions of DSGD and MG-DSGD as follows.

$$\text{(DSGD)} \quad \mathbf{x}^{(k+1)} = W\big(\mathbf{x}^{(k)} - \gamma\nabla F(\mathbf{x}^{(k)};\xi^{(k)})\big) \tag{70}$$

$$\text{(MG-DSGD)} \quad \mathbf{x}^{(k+1)} = \bar{M}\big(\mathbf{x}^{(k)} - \gamma\mathbf{g}^{(k)})\big), \quad \text{where} \quad \mathbf{g}^{(k)} = \frac{1}{R}\sum_{r=1}^{R}\nabla F(\mathbf{x}^{(k)};\xi^{(k,r)}) \tag{71}$$

where $\bar{M}$ is doubly stochastic (see Prop. 1). MG-DSGD has two differences from the vanilla DSGD. First, the weight matrix W is replaced with $\bar{M}$. Second, the stochastic gradient $\mathbf{g}^{(k)}$ is achieved via gradient accumulation. This implies that the convergence analysis of MG-DSGD can follow that of vanilla DSGD. We only need to pay attentions to the influence of $\bar{M}$ obtained by fast gossip averaging and the $\mathbf{g}^{(k)}$ achieved by gradient accumulation. To proceed, we notice that

$$\mathbb{E}[\|g_i^{(k)} - \nabla f_i(x)\|^2] = \frac{1}{R}\mathbb{E}[\|\nabla F(x_i^{(k)};\xi_i^{(k,r)}) - \nabla f_i(x)\|^2] \leq \frac{\sigma^2}{R}$$

because $\{\xi_i^{(k,r)}\}_{r=1}^R$ are sampled independently. We introduce $\tilde{\sigma}^2 \triangleq \sigma^2/R$ for notation simplicity. In addition, we know from Prop. 1 that

$$\|\bar{M} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T\|_2 \leq \tilde{\beta}, \quad \text{where} \quad \tilde{\beta} := \sqrt{2}\Big(1 - \sqrt{1-\beta}\Big)^R.$$

If we let

$$R = \left\lceil \frac{\max\{\ln(2), \frac{1}{2}\ln(n\max\{1, \frac{b^2}{\sigma^2\sqrt{1-\beta}}\})\}}{\sqrt{1-\beta}} \right\rceil = \tilde{O}\left(\frac{1}{\sqrt{1-\beta}}\right), \tag{72}$$

then it holds that

$$\tilde{\beta} \triangleq \sqrt{2}(1-\sqrt{1-\beta})^R \leq \sqrt{2}e^{-\sqrt{1-\beta}R} \leq \min\left\{\frac{1}{\sqrt{2}}, \sqrt{2}\left(n\max\left\{1, \frac{b^2}{\sigma^2\sqrt{1-\beta}}\right\}\right)^{-\frac{1}{2}}\right\}. \tag{73}$$

We thus immediately have $1-\tilde{\beta}=\Omega(1)$ and

$$\tilde{\beta}^2 n \max\{1, b^2R/\sigma^2\} = \tilde{O}\left(\tilde{\beta}^2 n \max\left\{1, \frac{b^2}{\sigma^2\sqrt{1-\beta}}\right\}\right) = \tilde{O}(1). \tag{74}$$

Theorem 4 (Formal version). *Given $L>0$, $n\geq 2$, $\beta\in[0,1)$, $\sigma>0$, and let A denote Algorithm 1. Assuming that $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x)-\nabla f(x)\|^2 \leq b^2$ for any $x\in\mathbb{R}^n$ and the number of gossip rounds R is set as in (72), the convergence of A can be bounded for any $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_L$ and any $W\in\mathcal{W}_{n,\beta}$ that*

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = \tilde{O}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nT}} + \frac{L\Delta}{\sqrt{1-\beta}T}\right) \quad \textit{where} \quad \bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^n x_i^{(k)} \tag{75}$$

and $\Delta = \mathbb{E}[f(\bar{x}^{(0)})] - f^\star$.

Remark 10. *Comparing with the established lower bound* (33)*, we find the upper bound* (75) *matches it tightly in all constants.*

Proof. The convergence analysis of DSGD applies to MG-DSGD by replacing σ^2 with $\tilde{\sigma}^2$, and β with $\tilde{\beta}$. As a result, we know from Lemma 10 that if the learning rate is set as

$$\gamma = \min\{\frac{\sqrt{\Delta n}}{\tilde{\sigma}\sqrt{L(K+1)}}, \frac{1-\tilde{\beta}}{6\tilde{\beta}L}, \frac{1}{4L}\} = \tilde{O}\left(\min\{\frac{\sqrt{2\Delta n}}{\sigma\sqrt{(1-\beta)LT}}, \frac{1}{L}\}\right) \tag{76}$$

where the last equality holds since $T=KR$ (where T is the number of total gradient queries or communication rounds), it holds that

$$\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^K \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 &= \mathcal{O}\left(\frac{\tilde{\sigma}\sqrt{\Delta L}}{\sqrt{nK}} + \frac{\tilde{\beta}^2 L\Delta n}{(1-\tilde{\beta})K} + \frac{\tilde{\beta}^2 L\Delta n b^2}{(1-\tilde{\beta})^2 K\tilde{\sigma}^2} + \frac{\tilde{\beta}L\Delta}{(1-\tilde{\beta})K} + \frac{L\Delta}{K}\right) \\
&\overset{\tilde{\beta}\leq\frac{1}{\sqrt{2}};\, T=KR}{=} \mathcal{O}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nT}} + \frac{\tilde{\beta}^2 RL\Delta n}{T} + \frac{\tilde{\beta}^2 R^2 L\Delta n b^2}{T\sigma^2} + \frac{RL\Delta}{T}\right) \\
&\overset{(72)}{=} \tilde{\mathcal{O}}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nT}} + \frac{\tilde{\beta}^2 RL\Delta n}{T} + \frac{\tilde{\beta}^2 RL\Delta n \frac{b^2}{\sigma^2\sqrt{1-\beta}}}{T} + \frac{RL\Delta}{T}\right) \\
&\overset{(74)}{=} \tilde{\mathcal{O}}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nT}} + \frac{RL\Delta}{T}\right) \overset{(72)}{=} \tilde{\mathcal{O}}\left(\frac{\sigma\sqrt{\Delta L}}{\sqrt{nT}} + \frac{L\Delta}{\sqrt{1-\beta}T}\right).
\end{aligned}$$

□

E.2 Smooth and non-convex setting under PL condition

Similar to Appendix E.1, we first establish the convergence of DSGD under the PL condition, and then derive the convergence of MG-DSGD.

E.2.1 Convergence rate of DSGD

Substituting (3) to inequality (59), we achieve the following descent lemma.

Lemma 11 (DESCENT LEMMA)**.** *If $\{f_i(x)\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}$, $\{\tilde{g}_i\}\subseteq O_{\sigma^2}$ and learning rate $\gamma\leq\frac{1}{4L}$, it holds for $k=0,1,2,\cdots$ that*

$$\mathbb{E}[f(\bar{x}^{(k+1)})] - f^\star \leq (1-\frac{\mu\gamma}{2})(\mathbb{E}[f(\bar{x}^{(k)})] - f^\star) + \frac{3\gamma L^2}{4n}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + \frac{\gamma^2\sigma^2 L}{2n}. \tag{77}$$

The following lemma establishes the consensus lemma for smooth and non-convex loss functions satisfying the PL condition.

Lemma 12 (CONSENSUS LEMMA). *If $\{f_i(x)\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}$, $\{\tilde{g}_i\} \subseteq O_{\sigma^2}$, $W \in \mathcal{W}_{n,\beta}$, $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \leq b^2$ for any $x \in \mathbb{R}^d$, and learning rate $\gamma \leq \frac{1-\beta}{3\beta L}$, it holds for $k = 0, 1, 2, \cdots$ that*

$$\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|^2 \leq \big(\frac{1+\beta}{2}\big)\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{6n\beta^2\gamma^2 L}{1-\beta}\big(\mathbb{E}f(\bar{x}^{(k)}) - f^\star\big) + n\gamma^2\beta^2\sigma^2 + \frac{3n\beta^2\gamma^2 b^2}{1-\beta}. \tag{78}$$

Proof. When $\{f_i\}_{i=1}^L \subseteq \mathcal{F}_L$, it is know from [13, Eq. (78)] that

$$\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|^2 \leq \big(\beta + \frac{3\beta^2\gamma^2 L^2}{1-\beta}\big)\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 + \frac{3n\beta^2\gamma^2\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2}{1-\beta} + n\gamma^2\beta^2\sigma^2 + \frac{3n\beta^2\gamma^2 b^2}{1-\beta}. \tag{79}$$

If $\gamma \leq (1-\beta)/(3\beta L)$, it holds that

$$\beta + \frac{3\beta^2\gamma^2 L^2}{1-\beta} \leq \frac{1+\beta}{2}. \tag{80}$$

Moreover, since $\{f_i\}_{i=1}^L \subseteq \mathcal{F}_L$ and hence $f(x) \in \mathcal{F}_L$, we have

$$\|\nabla f(\bar{x}^{(t)})\|^2 \leq 2L\big(f(\bar{x}^{(t)}) - f^\star\big). \tag{81}$$

Substituting (80) and (81) to (79), we achieve (78). □

The following lemma establishes the convergence rate of DSGD under the PL condition. Using different proof techniques from those for the strongly-convex scenario in [28], we can show how fast the last-iterate variable, i.e., $f(\bar{x}^{(K)}) - f^\star$, converges. In contrast, authors in [28] show the ergodic convergence, i.e., $\frac{1}{H_K}\sum_{i=1}^K h_k f(\bar{x}^{(k)}) - f^\star$ where $h_k > 0$ is some positive weight and $H_K = \sum_{k=1}^K h_k$. Moreover, the β appearing in the numerator of the third term in (83) can significantly reduce the influence of data heterogeneity b^2 when $\beta \to 0$.

Lemma 13 (DSGD CONVERGENCE). *If $\{f_i(x)\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}$, $\{\tilde{g}_i\} \subseteq O_{\sigma^2}$, $W \in \mathcal{W}_{n,\beta}$, $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \leq b^2$ for any $x \in \mathbb{R}^d$, and learning rate γ is set as*

$$\gamma = \min\left\{\frac{4}{\mu K}\ln(\frac{\Delta\mu^2 nK}{\sigma^2 L}), \frac{1-\beta}{4L}, \frac{\mu(1-\beta)}{24n\beta L^2}\right\}, \tag{82}$$

it holds that

$$\begin{aligned}&\mathbb{E}[f(\bar{x}^{(K)})] - f^\star + L\mathbb{E}\|\mathbf{x}^{(K)} - \bar{\mathbf{x}}^{(K)}\|^2 \qquad (83)\\ =&\tilde{\mathcal{O}}\Big(\frac{\sigma^2 L}{\mu^2 nK} + \frac{c_1\beta^2\sigma^2}{\mu^2 K^2(1-\beta)} + \frac{c_2\beta^2 b^2}{\mu^2 K^2(1-\beta)^2} + \Delta\exp(-\frac{\mu(1-\beta)K}{L}) + \Delta\exp(-\frac{\mu^2(1-\beta)K}{n\beta L^2})\Big)\end{aligned}$$

where $\Delta = \mathbb{E}[f(x^{(0)})] - f^\star$, c_1 and c_2 are constants defined as follows

$$c_1 = \frac{6L}{\mu} + \frac{24L^3}{\mu^2} + 4nL, \quad c_2 = \frac{18L^2}{\mu} + 12nL. \tag{84}$$

Proof. With Lemmas 11 and 12, if γ satisfies

$$\gamma \leq \frac{1}{4L}, \quad \text{and} \quad \gamma \leq \frac{1-\beta}{3\beta L}, \tag{85}$$

then we have

$$\underbrace{\begin{bmatrix} \mathbb{E}[f(\bar{x}^{(k+1)})] - f^\star \\ L\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|^2 \end{bmatrix}}_{:=z^{(k+1)}} \leq \underbrace{\begin{bmatrix} 1-\frac{\mu\gamma}{2} & \frac{3\gamma L}{4n} \\ \frac{6n\beta^2\gamma^2L^2}{1-\beta} & \frac{1+\beta}{2} \end{bmatrix}}_{:=A} \underbrace{\begin{bmatrix} \mathbb{E}[f(\bar{x}^{(k)})] - f^\star \\ L\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 \end{bmatrix}}_{:=z^{(k)}} + \underbrace{\begin{bmatrix} \frac{\gamma^2\sigma^2L}{2n} \\ nL\gamma^2\beta^2\sigma^2 + \frac{3nL\beta^2\gamma^2b^2}{1-\beta} \end{bmatrix}}_{:=s} \tag{86}$$

By recursing the above inequality, we have

$$z^{(k)} \leq A^k z^{(0)} + \sum_{\ell=0}^{k-1} A^\ell s \leq A^k z^{(0)} + (I-A)^{-1}s \tag{87}$$

where the last inequality holds because A is nonnegative and

$$\rho(A) \leq \|A\|_1 \leq \max\{1 - \frac{\mu\gamma}{2} + \frac{6n\beta^2\gamma^2L^2}{1-\beta}, \frac{3\gamma L}{4n} + \frac{1+\beta}{2}\} \leq 1 - \frac{\mu\gamma}{4}. \tag{88}$$

The last inequality in (88) holds when γ satisfies

$$\gamma \leq \frac{(1-\beta)n}{3L} \quad \Longrightarrow \quad \frac{3\gamma L}{4n} + \frac{1+\beta}{2} \leq \frac{3+\beta}{4} \tag{89}$$

$$\gamma \leq \frac{1-\beta}{\mu} \quad \Longrightarrow \quad \frac{3+\beta}{4} \leq 1 - \frac{\mu\gamma}{4} \tag{90}$$

$$\gamma \leq \frac{\mu(1-\beta)}{24n\beta^2L^2} \quad \Longrightarrow \quad 1 - \frac{\mu\gamma}{2} + \frac{6n\beta^2\gamma^2L^2}{1-\beta} \leq 1 - \frac{\mu\gamma}{4} \tag{91}$$

Furthermore, we can derive from (87) that

$$\begin{aligned} \|z^{(k)}\|_1 &\leq \|A^k\|_1\|z^{(0)}\|_1 + \|(I-A)^{-1}s\|_1 \\ &\leq \|A\|_1^k\|z^{(0)}\|_1 + \|(I-A)^{-1}s\|_1 \\ &\overset{(88)}{\leq} (1-\frac{\mu\gamma}{4})^k\|z^{(0)}\|_1 + \|(I-A)^{-1}s\|_1. \end{aligned} \tag{92}$$

It is easy to verify that

$$\begin{aligned} (I-A)^{-1}s &= \frac{1}{\frac{\mu\gamma(1-\beta)}{4} - \frac{9\beta^2\gamma^3L^3}{2(1-\beta)}} \begin{bmatrix} \frac{1-\beta}{2} & \frac{3\gamma L}{4n} \\ \frac{6n\beta^2\gamma^2L^2}{1-\beta} & \frac{\mu\gamma}{2} \end{bmatrix} \begin{bmatrix} \frac{\gamma^2\sigma^2L}{2n} \\ nL\gamma^2\beta^2\sigma^2 + \frac{3nL\beta^2\gamma^2b^2}{1-\beta} \end{bmatrix} \\ &\overset{(a)}{\leq} \frac{8}{\mu\gamma(1-\beta)} \begin{bmatrix} \frac{1-\beta}{2} & \frac{3\gamma L}{4n} \\ \frac{6n\beta^2\gamma^2L^2}{1-\beta} & \frac{\mu\gamma}{2} \end{bmatrix} \begin{bmatrix} \frac{\gamma^2\sigma^2L}{2n} \\ nL\gamma^2\beta^2\sigma^2 + \frac{3nL\beta^2\gamma^2b^2}{1-\beta} \end{bmatrix} \\ &= \begin{bmatrix} \frac{2\gamma\sigma^2L}{\mu n} + \frac{6L^2\gamma^2\beta^2\sigma^2}{\mu(1-\beta)} + \frac{18L^2\beta^2\gamma^2b^2}{\mu(1-\beta)^2} \\ \frac{24\beta^2\gamma^3\sigma^2L^3}{\mu(1-\beta)^2} + \frac{4nL\gamma^2\beta^2\sigma^2}{1-\beta} + \frac{12nL\beta^2\gamma^2b^2}{(1-\beta)^2} \end{bmatrix}, \end{aligned} \tag{93}$$

where (a) holds when

$$\gamma \leq \frac{1-\beta}{6\beta L}\sqrt{\frac{\mu}{L}}. \tag{94}$$

Inequality (93) implies that

$$\begin{aligned} \|(I-A)^{-1}s\|_1 &\leq \frac{2\gamma\sigma^2L}{\mu n} + \big(\frac{6L^2}{\mu} + \frac{24\gamma L^3}{\mu(1-\beta)} + 4nL\big)\frac{\gamma^2\beta^2\sigma^2}{1-\beta} + \big(\frac{18L^2}{\mu} + 12nL\big)\frac{\gamma^2\beta^2b^2}{(1-\beta)^2} \\ &\overset{(a)}{\leq} \frac{2\gamma\sigma^2L}{\mu n} + \big(\frac{6L}{\mu} + \frac{24L^3}{\mu^2} + 4nL\big)\frac{\gamma^2\beta^2\sigma^2}{1-\beta} + \big(\frac{18L^2}{\mu} + 12nL\big)\frac{\gamma^2\beta^2b^2}{(1-\beta)^2} \\ &\overset{(b)}{=} \frac{2\gamma\sigma^2L}{\mu n} + \frac{c_1\gamma^2\beta^2\sigma^2}{1-\beta} + \frac{c_2\gamma^2\beta^2b^2}{(1-\beta)^2} \end{aligned} \tag{95}$$

where (a) holds because γ satisfies (90), and (b) holds by introducing

$$c_1 := \frac{6L}{\mu} + \frac{24L^3}{\mu^2} + 4nL, \quad c_2 := \frac{18L^2}{\mu} + 12nL. \tag{96}$$

Substituting (95) to (92), and recalling the definition of $z^{(k)}$ in (86), we have (at iteration K)

$$\begin{aligned}
&\mathbb{E}[f(\bar{x}^{(K)})] - f^\star + L\mathbb{E}\|\mathbf{x}^{(K)} - \bar{\mathbf{x}}^{(K)}\|^2 \\
\leq\ & \Delta(1 - \frac{\mu\gamma}{4})^K + \frac{2\gamma\sigma^2 L}{\mu n} + \frac{c_1\gamma^2\beta^2\sigma^2}{1-\beta} + \frac{c_2\gamma^2\beta^2 b^2}{(1-\beta)^2} \\
\leq\ & \Delta\exp(-\frac{\mu\gamma K}{4}) + \frac{2\gamma\sigma^2 L}{\mu n} + \frac{c_1\gamma^2\beta^2\sigma^2}{1-\beta} + \frac{c_2\gamma^2\beta^2 b^2}{(1-\beta)^2}
\end{aligned} \tag{97}$$

where $\Delta = \mathbb{E}[f(\bar{x}^{(0)})] - f^\star + L\mathbb{E}\|\mathbf{x}^{(0)} - \bar{\mathbf{x}}^{(0)}\|^2 = \mathbb{E}[f(x^{(0)})] - f^\star$ if $x_i^{(0)} = 0$ for any $i \in [n]$. Next we let

$$\gamma_1 = \frac{4}{\mu K}\ln(\frac{\Delta\mu^2 nK}{\sigma^2 L}), \quad \gamma_2 = \frac{1-\beta}{4L}, \quad \gamma_3 = \frac{\mu(1-\beta)}{24n\beta L^2}, \quad \gamma = \min\{\gamma_1, \gamma_2, \gamma_3\}. \tag{98}$$

It is easy to verify that γ satisfies all conditions in (85), (89), (90), (91) and (94). With (98), we have

$$\begin{aligned}
\Delta\exp(-\frac{\mu\gamma K}{4}) &\leq \Delta\big[\exp(-\frac{\mu\gamma_1 K}{4}) + \exp(-\frac{\mu\gamma_2 K}{4}) + \exp(-\frac{\mu\gamma_3 K}{4})\big] \\
&= \mathcal{O}\Big(\frac{\sigma^2 L}{\mu^2 nK} + \Delta\exp(-\frac{\mu(1-\beta)K}{L}) + \Delta\exp(-\frac{\mu^2(1-\beta)K}{n\beta L^2})\Big)
\end{aligned} \tag{99}$$

and

$$\begin{aligned}
\frac{\gamma\sigma^2 L}{\mu n} + \frac{c_1\gamma^2\beta^2\sigma^2}{1-\beta} + \frac{c_2\gamma^2\beta^2 b^2}{(1-\beta)^2} &\leq \frac{\gamma_1\sigma^2 L}{\mu n} + \frac{c_1\gamma_1^2\beta^2\sigma^2}{1-\beta} + \frac{c_2\gamma_1^2\beta^2 b^2}{(1-\beta)^2} \\
&= \tilde{\mathcal{O}}\Big(\frac{\sigma^2 L}{\mu^2 nK} + \frac{c_1\beta^2\sigma^2}{\mu^2 K^2(1-\beta)} + \frac{c_2\beta^2 b^2}{\mu^2 K^2(1-\beta)^2}\Big)
\end{aligned} \tag{100}$$

Combining (97), (99) and (100), we achieve (83).

□

E.2.2 Convergence rate of MG-DSGD

As discussed in Appendix E.1.2, it holds that $\tilde{\sigma}^2 \triangleq \sigma^2/R$ and $\tilde{\beta} \triangleq \sqrt{2}(1-\sqrt{1-\beta})^R$. If we let

$$R = \left\lceil \frac{\max\{\ln(2), \ln(\frac{nL}{\mu}), \frac{1}{2}\ln(\frac{n}{L}\max\{c_1, \frac{c_2 b^2}{\sigma^2\sqrt{1-\beta}}\})\}}{\sqrt{1-\beta}} \right\rceil = \tilde{O}\left(\frac{1}{\sqrt{1-\beta}}\right), \tag{101}$$

we then have

$$\tilde{\beta} \triangleq \sqrt{2}(1-\sqrt{1-\beta})^R \leq \sqrt{2}e^{-\sqrt{1-\beta}R} \leq \min\left\{\frac{1}{\sqrt{2}}, \sqrt{2}\frac{\mu}{nL}, \left(2\frac{L}{n}\min\left\{\frac{1}{c_1}, \frac{\sigma^2\sqrt{1-\beta}}{b^2 c_2}\right\}\right)^{\frac{1}{2}}\right\}. \tag{102}$$

which immediately leads to $1-\tilde{\beta} = \Omega(1)$ and

$$\tilde{\beta}^2\max\{c_1, c_2 R\frac{b^2}{\sigma^2}\} = \tilde{O}\left(\tilde{\beta}^2\max\left\{c_1, c_2\frac{b^2}{\sqrt{1-\beta}\sigma^2}\right\}\right) = \tilde{O}(\frac{L}{n}). \tag{103}$$

Theorem 5 (Formal version). *Under the same assumptions as in Theorem 4 and setting R as in* (101)*, the convergence of A can be bounded for any loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_{L,\mu}$, and any $W \in \mathcal{W}_{n,\beta}$ by*

$$\mathbb{E}[f(\bar{x}^{(K)}) - f^\star] = \tilde{\mathcal{O}}\left(\frac{\sigma^2 L}{\mu^2 nT} + \Delta\exp(-\frac{\mu\sqrt{1-\beta}T}{L})\right). \tag{104}$$

where $\Delta = \mathbb{E}[f(x^{(0)})] - f^\star$.

Remark 11. *Comparing with the lower bound established in* (44)*, we find the upper bound* (104) *matches it tightly in terms of T, n, σ^2, and β, but is a little bit worse in dependence on L and μ. We will leave it as a future work to develop new algorithms that can match* (44) *tightly in all constants.*

Proof. The convergence analysis of DSGD applies to MG-DSGD by replacing σ^2 with $\tilde{\sigma}^2$, and β with $\tilde{\beta}$. As a result, we know from Lemma 13 that if the learning rate is set as

$$\gamma = \min\left\{\frac{4}{\mu K}\ln(\frac{\Delta\mu^2 nK}{\tilde{\sigma}^2 L}), \frac{1-\tilde{\beta}}{4L}, \frac{\mu(1-\tilde{\beta})}{24n\tilde{\beta}L^2}\right\} = \tilde{O}\left(\min\{\frac{1}{\mu\sqrt{1-\beta}T}, \frac{1}{L}\}\right) \tag{105}$$

it holds that

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 \\
=&\tilde{\mathcal{O}}\Big(\frac{\tilde{\sigma}^2 L}{\mu^2 nK} + \frac{c_1\tilde{\beta}^2\tilde{\sigma}^2}{\mu^2K^2(1-\tilde{\beta})} + \frac{c_2\tilde{\beta}^2 b^2}{\mu^2K^2(1-\tilde{\beta})^2} + \Delta\exp(-\frac{\mu(1-\tilde{\beta})K}{L}) + \Delta\exp(-\frac{\mu^2(1-\tilde{\beta})K}{n\tilde{\beta}L^2})\Big) \\
\overset{\tilde{\beta}\le\frac{1}{\sqrt{2}};\ T=KR\ge R}{=}& \mathcal{O}\left(\frac{\sigma^2 L}{\mu^2 nT} + \frac{c_1\tilde{\beta}^2\sigma^2}{\mu^2 T} + \frac{c_2\tilde{\beta}^2 Rb^2}{\mu^2 T} + \Delta\exp(-\frac{\mu T}{RL}) + \Delta\exp(-\frac{\mu^2 T}{n\tilde{\beta}RL^2})\right) \\
\overset{(103),(102)}{=}& \tilde{\mathcal{O}}\left(\frac{\sigma^2 L}{\mu^2 nT} + \Delta\exp(-\frac{\mu T}{RL})\right) \overset{(101)}{=} \tilde{\mathcal{O}}\left(\frac{\sigma^2 L}{\mu^2 nT} + \Delta\exp(-\sqrt{1-\beta}\frac{\mu T}{L})\right).
\end{aligned}$$

□

F Experiments

F.1 Performance in terms of epochs

Fig. 2 depicts the performance of several algorithms in terms of the communicated messages. In Fig. 3 we illustrate their performances in terms of epochs. It is observed that MG-DSGD achieves slightly better validation accuracy than other baselines in both CIFAR-10 and ImageNet dataset.

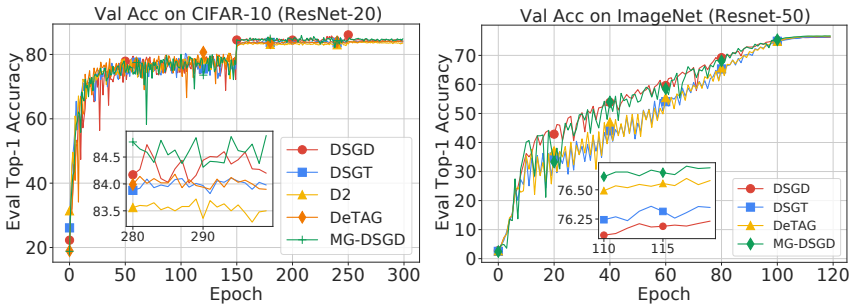

Figure 3: Convergence results in terms of validation accuracy. Left: CIFAR-10 with heterogeneous data ($\alpha = 1$); Right: ImageNet with homogeneous data.

F.2 Effects of accumulation rounds

We further investigate the performance of the proposed MG-DSGD and DeTAG over different accumulation rounds. The "M-" prefix indicates methods with momentum acceleration. As shown in Table 9, MG-DSGD reaches almost the same performance as DeTAG with less communications. Both methods have performance degradation as accumulation round scales up. We conjecture that while gradient accumulation, which amounts to using large-batch samples in gradient evaluation, can help in the optimization and training stage, it may hurt the generalization performance because gradient with less variance can lead the algorithm to a sharp local minimum. As a result, we recommend using MG-DSGD in applications that are friendly to large-batch training.

Table 9: Effects of round numbers for CIFAR-10 dataset

MODELS	RESNET18				RESNET20			
ROUNDS	2	3	4	5	2	3	4	5
M-DETAG	94.40±.13	93.50±.75	92.88±.50	92.57±.66	91.77±.25	91.50±.14	91.19±.26	90.62±.31
M-OURS	94.57±.05	93.95±.11	93.15±.25	92.01±.91	91.77±.09	91.17±.07	91.19±.13	90.44±.36
DETAG	93.17±.18	92.59±.03	92.26±.23	91.48±.27	88.97±.11	89.02±.08	88.77±.09	88.37±.24
OURS	93.75±.12	92.72±.20	92.43±.32	91.78±.28	89.18±.08	88.92±.10	88.80±.15	88.52±.20