On the Complexity of Finite-Sum Smooth Optimization under the Polyak–Łojasiewicz Condition

Yunyan Bai [1]   Yuxing Liu [1]   Luo Luo [1] [2]

Abstract

This paper considers the optimization problem of the form $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x})$, where $f(\cdot)$ satisfies the Polyak–Łojasiewicz (PL) condition with parameter μ and $\{f_i(\cdot)\}_{i=1}^{n}$ is L-mean-squared smooth. We show that any gradient method requires at least $\Omega(n + \kappa\sqrt{n}\log(1/\epsilon))$ incremental first-order oracle (IFO) calls to find an ϵ-suboptimal solution, where $\kappa \triangleq L/\mu$ is the condition number of the problem. This result nearly matches the upper bounds of IFO complexity for best-known first-order methods. We also study the problem of minimizing the PL function in the distributed setting such that the individuals $f_1(\cdot), \ldots, f_n(\cdot)$ are located on a connected network of n agents. We provide lower bounds of $\Omega(\kappa/\sqrt{\gamma}\log(1/\epsilon))$, $\Omega((\kappa + \tau\kappa/\sqrt{\gamma})\log(1/\epsilon))$ and $\Omega(n + \kappa\sqrt{n}\log(1/\epsilon))$ for communication rounds, time cost and local first-order oracle calls respectively, where $\gamma \in (0, 1]$ is the spectral gap of the mixing matrix associated with the network and $\tau > 0$ is the time cost of per communication round. Furthermore, we propose a decentralized first-order method that nearly matches above lower bounds in expectation.

1. Introduction

We study the optimization problem of the form

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}), \qquad (1)$$

where $\{f_i(\cdot)\}_{i=1}^{n}$ is L-mean-squared smooth but each $f_i(\cdot)$ is possibly nonconvex. The complexity of finding stationary points in Problem (1) has been widely studied in recent

years (Allen-Zhu & Hazan, 2016; Reddi et al., 2016; Zhou & Gu, 2019; Fang et al., 2018; Li et al., 2021). However, finding the global solution is intractable for the general non-convex smooth optimization (Nemirovskij & Yudin, 1983; Roux et al., 2012). This paper focuses on the minimization Problem (1) under the Polyak–Łojasiewicz condition (PL) (Polyak, 1963; Łojasiewicz, 1963), i.e., the objective function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$f(\mathbf{x}) - \inf_{\mathbf{y} \in \mathbb{R}^d} f(\mathbf{y}) \leq \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2$$

for any $\mathbf{x} \in \mathbb{R}^d$, where $\mu > 0$ is a constant. This inequality suggests the function value gap $f(\mathbf{x}) - f^*$ is dominated by the square of gradient norm, which leads to the gradient descent (GD) method linearly converging to the global minimum without the convexity (Karimi et al., 2016). The PL condition covers a lot of popular applications, such as deep neural networks (Liu et al., 2022; Allen-Zhu et al., 2019; Zeng et al., 2018), reinforcement learning (Fazel et al., 2018; Agarwal et al., 2021; Mei et al., 2020; Yuan et al., 2022b), optimal control (Bu et al., 2019; Fatkhullin & Polyak, 2021) and matrix recovery (Hardt & Ma, 2016; Li et al., 2018; Bi et al., 2022).

We typically solve the finite-sum optimization Problem (1) by incremental first-order oracle (IFO) methods (Agarwal & Bottou, 2015), which can access the pair $\{f_i(\mathbf{x}), \nabla f_i(\mathbf{x})\}$ for given $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$. This class of methods can leverage the structure of the objective (Johnson & Zhang, 2013; Defazio et al., 2014; Schmidt et al., 2017; Zhang et al., 2013) to iterate with one or mini-batch individual gradient, which is more efficient than the iteration with the full-batch gradient. IFO methods have received a lot of attention in recent years. For example, the stochastic variance reduced gradient (SVRG) methods with negative momentum (Qian et al., 2021; Allen-Zhu, 2017; Kovalev et al., 2020a; Woodworth & Srebro, 2016; Agarwal & Bottou, 2015) achieve the (near) optimal IFO complexity for convex optimization; the stochastic recursive gradient methods (Nguyen et al., 2017; Fang et al., 2018; Li et al., 2021; Pham et al., 2020; Wang et al., 2019; Zhou & Gu, 2019; Carmon et al., 2020) achieve the optimal IFO complexity for finding approximate stationary points in general nonconvex optimization. For the PL condition, Reddi et al. (2016); Lei et al. (2017)

[1] School of Data Science, Fudan University, Shanghai, China [2] Shanghai Key Laboratory for Contemporary Applied Mathematics, Shanghai, China. Correspondence to: Luo Luo <luoluo@fudan.edu.cn>.

proposed SVRG-type methods that find the ϵ-suboptimal solution within at most $\mathcal{O}((n + n^{2/3}\kappa) \log(1/\epsilon))$ IFO calls, where $\kappa \triangleq L/\mu$ is the condition number. Later, Zhou et al. (2019); Wang et al. (2019); Li et al. (2021) improved the upper bound to $\mathcal{O}((n+\kappa\sqrt{n}) \log(1/\epsilon))$ by stochastic recursive gradient estimator. Recently, Yue et al. (2023) established a tight lower complexity bound of the full-batch gradient methods for minimizing the PL function. However, the optimality of existing IFO methods for the finite-sum setting is still an open problem.

For large-scale optimization problems, we are interested in designing distributed algorithms. Specifically, we allocate individuals $f_1(\cdot), \ldots, f_n(\cdot)$ on n different agents and desire the agents to solve the problem collaboratively. We focus on the decentralized setting where agents are linked by a connected network so that each agent can only access its own local first-order oracle (LFO) and exchange messages with its neighbours. Besides the LFO complexity, we also require considering the communication complexity and the time complexity. It is worth noting that the time complexity in distributed optimization does not directly correspond to the weighted sum of the LFO complexity and the communication complexity, since some agents may skip the computation of local gradient during the iterations (Maranjyan et al., 2022; Mishchenko et al., 2022). Most of the work for decentralized optimization focus on the convex case (Shi et al., 2015; Nedic & Ozdaglar, 2009; Qu & Li, 2017; Scaman et al., 2017; Kovalev et al., 2020b; Song et al., 2023; Ye et al., 2023; Hendrikx et al., 2021; Li et al., 2022b) or the general nonconvex case (Luo & Ye, 2022; Li et al., 2022a; Xin et al., 2022; Sun et al., 2020; Lu & De Sa, 2021; Zhan et al., 2022). Recently, Yuan et al. (2022a) studied decentralized optimization under the PL condition in online setting, but their result is not tight to condition number.

In this paper, we provide the nearly tight lower bounds for the finite-sum optimization problem under the PL condition. We summarize our contributions as follows:

- We provide the lower bound of $\Omega(n + \kappa\sqrt{n} \log(1/\epsilon))$ for IFO complexity, which nearly matches the upper complexity bound of $\mathcal{O}((n + \kappa\sqrt{n}) \log(1/\epsilon))$ (Zhou et al., 2019; Wang et al., 2019; Li et al., 2021).
- We provide the lower bounds of $\Omega(\kappa/\sqrt{\gamma} \log(1/\epsilon))$, $\Omega((\kappa + \tau\kappa/\sqrt{\gamma}) \log(1/\epsilon))$ and $\Omega(n + \kappa\sqrt{n} \log(1/\epsilon))$ for communication complexity, time complexity and LFO complexity for decentralized setting, where γ is the spectral gap of the mixing matrix for the network and τ is the time cost of per communication round.
- We propose a decentralized first-order algorithm within the communication complexity of $\tilde{\mathcal{O}}(\kappa/\sqrt{\gamma} \log(1/\epsilon))$, the time complexity of $\tilde{\mathcal{O}}((\kappa + \tau\kappa/\sqrt{\gamma}) \log(1/\epsilon))$ and the LFO complexity of $\mathcal{O}((n + \kappa\sqrt{n}) \log(1/\epsilon))$ in expectation, nearly matching the lower bounds.

We compare our results with related work in Table 1-2.

2. Preliminaries

In this section, we formalize the problem setting and the complexity of the finite-sum optimization.

2.1. Notation and Assumptions

Given vector $\mathbf{x} \in \mathbb{R}^d$, we denote x_i as the i-th entry of $\mathbf{x}$ for $i \in [d]$ and denote $\text{supp}(\mathbf{x})$ as the index set for nonzero entries of $\mathbf{x}$. Given matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we denote $a_{i,j}$ as the (i, j)-th entry of $\mathbf{A}$ for $i \in [m]$ and $j \in [n]$. We let $\mathbf{1}$ be the vectors (or matrices) of all ones and $\mathbf{0}$ be the vector (or matrix) of all zeros. Additionally, we let $\mathbf{I}$ be the identity matrix and denote its i-th column as $\mathbf{e}_i$. We use $\|\cdot\|$ to present the Euclidean norm of a vector or the Frobenius norm of a matrix.

We consider the following assumptions for the finite-sum optimization Problem (1).

Assumption 2.1. We suppose the objective function $f(\cdot)$ is lower bounded, i.e, we have

$$f^* = \inf_{\mathbf{y} \in \mathbb{R}^d} f(\mathbf{y}) > -\infty.$$

Assumption 2.2. We suppose the function set $\{f_i(\cdot)\}_{i=1}^n$ is L-mean-squared smooth for some $L > 0$, i.e., we have

$$\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \leq L^2 \|\mathbf{x} - \mathbf{y}\|^2,$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Assumption 2.3. We suppose objective function $f(\cdot)$ is μ-PL for some $\mu > 0$, i.e., we have

$$f(\mathbf{x}) - \inf_{\mathbf{y} \in \mathbb{R}^d} f(\mathbf{y}) \leq \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2$$

for any $\mathbf{x} \in \mathbb{R}^d$.

Based on the above assumptions, we define condition number and ϵ-suboptimal solution of our problem.

Definition 2.4. We define $\kappa \triangleq L/\mu$ as the condition number of problem (1).

Definition 2.5. We say $\hat{\mathbf{x}}$ is an ϵ-suboptimal solution of Problem (1) if it holds that $f(\hat{\mathbf{x}}) - f^* \leq \epsilon$.

We use the notation $\tilde{\mathcal{O}}(\cdot)$ to hide the logarithmic dependence on condition number κ and individual number n.

For the decentralized setting, the individual $f_i(\cdot)$ presents the local function on the i-th agent. We define the aggregate variable and the corresponding aggregated gradient as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}(1) \\ \vdots \\ \mathbf{x}(n) \end{bmatrix} \in \mathbb{R}^{n \times d} \text{ and } \nabla\mathbf{F}(\mathbf{X}) = \begin{bmatrix} \nabla f_1(\mathbf{x}(1)) \\ \vdots \\ \nabla f_n(\mathbf{x}(n)) \end{bmatrix} \in \mathbb{R}^{n \times d}$$

Table 1. We present the IFO complexity for finding an ϵ-suboptimal solution of Problem (1) on single machine.

Algorithm	IFO	Reference
GD	$\mathcal{O}\big(\kappa n \log(1/\epsilon)\big)$	Karimi et al. (2016)
SVRG / SCSG	$\mathcal{O}\big((n + \kappa n^{2/3})\log(1/\epsilon)\big)$	Reddi et al. (2016); Lei et al. (2017)
SPIDER / PAGE	$\mathcal{O}\big((n + \kappa\sqrt{n})\log(1/\epsilon)\big)$	Zhou et al. (2019); Wang et al. (2019); Li et al. (2021)
Lower Bound	$\Omega\big(n + \kappa\sqrt{n}\log(1/\epsilon)\big)$	Corollary 3.6

respectively, where $\mathbf{x}(i) \in \mathbb{R}^{1 \times d}$ is the local variable on the i-th agent. For given $\mathbf{X} \in \mathbb{R}^{n \times d}$, we also introduce the mean vector $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{1}^\top \mathbf{X} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}(i) \in \mathbb{R}^{1 \times d}$. For ease of presentation, we let the input of a function can also be organized as a row vector, such as $f(\bar{\mathbf{x}})$ and $f_i(\mathbf{x}(i))$.

We describe one communication round by multiplying the mixing matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ on the aggregated variable. We give the following assumption for matrix $\mathbf{W}$.

Assumption 2.6. We suppose mixing matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ has the following properties: (a) We have $\mathbf{W} = \mathbf{W}^\top$, $\mathbf{W}\mathbf{1}_n = \mathbf{1}_n$ and $\mathbf{0} \preceq \mathbf{W} \preceq \mathbf{I}$; (b) The entry of $\mathbf{W}$ holds that $w_{i,j} > 0$ if and only if the i-th agent and the j-th agent are connected or $i = j$, otherwise it holds that $w_{ij} = 0$. (c) The spectral gap of $\mathbf{W}$ is lower bounded by $\gamma \in (0, 1]$, i.e., it holds that $1 - \lambda_2(\mathbf{W}) \geq \gamma$ for some $\gamma \in (0, 1]$, where $\lambda_2(\mathbf{W})$ is the second-largest eigenvalue of $\mathbf{W}$.

2.2. The Finite-Sum Optimization

The complexity of first-order methods for solving the finite-sum optimization Problem (1) on a single machine mainly depends on the number of access to the incremental first-order oracle (IFO), which is defined as follows (Agarwal & Bottou, 2015; Woodworth & Srebro, 2016).

Definition 2.7. The incremental first-order oracle (IFO) takes the input $i \in [n]$ and $\mathbf{x}$, and returns the pair $(f_i(\mathbf{x}), \nabla f_i(\mathbf{x}))$.

We focus on the complexity of linear span methods and formally define the IFO algorithm as follows.

Definition 2.8. An IFO algorithm for given initial point $\mathbf{x}^0$ is defined as a measurable mapping from functions $\{f_i(\cdot)\}_{i=1}^n$ to an infinite sequence of point and index pairs $\{(\mathbf{x}^t, i_t)\}_{t=0}^\infty$ with random variable $i_t \in [n]$, which satisfies

$$\mathbf{x}^t \in \mathrm{Lin}\big(\{\mathbf{x}^0, \ldots, \mathbf{x}^{t-1}, \nabla f_{i_0}(\mathbf{x}^0), \ldots, \nabla f_{i_{t-1}}(\mathbf{x}^{t-1})\}\big),$$

where $\mathrm{Lin}(\cdot)$ denotes the linear span and i_t denotes the index of individual function chosen at the t-th step.

For the distributed optimization over a network of n agents, the i-th agent can only perform the computation on its local function $f_i(\cdot)$ directly. Hence, we describe the complexity

of computational cost by the number of access to the local first-order oracle (LFO).

Definition 2.9. The local first-order oracle (IFO) takes the input $i \in [n]$ and $\mathbf{x}$, and returns the pair $(f_i(\mathbf{x}), \nabla f_i(\mathbf{x}))$.

Recall that agents on the network can only communicate with their neighbours, which means the agent in decentralized algorithms cannot arbitrarily establish the linear space of all local gradients. Additionally, one iteration of the algorithm allows a mini-batch of agents to compute their local gradient in parallel. Therefore, besides the LFO complexity, we also need to separately consider the communication complexity and the time complexity. This motivates the following definition for decentralized first-order oracle algorithm (DFO) (Scaman et al., 2017).

Definition 2.10. A decentralized first-order oracle (DFO) algorithm over a network of n agents satisfies the following constraints:

- **Local memory:** Each agent i can store past values in a local memory $\mathcal{M}_{i,s}$ at time $s > 0$. These values can be accessed and used at time s by running the algorithm on agent i. Additionally, for all $i \in [n]$, we have

$$\mathcal{M}_i^s \subset \mathcal{M}_{\mathrm{comp},i}^s \bigcup \mathcal{M}_{\mathrm{comm},i}^s,$$

  where $\mathcal{M}_{\mathrm{comp},i}^s$ and $\mathcal{M}_{\mathrm{comm},i}^s$ are the values that come from the computation and communication respectively.

- **Local computation:** Each agent i can access its local first-order oracle $\{f_i(\mathbf{x}), \nabla f_i(\mathbf{x})\}$ for given $\mathbf{x} \in \mathcal{M}_{i,s}$ at time s. That is, for all $i \in [n]$, we have

$$\mathcal{M}_{\mathrm{comp},i}^s = \mathrm{Lin}\big(\{\mathbf{x}, \nabla f_i(\mathbf{x}) : x \in \mathcal{M}_i^{s-1}\}\big).$$

- **Local communication:** Each agent i can share its value to all or part of its neighbors at time s. That is, for all $i \in [n]$, we have

$$\mathcal{M}_{\mathrm{comm},i}^s = \mathrm{Lin}\left(\bigcup_{j \in \mathrm{nbr}(i)} \mathcal{M}_j^{s-\tau}\right),$$

  where $\mathrm{nbr}(i)$ is the set consisting of the indices for the neighbours of agent i and $\tau < s$.

- **Output value:** Each agent i can specify one vector in its memory as the local output of the algorithm at time s. That is, for all $i \in [n]$, we have $\mathbf{x}_i^s \in \mathcal{M}_i^s$.

Table 2. We present the complexity for finding an ϵ-suboptimal solution of Problem (1) in decentralized setting.

Algorithm	Communication	Time	LFO	Reference
DGD-GT	$\tilde{\mathcal{O}}\left(\kappa/\sqrt{\gamma}\log(1/\epsilon)\right)$	$\tilde{\mathcal{O}}\left(\kappa\left(1+\tau/\sqrt{\gamma}\right)\log(1/\epsilon)\right)$	$\mathcal{O}\left(\kappa n\log(1/\epsilon)\right)$	Theorem 5.1
DRONE	$\tilde{\mathcal{O}}\left(\kappa/\sqrt{\gamma}\log(1/\epsilon)\right)$	$\tilde{\mathcal{O}}\left(\kappa\left(1+\tau/\sqrt{\gamma}\right)\log(1/\epsilon)\right)$	$\mathcal{O}\left((n+\kappa\sqrt{n}\,)\log(1/\epsilon)\right)$	Corollary 5.3
Lower Bound	$\Omega\left(\kappa/\sqrt{\gamma}\log(1/\epsilon)\right)$	$\Omega\left(\kappa\left(1+\tau/\sqrt{\gamma}\right)\log(1/\epsilon)\right)$	$\Omega\left(n+\kappa\sqrt{n}\log(1/\epsilon)\right)$	Theorem 4.3, Corollary 4.4

3. The Lower Bound on IFO Complexity

This section provides the lower bound on IFO complexity to show the optimality (up to logarithmic factors) of existing first-order methods (Zhou et al., 2019; Wang et al., 2019; Li et al., 2021). Without loss of generality, we always assume the IFO algorithm iterates with the initial point $\mathbf{x}^{(0)} = \mathbf{0}$ in our analysis for the lower bound. Otherwise, we can take the functions $\{f_i(\mathbf{x} + \mathbf{x}^{(0)})\}_{i=1}^n$ into consideration.

We first consider the case of $n = \mathcal{O}(\kappa^2)$. We introduce the functions $\psi_\theta : \mathbb{R} \to \mathbb{R}$, $q_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ and $g_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ provided by Yue et al. (2023), that is

$$\psi_\theta(x) = \begin{cases} \frac{1}{2}x^2, & x \le \frac{31}{32}\theta, \\ \frac{1}{2}x^2 - 16(x - \frac{31}{32}\theta)^2, & \frac{31}{32}\theta < x \le \theta, \\ \frac{1}{2}x^2 - \frac{1}{32}\theta^2 + 16(x - \frac{33}{32}\theta)^2, & \theta < x \le \frac{33}{32}\theta, \\ \frac{1}{2}x^2 - \frac{1}{32}\theta^2, & x > \frac{33}{32}\theta, \end{cases}$$

$$q_{T,t}(\mathbf{x}) = \frac{1}{2}\sum_{i=0}^{t-1}\left(\left(\frac{7}{8}x_{iT} - x_{iT+1}\right)^2 + \sum_{j=1}^{T-1}(x_{iT+j+1} - x_{iT+j})^2\right),$$

and $\quad g_{T,t}(\mathbf{x}) = q_{T,t}(\mathbf{b} - \mathbf{x}) + \sum_{i=1}^{Tt}\psi_{b_i}(b_i - x_i),$

where we define $x_0 = 0$ and $\mathbf{b} \in \mathbb{R}^{Tt}$ with $b_{kT+\tau} = (7/8)^k$ for $k \in \{0\} \cup [T-1]$ and $\tau \in [T]$. We can verify that

$$g_{T,t}^* \triangleq \inf_{\mathbf{y} \in \mathbb{R}^{Tt}} g_{T,t}(\mathbf{y}) = 0.$$

The following lemma shows the function $g_{T,t}$ holds the zero-chain property (Nesterov, 2018; Carmon & Duchi, 2020) and describes its smoothness, PL parameter and optimal function value gap, which results the tight lower bound of full-batch first-order methods (Yue et al., 2023, Section 4).

Lemma 3.1. *The function $g_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ holds that:*

(a) *For any $\mathbf{x} \in \mathbb{R}^{Tt}$ satisfying $\mathrm{supp}(\mathbf{x}) \subseteq \{1, 2, \cdots, k\}$, it holds $\mathrm{supp}(\nabla g_{T,t}(\mathbf{x})) \subseteq \{1, 2, \cdots, k+1\}$.*

(b) *The function $g_{T,t}$ is 37-smooth.*

(c) *The function $g_{T,t}$ is $1/(aT)$-PL with $a = 19708$.*

(d) *The function $g_{T,t}$ satisfies that $g_{T,t}(\mathbf{0}) - g_{T,t}^* \le 3T$.*

(e) *For any $\delta < 0.01$, $t = 2\lfloor \log_{8/7} 2/(3\delta) \rfloor$ and $\mathbf{x} \in \mathbb{R}^{Tt}$ satisfying $\mathrm{supp}(\mathbf{x}) \subseteq \{1, 2, \cdots, Tt/2\}$, it holds that $g_{T,t}(\mathbf{x}) - g_{T,t}^* > 3T\delta$.*

We can establish the mean-squared smooth functions by the composition of orthogonal transformation. Compared with the study on convex and general nonconvex problem (Carmon et al., 2020; Zhou & Gu, 2019), we present the following lemma by additionally considering the PL condition.

Lemma 3.2. *Given a function $g : \mathbb{R}^m \to \mathbb{R}$ that is $\hat{L}$-smooth and $\hat{\mu}$-PL, define $f_i(\mathbf{x}) = g(\mathbf{U}^{(i)}\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^{mn}$, $i \in [n]$ and $\mathbf{U}^{(i)} = [\mathbf{e}_{(i-1)m+1}, \cdots, \mathbf{e}_{im}]^\top \in \mathbb{R}^{m \times mn}$. Then the function set $\{f_i : \mathbb{R}^{mn} \to \mathbb{R}\}_{i=1}^n$ is $\hat{L}/\sqrt{n}$-mean-squared smooth, and the function $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is $\hat{\mu}/n$-PL with $f(\mathbf{0}) - \inf_{\mathbf{y} \in \mathbb{R}} f(\mathbf{y}) = g(\mathbf{0}) - \inf_{\mathbf{y} \in \mathbb{R}} g(\mathbf{y})$.*

To achieve the hard instance functions with the desired smoothness and PL parameters, we also require the scaling lemma as follows.

Lemma 3.3. *Suppose the function $g : \mathbb{R}^m \to \mathbb{R}$ is $\hat{L}$-smooth, $\hat{\mu}$-PL and has lower bound $g^* = \inf_{\mathbf{y} \in \mathbb{R}^m} g(\mathbf{y})$, then the function $\hat{g}(\mathbf{x}) = \alpha g(\beta\mathbf{x})$ is $\alpha\beta^2\hat{L}$-smooth, $\alpha\beta^2\hat{\mu}$-PL and satisfies that $\hat{g}(\mathbf{0}) - \hat{g}^* = \alpha(g(\mathbf{0}) - g^*)$ for any $\alpha, \beta > 0$, where $\hat{g}^* = \inf_{\mathbf{y} \in \mathbb{R}^m} \hat{g}(\mathbf{y})$.*

Based on above lemmas 3.1, 3.2 and 3.3, we provide the lower bound of $\Omega(\kappa\sqrt{n}\log(1/\epsilon))$ on IFO complexity for large κ.

Theorem 3.4. *For any L, μ, n, Δ and ϵ with $\epsilon < 0.005\Delta$ and $L \ge 37a\sqrt{n}\mu$, there exists L-mean-squared smooth function set $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ with $d = \mathcal{O}(\kappa\sqrt{n}\log(1/\epsilon))$ such that the function $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is μ-PL with $f(\mathbf{x}^0) - f^* \le \Delta$. In order to find an ϵ-suboptimal solution of problem $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, any IFO algorithm needs at least $\Omega(\kappa\sqrt{n}\log(1/\epsilon))$ IFO calls.*

We present the proof sketch of Theorem 3.4 as follows and defer the details in Appendix A.3. Specifically, we take the function $g_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ with $T = \lfloor L/(37a\sqrt{n}\mu) \rfloor$ and $t = 2\lfloor \log_{8/7} \Delta/(3\epsilon) \rfloor$, and let $\hat{g}(\mathbf{x}) = \alpha g_{T,t}(\beta\mathbf{x})$ with $\alpha = \Delta/(3T)$ and $\beta = \sqrt{3\sqrt{n}LT/(37\Delta)}$. We also define function set $\{f_i : \mathbb{R}^{nTt} \to \mathbb{R}\}_{i=1}^n$ by following Lemma 3.2 with $g(\mathbf{x}) = \hat{g}(\mathbf{x})$ and $m = Tt$. Then statements (b)-(d) of Lemma 3.1 and the scaling property shown in Lemma 3.3 means such construction results that the condition number of the problem is $\kappa = L/\mu$ and the optimal function value is gap Δ. Finally, the statements (a) and (e) of Lemma 3.1 lead to the desired lower complexity bound.

4

We then consider the case of $n = \Omega(\kappa^2)$. Following the analysis by Li et al. (2021, Theorem 2), we establish the lower bound of $\Omega(n)$ for IFO complexity by introducing $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ with $f_i(\mathbf{x}) = c\langle \mathbf{u}_i, \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x}\|^2$, where $\mathbf{u}_i = [\mathbb{I}(\lceil 1/(2n) \rceil = i), \cdots, \mathbb{I}(\lceil 2n^2/(2n) \rceil = i)]^\top \in \mathbb{R}^{2n^2}$, $d = 2n^2$, $c = \sqrt{L\Delta}$ and $\mathbb{I}(\cdot)$ is the indicator function. We can verify $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ is L-mean-squared smooth and $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is μ-PL with the optimal function value gap Δ. Additionally, each $f_i(\cdot)$ holds zero-chain property, which leads to the lower bound of $\Omega(n)$ on IFO complexity. We formally present this result in the following theorem and leave the detailed proof in Appendix A.4.

Theorem 3.5. *For any L, μ, n, Δ and ϵ with $\epsilon < \Delta/2$ and $L \geq \mu$, there exists L-mean-squared smooth function set $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ with $d = \mathcal{O}(n^2)$ such that the function $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is μ-PL with $f(\mathbf{x}^0) - f^* \leq \Delta$. In order to find an ϵ-suboptimal solution of problem $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, any IFO algorithm needs at least $\Omega(n)$ IFO calls.*

We combine Theorem 3.4 and 3.5 to achieve the lower bound on IFO complexity for the first-order finite-sum optimization under the PL condition.

Corollary 3.6. *For any L, μ, n, Δ and ϵ with $\epsilon < 0.005\Delta$ and $L \geq \mu$. there exists L-mean-squared smooth function set $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ with $d = \mathcal{O}(n^2 + \kappa\sqrt{n}\log(1/\epsilon))$ such that the function $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is μ-PL and satisfies $f(\mathbf{x}^0) - f^* \leq \Delta$. In order to find an ϵ-suboptimal solution of problem $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, any IFO algorithm needs at least $\Omega(n + \kappa\sqrt{n}\log(1/\epsilon))$ IFO calls.*

Noticing that the lower bound of $\Omega(n + \kappa\sqrt{n}\log(1/\epsilon))$ on IFO complexity shown in Corollary 3.6 nearly matches the upper bound of $\mathcal{O}((n + \kappa\sqrt{n})\log(1/\epsilon))$ achieved by stochastic recursive gradient algorithms (Zhou et al., 2019; Wang et al., 2019; Li et al., 2021).

4. Lower Bounds in Decentralized Setting

This section provides lower bounds for the decentralized setting. The main idea of our construction is splitting the function $g_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ (defined in Section 3) as follows

$$g_{T,t}(\mathbf{x}) = q_1(\mathbf{b} - \mathbf{x}) + q_2(\mathbf{b} - \mathbf{x}) + r(\mathbf{x}),$$

where we define the functions $q_1 : \mathbb{R}^{Tt} \to \mathbb{R}$, $q_2 : \mathbb{R}^{Tt} \to \mathbb{R}$ and $r : \mathbb{R}^{Tt} \to \mathbb{R}$ as

$$q_1(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{Tt/2}(x_{2i-1} - x_{2i})^2,$$

$$q_2(\mathbf{x}) = \frac{1}{2}\sum_{i=0}^{t-1}\left[\left(\frac{7}{8}x_{iT} - x_{iT+1}\right)^2 + \sum_{j=iT/2+1}^{(i+1)T/2-1}(x_{2j} - x_{2j+1})^2\right],$$

and $r(\mathbf{x}) = \sum_{i=1}^{Tt}\psi_{b_i}(b_i - x_i),$

Based on the above decomposition for $g_{T,t}$, we can establish a hard instance of n individual functions for communication complexity and time complexity. We let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be the graph associated to the network of the agents, where the node set $\mathcal{V} = \{1, \ldots, n\}$ corresponds to the n agents and the edge set $\mathcal{E} = \{(i,j) : \text{node } i \text{ and node } j \text{ are connected}\}$ describes the topology for the network of agents. For given a subset $\mathcal{C} \subseteq \mathcal{V}$, we define the function $h_i^{\mathcal{C}}(\mathbf{x}) : \mathbb{R}^{Tt} \to \mathbb{R}$ as

$$h_i^{\mathcal{C}}(\mathbf{x}) = \begin{cases} \frac{r(\mathbf{x})}{n} + \frac{q_1(\mathbf{b} - \mathbf{x})}{|\mathcal{C}|}, & i \in \mathcal{C}, \\ \frac{r(\mathbf{x})}{n} + \frac{q_2(\mathbf{b} - \mathbf{x})}{|\mathcal{C}_\sigma|}, & i \in \mathcal{C}_\sigma, \\ \frac{r(\mathbf{x})}{n}, & \text{otherwise}, \end{cases}$$

where $\mathcal{C}_\sigma = \{v \in \mathcal{V} : \text{dis}(\mathcal{C}, v) \geq \sigma\}$ and $\text{dis}(\mathcal{C}, v)$ is the distance between set $\mathcal{C}$ and node v.

We can verify the function sets $\{h_i^{\mathcal{C}} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ has the following properties.

Lemma 4.1. *We define $h(\cdot) = \frac{1}{n}\sum_{i=1}^n h_i^{\mathcal{C}}(\cdot)$, then we have:*

(a) *The function sets $\{h_i^{\mathcal{C}} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ is mean-squared smooth with parameter $33/n + \max\{2/|\mathcal{C}|, 2/|\mathcal{C}_\sigma|\}$.*

(b) *The function $h : \mathbb{R}^{Tt} \to \mathbb{R}$ is $1/(anT)$-PL.*

(c) *The function holds $h(\mathbf{0}) - \inf_{\mathbf{y} \in \mathbb{R}^{Tt}} h(\mathbf{y}) \leq 3T/n$.*

Now we provide the lower bounds for communication complexity and time complexity based on the scaling on the functions $\{h_i^{\mathcal{C}} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$.

Lemma 4.2. *We let $f_i(\mathbf{x}) = \alpha h_i^{\mathcal{C}}(\beta\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^{Tt}$ and $t = 2\lfloor \log_{8/7}(2/3\delta) \rfloor$ for any $i \in [n]$ and some $\alpha, \beta > 0$. For given $\delta < 0.01$, any DFO algorithm takes at least $Tt\sigma/2$ communication complexity and $Tt(1 + \sigma\tau)/2$ time complexity to achieve an $3\alpha T\delta/n$-suboptimal solution of problem $\min_{\mathbf{x} \in \mathbb{R}^{Tt}} \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$.*

For given spectral gap $\gamma \in (0, 1]$, we consider the linear graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with the node set $\mathcal{V} = \{1, \ldots, n\}$ and the edge set $\mathcal{E} = \{(i,j) : |i - j| = 1, i \in \mathcal{V} \text{ and } j \in \mathcal{V}\}$. Combining with the function $f_i(\mathbf{x}) = \alpha h_i^{\mathcal{C}}(\beta\mathbf{x})$ defined in Lemma 4.2, we achieve the lower bounds of communication complexity and time complexity as follows.

Theorem 4.3. *For any L, μ, Δ, γ and ϵ with $L \geq 194a\mu$, $\gamma \in (0, 1]$ and $\epsilon < 0.01\Delta$, there exist matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ with $1 - \lambda_2(\mathbf{W}) \geq \gamma$ and L-mean-squared smooth function set $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ with $d = \mathcal{O}(\kappa\log(1/\epsilon))$ such that $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is μ-PL with $f(\mathbf{0}) - f^* \leq \Delta$ and $\gamma(\mathbf{W}) = \gamma$. In order to find an ϵ-suboptimal solution of problem $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, any DFO algorithm needs at least $\Omega(\kappa/\sqrt{\gamma}\log(1/\epsilon))$ communication rounds and $\Omega(\kappa\log(1/\epsilon)(1 + \tau/\sqrt{\gamma}))$ time cost.*

The lower complexity bound on LFO complexity for the decentralized setting can be achieved by applying Corollary 3.6 on a fully connected network.

Algorithm 1 GD

1: **Input:** initial point $\mathbf{x}^0 \in \mathbb{R}^d$, iteration number T and stepsize $\eta > 0$
2: **for** $t = 0, 1, \ldots, T - 1$ **do**
3: $\quad \mathbf{x}^{t+1} = \mathbf{x}^t - \eta \nabla f(\mathbf{x}^t)$
4: **end for**
5: **Output:** $\mathbf{x}^T$

Algorithm 2 AccGossip($\mathbf{Y}^0, \mathbf{W}, K$)

1: **Initialize:** $\mathbf{Y}^{-1} = \mathbf{Y}^0$
2: $\eta_y = 1/\left(1 + \sqrt{1 - \lambda_2^2(\mathbf{W})}\right)$
3: **for** $k = 0, 1, \ldots, K$ **do**
4: $\quad \mathbf{Y}^{k+1} = (1 + \eta_y)\mathbf{W}\mathbf{Y}^k - \eta_y \mathbf{Y}^{k-1}$
5: **end for**
6: **Output:** $\mathbf{Y}^K$

Corollary 4.4. *For any $L, \mu, n, \Delta, \gamma$ and ϵ with $\epsilon < 0.005\Delta$, $L \geq \mu$ and $\gamma \in (0, 1]$, there exist mixing matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ with $1 - \lambda_2(\mathbf{W}) \geq \gamma$ and L-mean-squared smooth function set $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ with $d = \mathcal{O}(n^2 + \kappa\sqrt{n}\log(1/\epsilon))$ such that the function $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is μ-PL and satisfies $f(\mathbf{x}^0) - f^* \leq \Delta$. In order to find an ϵ-suboptimal solution of problem $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, any DFO algorithm needs at least $\Omega\left(n + \kappa\sqrt{n}\log(1/\epsilon)\right)$ LFO calls.*

5. Decentralized First-Order Algorithms

It is well-known that GD achieves the linear convergence for minimizing the PL function (Karimi et al., 2016) on a single machine. For decentralized optimization, it is natural to integrate GD with gradient tracking (Shi et al., 2015; Nedic & Ozdaglar, 2009; Qu & Li, 2017) and Chebyshev acceleration (Arioli & Scott, 2014; Scaman et al., 2017; Song et al., 2023; Ye et al., 2023), leading to Algorithm 3 which is called decentralized gradient descent with gradient tracking (DGD-GT). The following theorem shows that the communication complexity and the time complexity of DGD-GT nearly match the lower bounds shown in Corollary 4.3.

Theorem 5.1. *We suppose that Assumption 2.1–2.6 hold, then running Algorithm 3 (DGD-GT) with appropriate parameters setting achieves $\mathbb{E}[f(\mathbf{x}^{\mathrm{out}}) - f^*] \leq \epsilon$ within communication complexity of $\tilde{\mathcal{O}}\left(\kappa/\sqrt{\gamma}\log(1/\epsilon)\right)$, time complexity of $\tilde{\mathcal{O}}\left(\kappa(1 + \tau/\sqrt{\gamma})\log(1/\epsilon)\right)$ and LFO complexity of $\mathcal{O}\left(n\kappa\log(1/\epsilon)\right)$ in expectation.*

However, the upper bound on LFO complexity of DGD-GT (Algorithm 3) shown in Theorem 5.1 does not match the lower bound of $\Omega\left(n + \kappa\sqrt{n}\log(1/\epsilon)\right)$ provided by Corollary 4.4. Recall that the optimal IFO methods for the non-distributed setting are based on the stochastic recursive 'gra-

Algorithm 3 DGD-GT

1: **Input:** initial point $\bar{\mathbf{x}}_0 \in \mathbb{R}^{1 \times d}$, iteration number T, stepsize $\eta > 0$ and communication numbers K
2: $\mathbf{X}^0 = \mathbf{1}\bar{\mathbf{x}}^0$
3: $\mathbf{S}^0 = \nabla \mathbf{F}(\mathbf{X}^0)$
4: **for** $t = 0, \ldots, T - 1$ **do**
5: $\quad \mathbf{X}^{t+1} = \mathrm{AccGossip}(\mathbf{X}^t - \eta\mathbf{S}^t, \mathbf{W}, K)$
6: $\quad \mathbf{S}^{t+1} = \mathrm{AccGossip}(\mathbf{S}^t + \nabla\mathbf{F}(\mathbf{X}^{t+1}) - \nabla\mathbf{F}(\mathbf{X}^t), \mathbf{W}, K)$
7: **end for**
8: **Output:** uniformly sample $\mathbf{x}^{\mathrm{out}}$ from $\{\mathbf{x}^T(i)\}_{i=1}^n$

Algorithm 4 DRONE

1: **Input:** initial point $\bar{\mathbf{x}}^0 \in \mathbb{R}^{1 \times d}$, mini-batch size b, iteration number T, stepsize $\eta > 0$, probability $p, q \in (0, 1]$ and communication numbers K.
2: $\mathbf{X}^0 = \mathbf{1}\bar{\mathbf{x}}^0$
3: $\mathbf{S}^0 = \mathbf{G}^0 = \nabla\mathbf{F}(\mathbf{X}^0)$
4: **for** $t = 0, \ldots, T - 1$ **do**
5: $\quad \zeta^t \sim \mathrm{Bernoulli}(p)$
6: $\quad [\xi_1^t, \cdots, \xi_n^t]^\top \sim \mathrm{Multinomial}(b, q\mathbf{1})$
7: $\quad \mathbf{X}^{t+1} = \mathrm{AccGossip}(\mathbf{X}^t - \eta\mathbf{S}^t, \mathbf{W}, K)$
8: $\quad$ **parallel for** $i = 1, \ldots, n$ **do**
9: $\quad\quad$ **if** $\zeta^t = 1$ **then**
10: $\quad\quad\quad \mathbf{g}^{t+1}(i) = \nabla f_i(\mathbf{x}^{t+1}(i))$
11: $\quad\quad$ **else**
12: $\quad\quad\quad \mathbf{g}^{t+1}(i) = \mathbf{g}^t(i) + \frac{\xi_i^t}{bq}\left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right)$
13: $\quad\quad$ **end if**
14: $\quad$ **end parallel for**
15: $\quad \mathbf{S}^{t+1} = \mathrm{AccGossip}(\mathbf{S}^t + \mathbf{G}^{t+1} - \mathbf{G}^t, \mathbf{W}, K)$
16: **end for**
17: **Output:** uniformly sample $\mathbf{x}^{\mathrm{out}}$ from $\{\mathbf{x}^T(i)\}_{i=1}^n$

---

dient (Li et al., 2021; Zhou et al., 2019; Wang et al., 2019). We borrow this idea to construct the recursive gradient with respect to local agents, i.e., we update the local gradient estimator by

$$\mathbf{g}^{t+1}(i) = \begin{cases} \nabla f_i(\mathbf{x}^{t+1}(i)), & \zeta^t = 1, \\ \mathbf{g}^t(i) + \frac{\xi_i^t}{bq}\left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right), & \zeta^t = 0, \end{cases}$$

where we introduce random variables $\zeta^t \sim \mathrm{Bernoulli}(p)$ and $[\xi_1^t, \cdots, \xi_n^t]^\top \sim \mathrm{Multinomial}(b, q\mathbf{1})$ with some small probabilities p and q which encourage only few of agents compute local gradients in most of iterations. Similar to the procedure of DGD-GT (Algorithm 3), we can also introduce steps of gradient tracking and Chebyshev acceleration to improve communication efficiency. Finally, we achieve decentralized recursive local gradient descent (DRONE) method, which is formally presented in Algorithm 4.

We analyze the complexity of DRONE by the following Lyapunov function

$$\Phi^t = \mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + \alpha U^t + \beta V^t + LC^t,$$

where $\alpha = 2\eta/p$, $\beta = 8L\rho^2 n\eta^2$,

$$U^t = \mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^n \left(\mathbf{g}^t(i) - \nabla f_i(\mathbf{x}^t(i)) \right) \right\|^2,$$

$$V^t = \frac{1}{n}\mathbb{E}\|\mathbf{G}^t - \nabla\mathbf{F}(\mathbf{X}^t)\|^2,$$

$$\text{and} \quad C^t = \mathbb{E}\|\mathbf{X}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \eta^2\mathbb{E}\|\mathbf{S}^t - \mathbf{1}\bar{\mathbf{s}}^t\|^2.$$

Compared with the analysis of Luo & Ye (2022); Li et al. (2022a); Xin et al. (2022) for general nonconvex case, we additionally introduce the term of f^* into our Lyapunov function to show the desired linear convergence under the PL condition, i.e., we can show that

$$\Phi^T \leq (1 - \mu\eta)^T \Phi^0$$

by taking $\eta = \Theta(1/L)$. We present the convergence result of DRONE formally in the following theorem and corollary.

Theorem 5.2. *We suppose that Assumption 2.1–2.6 hold, then running DRONE (Algorithm 4) with parameters setting*

$$p \in \left[\frac{1}{n+1}, \frac{1}{2}\right], \quad b \in \left[\frac{1-p}{p}, n\right], \quad \eta \leq \min\left\{\frac{1}{20L}, \frac{p}{2\mu}\right\},$$

$$q = \frac{1}{n}, \quad K = \left\lceil \frac{\sqrt{2}(4 + \log n)}{(\sqrt{2}-1)\sqrt{\gamma}} \right\rceil \quad \text{and} \quad T \geq \left\lceil \frac{1}{\mu\eta} \log \frac{\Phi^0}{\epsilon} \right\rceil$$

achieves output satisfying $\mathbb{E}[f(\mathbf{x}^{\text{out}}) - f^*] \leq \epsilon$.

Corollary 5.3. *Under the setting of Theorem 5.2, running DRONE (Algorithm 4) by specifically taking*

$$p = \frac{1}{\min\{\sqrt{n}, \kappa\} + 1}, \quad b = \left\lceil \frac{1-p}{p} \right\rceil, \quad \eta = \min\left\{\frac{1}{20L}, \frac{p}{2\mu}\right\},$$

$$q = \frac{1}{n}, \quad K = \left\lceil \frac{\sqrt{2}(4 + \log n)}{(\sqrt{2}-1)\sqrt{\gamma}} \right\rceil \quad \text{and} \quad T = \left\lceil \frac{1}{\mu\eta} \log \frac{\Phi^0}{\epsilon} \right\rceil$$

achieves output satisfying $\mathbb{E}[f(\mathbf{x}^{\text{out}}) - f^*] \leq \epsilon$ within communication complexity of $\tilde{\mathcal{O}}\big(\kappa/\sqrt{\gamma}\log(1/\epsilon)\big)$, time complexity of $\tilde{\mathcal{O}}\big(\kappa(1 + \tau/\sqrt{\gamma})\log(1/\epsilon)\big)$ and LFO complexity of $\mathcal{O}\big((n + \kappa\sqrt{n})\log(1/\epsilon)\big)$ in expectation.

Note that the setting of $p = 1/(\min\{\sqrt{n}, \kappa\} + 1)$ leads to $b = \mathcal{O}(\min\{\sqrt{n}, \kappa\})$, $\eta = \Theta(1/L)$ and $T = \Theta(\kappa \log(1/\epsilon))$, guarantees the algorithm nearly match the LFO lower bound (Corollary 4.4) in both cases of $n = \mathcal{O}(\kappa^2)$ and $n = \Omega(\kappa^2)$.

As a comparison, the analysis of PAGE (Li et al., 2021) (for single machine optimization) takes $p = \Theta(1/\sqrt{n})$, which leads to that $b = \Theta(\sqrt{n})$, $\eta = \Theta(\min\{1/L, 1/(\mu\sqrt{n})\})$ and $T = \Theta((\kappa + \sqrt{n})\log(1/\epsilon))$. If we directly apply

these parameters to DRONE, it will result LFO complexity of $\mathcal{O}\big((n + \kappa\sqrt{n})\log(1/\epsilon)\big)$, communication complexity of $\tilde{\mathcal{O}}\big((\kappa + \sqrt{n})\log(1/\epsilon)/\sqrt{\gamma}\big)$ and time complexity of $\tilde{\mathcal{O}}\big((\kappa + \sqrt{n})(1 + \tau/\sqrt{\gamma})\log(1/\epsilon)\big)$ in expectation. In the case of $n = \Omega(\kappa^2)$, such communication complexity and time complexity do not match the corresponding lower bounds (Theorem 4.3). Intuitively, our analysis for DRONE considers the larger stepsize $\eta = \Theta(1/L) \geq \Theta(1/(\mu\sqrt{n}))$ than PAGE when $n = \Omega(\kappa^2)$, which is important to reduce the iteration numbers $T = \lceil(1/\mu\eta)\log(\Phi^0/\epsilon)\rceil$, also reduce the overall communication rounds KT and the overall time cost $(1 + K\tau)T$.

6. Numerical Experiments

We conduct numerical experiments to compare DRONE with centralized gradient descent (CGD) and DGD-GT, where CGD is a distributed extension of GD in client-server networks. Please see Appendix D for details.

We test the algorithms on the following three problems:

- Hard instance: We follow the instance in the proof of Theorem 4.3 (Appendix B.3) and specifically let

$$f_i(\mathbf{x}) = \frac{16}{3} h_i^{\mathcal{C}}\big(\sqrt{12a}\mathbf{x}\big)$$

  for formulation (1). We set $T = 2$, $t = 72$, $\mathcal{C} = \{1\}$ and $\sigma = 29$ for $h_i^{\mathcal{C}} : \mathbb{R}^{Tt} \to \mathbb{R}$.

- Linear regression: We consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^m \ell_j(\mathbf{x}) \text{ with } \ell_j(\mathbf{x}) = (\mathbf{a}_j^\top \mathbf{x} - b_j)^2 \quad (2)$$

  where $\mathbf{a}_j \in \mathbb{R}^d$ is the feature vector of the j-th sample and $b_j \in \mathbb{R}$ is its label. We allocate the m individual loss on the n agents, which leads to

$$f_i(\mathbf{x}) = \frac{n}{m} \sum_{j=\lceil m/n\rceil(i-1)+1}^{\min\{\lceil m/n\rceil i, m\}} (\mathbf{a}_j^\top \mathbf{x} - b_j)^2. \quad (3)$$

  We evaluate the algorithms on dataset "DrivFace" ($m = 606$, $d = 921{,}600$) (Diaz-Chito et al., 2016) for this problem.

- Logistic regression: The objective function and local functions of this problem are similar to the counterparts in formulation (2)–(3), but replace the loss function by

$$\hat{\ell}_j(\mathbf{x}) = \log\big(1 + \exp(-b_j \mathbf{a}_j^\top \mathbf{x})\big).$$

  and require $b_j \in \{1, -1\}$. We evaluate the algorithms on dataset "RCV1" ($m = 20{,}242$, $d = 47{,}236$) (Diaz-Chito et al., 2016) for this problem.

For all the above problems, we set $n = 32$ and use a linear graph for the network of DGD-GT and DRONE, leading to that $\gamma = (1 - \cos(\pi/32)) / (1 + \cos(\pi/32)) \approx 0.0024$.
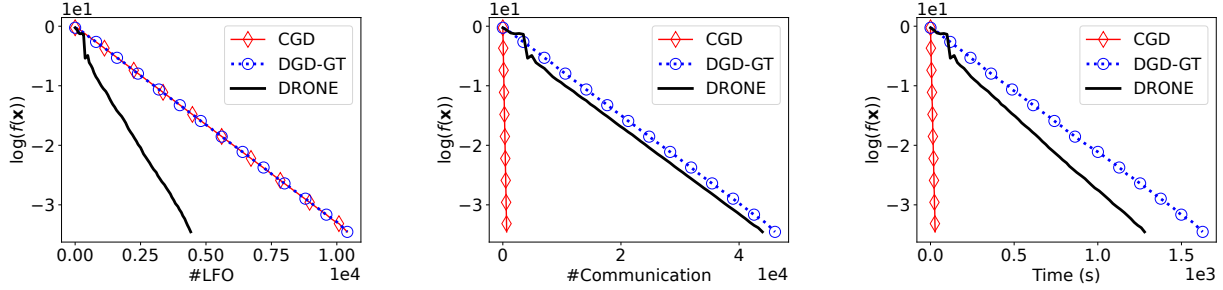
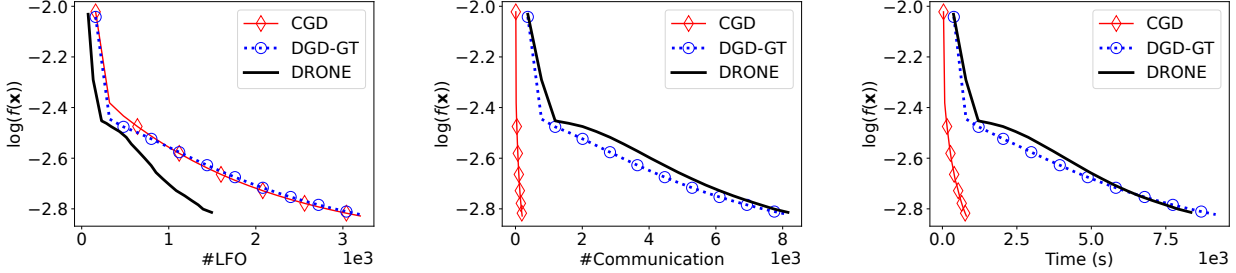*Figure 1.* The results for the hard instance in the proof of Theorem 4.3.

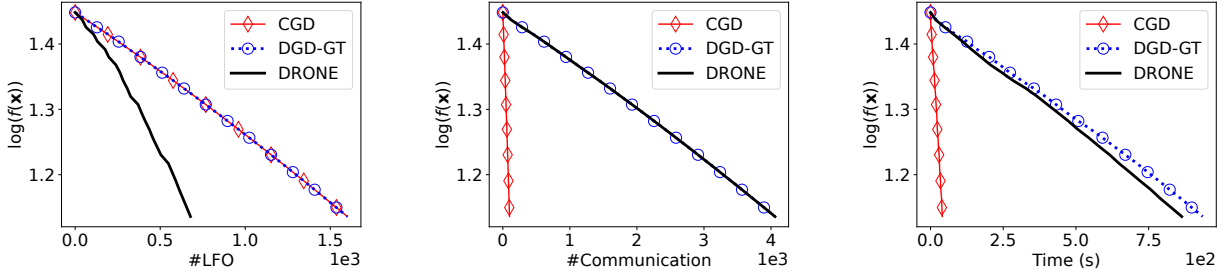*Figure 2.* The results for linear regression on dataset "DrivFace".

Figure 3. The results for logistic regression on dataset "RCV1".

We present empirical results for CGD, DGD-GT and DRONE on problems of hard instance, linear regression and logistic regression in Figure 1, 2 and 3, which includes the comparisons on LFO calls, communication rounds and running time.

We can observe that DRONE requires significantly fewer LFO calls, since we have shown only DRONE matches the lower complexity bound on LFO calls. We also observed CGD needs much fewer communication rounds than DGD-GT and DRONE, which also leads to less running time. This is because the client-server framework in CGD does not suffer from the consensus error which cannot be avoided in decentralized optimization. It also validates our theoretical analysis that the linear graph heavily affects the convergence rate of decentralized algorithms. Additionally, DGD-GT and DRONE have comparable communication rounds and running time for all of these problems, which also supports our theoretical results (see Table 2).

7. Conclusion

We provide the lower complexity bound for smooth finite-sum optimization under the PL condition, which implies the upper bound of IFO complexity archived by existing first-order methods (Wang et al., 2019; Li et al., 2021; Zhou et al., 2019) is nearly tight. We also construct the lower bounds of communication complexity, time complexity and LFO complexity for minimizing the PL function in the distributed setting and verify their tightness by proposing decentralized recursive local gradient descent.

In future work, we would like to study the lower bound in more general stochastic settings where the objective (or local functions) has the form of expectation (Yuan et al., 2022a; Arjevani et al., 2022). We are also interested in extending our results to address the functions that satisfy the Kurdyka–Łojasiewicz inequality (Bolte et al., 2014; 2007; Attouch & Bolte, 2009; Zhou et al., 2018; Fatkhullin et al., 2022; Jiang & Li, 2022).

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A. and Bottou, L. A lower bound for the optimization of finite sums. In *ICML*, 2015.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *ICML*, 2016.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019.

Arioli, M. and Scott, J. Chebyshev acceleration of iterative refinement. *Numerical Algorithms*, 66(3):591–608, 2014.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pp. 1–50, 2022.

Attouch, H. and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009.

Bi, Y., Zhang, H., and Lavaei, J. Local and global linear convergence of general low-rank matrix recovery problems. In *AAAI*, 2022.

Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth

problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

Bu, J., Mesbahi, A., Fazel, M., and Mesbahi, M. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

Carmon, Y. and Duchi, J. C. First-order methods for non-convex quadratic minimization. *SIAM Review*, 62(2):395–436, 2020.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.

Diaz-Chito, K., Hernández-Sabaté, A., and López, A. M. A reduced feature set for driver head pose estimation. *Applied Soft Computing*, 45:98–107, 2016.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *NeurIPS*, 2018.

Fatkhullin, I. and Polyak, B. Optimizing static linear feedback: Gradient method. *SIAM Journal on Control and Optimization*, 59(5):3887–3911, 2021.

Fatkhullin, I., Etesami, J., He, N., and Kiyavash, N. Sharp analysis of stochastic optimization under global Kurdyka-Łojasiewicz inequality. In *NeurIPS*, 2022.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. In *ICML*, 2018.

Hardt, M. and Ma, T. Identity matters in deep learning. In *ICLR*, 2016.

Hendrikx, H., Bach, F., and Massoulie, L. An optimal algorithm for decentralized finite-sum optimization. *SIAM Journal on Optimization*, 31(4):2753–2783, 2021.

Jiang, R. and Li, X. Hölderian error bounds and Kurdyka-Łojasiewicz inequality for the trust region subproblem. *Mathematics of Operations Research*, 47(4):3025–3050, 2022.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *ECML/PKDD*, 2016.

Kovalev, D., Horváth, S., and Richtárik, P. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *ALT*, 2020a.

Kovalev, D., Salim, A., and Richtárik, P. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *NeurIPS*, 2020b.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via SCSG methods. In *NIPS*, 2017.

Li, B., Li, Z., and Chi, Y. DESTRESS: Computation-optimal and communication-efficient decentralized nonconvex finite-sum optimization. *SIAM Journal on Mathematics of Data Science*, 4(3):1031–1051, 2022a.

Li, H., Lin, Z., and Fang, Y. Variance reduced EXTRA and DIGing and their optimal acceleration for strongly convex decentralized optimization. *Journal of Machine Learning Research*, 23:1–41, 2022b.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *COLT*, 2018.

Li, Z., Bao, H., Zhang, X., and Richtárik, P. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *ICML*, 2021.

Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Łojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Lu, Y. and De Sa, C. Optimal complexity in decentralized training. In *ICML*, 2021.

Luo, L. and Ye, H. An optimal stochastic algorithm for decentralized nonconvex finite-sum optimization. *arXiv preprint arXiv:2210.13931*, 2022.

Maranjyan, A., Safaryan, M., and Richtárik, P. Grad-Skip: Communication-accelerated local gradient methods with better computational complexity. *arXiv preprint arXiv:2210.16402*, 2022.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *ICML*, 2022.

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

Nemirovskij, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Nesterov, Y. *Lectures on convex optimization*, volume 137. Springer, 2018.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Qian, X., Qu, Z., and Richtárik, P. L-SVRG and L-Katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22(1):4991–5039, 2021.

Qu, G. and Li, N. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *ICML*, 2016.

Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, 2012.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *ICML*, 2017.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shi, W., Ling, Q., Wu, G., and Yin, W. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Song, Z., Shi, L., Pu, S., and Yan, M. Optimal gradient tracking for decentralized optimization. *Mathematical Programming*, pp. 1–53, 2023.

Sun, H., Lu, S., and Hong, M. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *ICML*, 2020.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. Spi-derBoost and momentum: Faster variance reduction algorithms. In *NeurIPS*, 2019.

Woodworth, B. E. and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *NIPS*, 2016.

Xin, R., Khan, U. A., and Kar, S. Fast decentralized non-convex finite-sum optimization with recursive variance reduction. *SIAM Journal on Optimization*, 32(1):1–28, 2022.

Ye, H., Luo, L., Zhou, Z., and Zhang, T. Multi-consensus decentralized accelerated gradient descent. *Journal of Machine Learning Research*, 24(306):1–50, 2023.

Yuan, K., Huang, X., Chen, Y., Zhang, X., Zhang, Y., and Pan, P. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. In *NeurIPS*, 2022a.

Yuan, R., Gower, R. M., and Lazaric, A. A general sample complexity analysis of vanilla policy gradient. In *AISTATS*, 2022b.

Yue, P., Fang, C., and Lin, Z. On the lower bound of minimizing Polyak-Łojasiewicz functions. In *COLT*, 2023.

Zeng, J., Ouyang, S., Lau, T. T.-K., Lin, S., and Yao, Y. Global convergence in deep learning with variable splitting via the Kurdyka-Łojasiewicz property. *arXiv preprint arXiv:1803.00225*, 9, 2018.

Zhan, W., Wu, G., and Gao, H. Efficient decentralized stochastic gradient descent method for nonconvex finite-sum optimization problems. In *AAAI*, 2022.

Zhang, L., Mahdavi, M., and Jin, R. Linear convergence with condition number independent access of full gradients. In *NIPS*, 2013.

Zhou, D. and Gu, Q. Lower bounds for smooth nonconvex finite-sum optimization. In *ICML*, 2019.

Zhou, P., Yuan, X.-T., and Feng, J. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. In *AISTATS*, 2019.

Zhou, Y., Wang, Z., and Liang, Y. Convergence of cubic regularization for nonconvex optimization under KL property. In *NeurIPS*, 2018.

We provide the proofs for results of Section 3, 4 and 5 in Appendix A, B and C respectively. In Section D, we present more details for our experiments.

A. The Proofs for Section 3

Without the loss of generality, we always assume the IFO algorithm iterates with initial point $\mathbf{x}^{(0)} = \mathbf{0}$. Otherwise, we can take functions $\{f_i(\mathbf{x} + \mathbf{x}^{(0)})\}_{i=1}^n$ into consideration.

Recall that we have defined the functions $\psi_\theta : \mathbb{R} \to \mathbb{R}$, $q_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ and $g_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ as (Yue et al., 2023)

$$
\psi_\theta(x) = \begin{cases}
\frac{1}{2}x^2, & x \le \frac{31}{32}\theta, \\
\frac{1}{2}x^2 - 16(x - \frac{31}{32}\theta)^2, & \frac{31}{32}\theta < x \le \theta, \\
\frac{1}{2}x^2 - \frac{1}{32}\theta^2 + 16(x - \frac{33}{32}\theta)^2, & \theta < x \le \frac{33}{32}\theta, \\
\frac{1}{2}x^2 - \frac{1}{32}\theta^2, & x > \frac{33}{32}\theta,
\end{cases}
\tag{4}
$$

$$
q_{T,t}(\mathbf{x}) = \frac{1}{2} \sum_{i=0}^{t-1} \left(\left(\frac{7}{8}x_{iT} - x_{iT+1} \right)^2 + \sum_{j=1}^{T-1} (x_{iT+j+1} - x_{iT+j})^2 \right),
$$

and

$$
g_{T,t}(\mathbf{x}) = q_{T,t}(\mathbf{b} - \mathbf{x}) + \sum_{i=1}^{Tt} \psi_{b_i}(b_i - x_i).
\tag{5}
$$

where $x_0 = 0$ and $\mathbf{b} \in \mathbb{R}^{Tt}$ with $b_{kT+\tau} = (7/8)^k$ for $k \in \{0\} \cup [t-1]$ and $\tau \in [T]$. We can verify that

$$
g_{T,t}^* \triangleq \inf_{\mathbf{y} \in \mathbb{R}^{Tt}} g_{T,t}(\mathbf{y}) = 0.
$$

A.1. The Proof of Lemma 3.2

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{mn}$, the smoothness of $g : \mathbb{R}^m \to \mathbb{R}$ implies

$$
\begin{aligned}
\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| &= \left\| (\mathbf{U}^{(i)})^\top \nabla g(U^{(i)}\mathbf{x}) - (\mathbf{U}^{(i)})^\top \nabla g(U^{(i)}\mathbf{y}) \right\| \\
&= \left\| \nabla g(\mathbf{U}^{(i)}\mathbf{x}) - \nabla g(\mathbf{U}^{(i)}\mathbf{y}) \right\| \\
&\le \hat{L} \left\| \mathbf{U}^{(i)}(\mathbf{x} - \mathbf{y}) \right\| \\
&\le \hat{L} \|\mathbf{x} - \mathbf{y}\|,
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 &= \frac{1}{n} \sum_{i=1}^n \left\| (\mathbf{U}^{(i)})^\top \nabla g(\mathbf{U}^{(i)}\mathbf{x}) - (\mathbf{U}^{(i)})^\top \nabla g(\mathbf{U}^{(i)}\mathbf{y}) \right\|^2 \\
&= \frac{1}{n} \sum_{i=1}^n \left\| \nabla g(\mathbf{U}^{(i)}\mathbf{x}) - \nabla g(\mathbf{U}^{(i)}\mathbf{y}) \right\|^2 \\
&= \frac{\hat{L}^2}{n} \sum_{i=1}^n \left\| \mathbf{U}^{(i)}(\mathbf{x} - \mathbf{y}) \right\|^2 \\
&\le \frac{\hat{L}^2}{n} \|\mathbf{x} - \mathbf{y}\|^2.
\end{aligned}
$$

This implies each $f_i : \mathbb{R}^{mn} \to \mathbb{R}$ is $\hat{L}$-smooth and $\{f_i : \mathbb{R}^{mn} \to \mathbb{R}\}_{i=1}^n$ is $\hat{L}/\sqrt{n}$-average smooth.

For any $\mathbf{x} \in \mathbb{R}^{mn}$, the PL condition of $g : \mathbb{R}^m \to \mathbb{R}$ implies

$$
\begin{aligned}
\|\nabla f(\mathbf{x})\|^2 &= \frac{1}{n^2} \Big\| \sum_{i=1}^{n} \nabla f_i(\mathbf{x}) \Big\|^2 \\
&= \frac{1}{n^2} \Big\| \sum_{i=1}^{n} (\mathbf{U}^{(i)})^\top \nabla g(\mathbf{U}^{(i)}\mathbf{x}) \Big\|^2 \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \big\| \nabla g(\mathbf{U}^{(i)}\mathbf{x}) \big\|^2 \\
&\geq \frac{\hat{\mu}}{n^2} \sum_{i=1}^{n} \Big(g(\mathbf{U}^{(i)}\mathbf{x}) - \inf_{\mathbf{x}\in\mathbb{R}^{mn}} g(\mathbf{U}^{(i)}\mathbf{x}) \Big) \\
&= \frac{\hat{\mu}}{n}(f(\mathbf{x}) - f^*),
\end{aligned}
$$

which means $f : \mathbb{R}^{mn} \to \mathbb{R}$ is $\hat{\mu}/n$-PL.

Consider the facts $f(\mathbf{0}) = g(\mathbf{0})$ and

$$
\inf_{\mathbf{x}\in\mathbb{R}^{mn}} \sum_{i=1}^{n} g(\mathbf{U}^{(i)}\mathbf{x}) = \sum_{i=1}^{n} \inf_{\mathbf{x}\in\mathbb{R}^{mn}} g(\mathbf{U}^{(i)}\mathbf{x}) = n \inf_{\mathbf{x}\in\mathbb{R}^m} g(\mathbf{x}),
$$

then we have

$$
f(\mathbf{0}) - \inf_{\mathbf{y}\in\mathbb{R}} f(\mathbf{y}) = g(\mathbf{0}) - \inf_{\mathbf{y}\in\mathbb{R}} g(\mathbf{y}).
$$

$\square$

A.2. The Proof of Lemma 3.3

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{mn}$, the smoothness of $g : \mathbb{R}^m \to \mathbb{R}$ implies

$$
\begin{aligned}
\|\nabla \hat{g}(\mathbf{x}) - \nabla \hat{g}(\mathbf{y})\| &= \alpha\beta \|\nabla g(\beta\mathbf{x}) - \nabla g(\beta\mathbf{y})\| \\
&\leq \alpha\beta\hat{L} \|\beta\mathbf{x} - \beta\mathbf{y}\| \\
&\leq \alpha\beta^2\hat{L} \|\mathbf{x} - \mathbf{y}\|,
\end{aligned}
$$

which means $\hat{g}$ is $\alpha\beta^2\hat{L}$-smooth.

For any $\mathbf{x} \in \mathbb{R}^{mn}$, the PL condition of $g : \mathbb{R}^m \to \mathbb{R}$ implies

$$
\begin{aligned}
\|\nabla \hat{g}(\mathbf{x})\|^2 &= \alpha^2\beta^2 \|\nabla g(\beta\mathbf{x})\|^2 \\
&\geq 2\alpha^2\beta^2\hat{\mu}(g(\beta\mathbf{x}) - g^*) \\
&= 2\alpha\beta^2\hat{\mu}(\hat{g}(\mathbf{x}) - \hat{g}^*),
\end{aligned}
$$

which means $\hat{g}$ is $\alpha\beta^2\hat{\mu}$-PL.

We can verify that $\hat{g}(\mathbf{0}) = \alpha g(\mathbf{0})$ and $\hat{g}^* = \alpha g^*$, which means

$$
\hat{g}(\mathbf{0}) - \hat{g}^* = \alpha(g(\mathbf{0}) - g^*).
$$

$\square$

A.3. The Proof of Theorem 3.4

Proof. We first take $g_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ by following equation (5) with

$$
T = \left\lfloor \frac{L}{37a\sqrt{n}\mu} \right\rfloor \qquad \text{and} \qquad t = 2\left\lceil \log_{7/8} \frac{3\epsilon}{\Delta} \right\rceil.
$$

The statements (b), (c) and (d) of Lemma 3.1 means the function $g_{T,t}$ is 37-smooth, $1/(aT)$-PL and satisfies

$$g_{T,t}(\mathbf{0}) - g_{T,t}^* \leq 3T.$$

We apply Lemma 3.3 with

$$g(\mathbf{x}) = g_{T,t}(\mathbf{x}), \quad m = Tt, \quad \alpha = \frac{\Delta}{3T} \quad \text{and} \quad \beta = \sqrt{\frac{3\sqrt{n}LT}{37\Delta}},$$

which means the function $\hat{g}(\mathbf{x}) = \alpha g_{T,t}(\beta\mathbf{x})$ is $37\alpha\beta^2$-smooth, $\alpha\beta^2/(aT)$-PL and satisfies

$$\alpha g_{T,t}(\mathbf{0}) - \alpha g_{T,t}^* \leq 3\alpha T.$$

Then we apply Lemma 3.2 with

$$g(\mathbf{x}) = \hat{g}(\mathbf{x}) = \alpha g_{T,t}(\beta\mathbf{x}) \qquad \text{and} \qquad m = Tt,$$

which achieves $f_i(\mathbf{x}) = \alpha g_{T,t}(\beta\mathbf{U}^{(i)}\mathbf{x})$ and $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ such that $\{f_i : \mathbb{R}^{nTt} \to \mathbb{R}\}_{i=1}^n$ is $37\alpha\beta^2/\sqrt{n}$-mean-squared smooth and $f : nTt$ is $\alpha\beta^2/(anT)$-PL with

$$f(\mathbf{x}^0) - f^* = \alpha g_{T,t}(\mathbf{0}) - \alpha g_{T,t}^* \leq 3\alpha T.$$

The choice of $\alpha = \Delta/(3T)$ and $\beta = \sqrt{3\sqrt{n}LT/(37\Delta)}$ and condition $L \geq 37a\sqrt{n}\mu$ implies

$$\frac{37\alpha\beta^2}{\sqrt{n}} = L, \qquad \frac{\alpha\beta^2}{anT} \geq \mu \qquad \text{and} \qquad 3\alpha T = \Delta.$$

Therefore, the function set $\{f_i\}_{i=1}^n$ is L-average smooth and the function f is μ-PL with $f(\mathbf{x}^0) - f^* \leq \Delta$.

Let $\delta = 2\epsilon/\Delta$, then we can write $t = 2\lfloor \log_{8/7} 2/(3\delta) \rfloor$. Moreover, the assumption $\epsilon < 0.005\Delta$ means $\delta < 0.01$. Then the statement (e) of Lemma 3.1 and definition $f_i(\mathbf{x}) = \alpha g_{T,t}(\beta\mathbf{U}^{(i)}\mathbf{x})$ implies if $\mathbf{x} \in \mathbb{R}^{nTt}$ satisfies

$$\text{supp}(\mathbf{U}^{(i)}\mathbf{x}) \subseteq \{1, 2, \cdots, Tt/2\},$$

then

$$\begin{aligned}
f_i(\mathbf{x}) - \alpha g_{T,t}^* &= \alpha g_{T,t}(\beta U^{(i)}\mathbf{x}) - \alpha g_{T,t}^* \\
&> 3\alpha T\delta \\
&= 2\epsilon.
\end{aligned}$$

Now we show that any IFO algorithm require at least $\lfloor nTt/4 \rfloor + 1 = \Omega(\kappa\sqrt{n}\log(1/\epsilon))$ IFO calls to achieve an ϵ-suboptimal solution $\hat{\mathbf{x}}$ of the problem. We consider the vector $\mathbf{x} \in \mathbb{R}^{nTt}$ achieved by an IFO algorithm with at most $\lfloor nTt/4 \rfloor$ IFO calls. The zero-chain property of g_{Tt} (statement (a) of Lemma 3.1) means the vector $\mathbf{x}$ has at most $\lfloor nTt/4 \rfloor$ non-zero entries. We partition $\mathbf{x} \in \mathbb{R}^{nTt}$ into n vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)} \in \mathbb{R}^{Tt}$ such that $\mathbf{y}^{(i)} = \mathbf{U}^{(i)}\mathbf{x} \in \mathbb{R}^{Tt}$. Then there at least $\lceil n/2 \rceil$ vectors in $\{\mathbf{y}^{(i)}\}_{i=1}^n$ such that each of them has at least $Tt/2$ zero entries. The zero-chain property means there exists index set $\mathcal{I} \subseteq [n]$ with $|\mathcal{I}| \geq \lceil n/2 \rceil$ such that each $i \in \mathcal{I}$ satisfies $y_{Tt/2+1}^{(i)} = \cdots = y_{Tt}^{(i)} = 0$. Therefore, the statement (e) of Lemma 3.1 implies

$$f_i(\mathbf{x}) - \alpha g_{T,t}^* > 2\epsilon,$$

which leads to

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}) - f^* &= \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}) - \alpha g_{T,t}^* \\
&\geq \frac{1}{n}\sum_{i\in\mathcal{I}}(f_i(\mathbf{x}) - \alpha g_{T,t}^*) \\
&> \frac{1}{n} \cdot \lceil \frac{n}{2} \rceil \cdot 2\epsilon \\
&\geq \epsilon.
\end{aligned}$$

Hence, finding an ϵ-suboptimal solution of the problem requires at least $\lfloor nTt/4 \rfloor + 1 = \Omega(\kappa\sqrt{n}\log(1/\epsilon))$ IFO calls.

$\square$

A.4. The Proof of Theorem 3.5

Proof. We prove this theorem by following Li et al. (2021, Theorem 2). For any $i \in [n]$, we define $f_i : \mathbb{R}^d \to \mathbb{R}$ as

$$f_i(\mathbf{x}) = c \langle \mathbf{u}_i, \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x}\|^2,$$

where

$$c = \sqrt{L\Delta}, \qquad d = 2n^2, \qquad \mathbf{u}_i = \left[\mathbb{I}\left(\left\lceil \frac{1}{2n} \right\rceil = i \right), \mathbb{I}\left(\left\lceil \frac{2}{2n} \right\rceil = i \right) \cdots, \mathbb{I}\left(\left\lceil \frac{2n^2}{2n} \right\rceil = i \right) \right]^\top \in \mathbb{R}^d$$

and $\mathbb{I}(\cdot)$ is the indicator function.

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$\nabla f_i(\mathbf{x}) = c\mathbf{u}_i + L\mathbf{x}$$

for any $i \in [n]$, which implies

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 &= \frac{1}{n} \sum_{i=1}^n \|(c\mathbf{u}_i + L\mathbf{x}) - (c\mathbf{u}_i + L\mathbf{y})\|^2 \\
&= \frac{1}{n} \sum_{i=1}^n \|L(\mathbf{x} - \mathbf{y})\|^2 \\
&= L^2 \|\mathbf{x} - \mathbf{y}\|^2.
\end{aligned}
$$

Hence, we conclude $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ is L-mean-squared smooth.

We also have $\nabla^2 f(\mathbf{x}) = L\mathbf{I} \succeq \mu\mathbf{I}$ for any $\mathbf{x} \in \mathbb{R}^d$. Hence, the function $\nabla^2 f(\mathbf{x})$ is μ-strongly convex, also is μ-PL.

We have

$$
\begin{aligned}
f^* &= \frac{1}{n} \sum_{i=1}^n \left(c\langle \mathbf{u}_i, \mathbf{x}^* \rangle + \frac{L}{2} \|\mathbf{x}^*\|^2 \right) \\
&= \frac{c}{n} \sum_{i=1}^n \langle \mathbf{u}_i, \mathbf{x}^* \rangle + \frac{L}{2} \|\mathbf{x}^*\|^2 \\
&= -\frac{c^2}{2Ln^2} \left\| \sum_{i=1}^n \mathbf{u}_i \right\|^2 \\
&= -\frac{c^2}{L},
\end{aligned}
$$

where $\mathbf{x}^* = -(c/Ln)\mathbf{1}$ is the minima of $f : \mathbb{R}^d \to \mathbb{R}$. Then the optimal function value gap holds

$$f(\mathbf{x}^0) - f^* = 0 - f^* = \frac{c^2}{L} = \Delta.$$

We consider any IFO algorithm with initial point $\mathbf{x}^0 = \mathbf{0}$. After t IFO calls, Definition 2.8 implies

$$\mathbf{x}^t \in \text{Lin}\left(\{ \nabla f_{i_0}(\mathbf{x}^0), \ldots, \nabla f_{i_{t-1}}(\mathbf{x}^{t-1}) \} \right) = \text{Lin}\left(\{ \mathbf{u}_{i_0}, \ldots, \mathbf{u}_{i_{t-1}} \} \right),$$

where $i_\tau \in [n]$ is the index of individual which is accessed at the τ-th IFO calls. Since each $\mathbf{u}_{i_\tau}$ has $2n$ nonzero entries, any vector $\mathbf{x} \in \mathbb{R}^d$ achieved by at most $n/2$ IFO calls has at least

$$d - \frac{n}{2} \cdot 2n = 2n^2 - n^2 = n^2$$

15

zero entries. Let $\mathcal{I}_0 = \{j \in [2n^2] : x_j = 0\}$, then we have $|\mathcal{I}| \geq n^2$. Based on the construction of f_i and $\mathbf{u}_i$, we have

$$
\begin{aligned}
f(\mathbf{x}) - f^* &= \frac{1}{n} \sum_{i=1}^{n} \left(c\langle \mathbf{u}_i, \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x}\|_2 \right) - \left(-\frac{c^2}{L} \right) \\
&= \sum_{j=1}^{2n^2} \left(\frac{c}{n} x_j + \frac{L}{2} x_j^2 + \frac{c^2}{2Ln^2} \right) \\
&= \sum_{j \in \mathcal{I}_0} \left(\frac{c}{n} x_j + \frac{L}{2} x_j^2 + \frac{c^2}{2Ln^2} \right) + \sum_{j \notin \mathcal{I}_0} \left(\frac{c}{n} x_j + \frac{L}{2} x_j^2 + \frac{c^2}{2Ln^2} \right) \\
&\geq n^2 \cdot \frac{c^2}{2Ln^2} + \sum_{j \notin \mathcal{I}_0} \left(x_j + \frac{c}{nL} \right)^2 \\
&\geq \frac{\Delta}{2} > \epsilon,
\end{aligned}
$$

Hence, achieving an ϵ-suboptimal solution requires at least $n/2 + 1 = \Omega(n)$ IFO calls. $\qquad\square$

A.5. The Proof of Corollary 3.6

Proof. This result can be achieved by directly combining Theorem 3.4 and 3.5. $\qquad\square$

B. The Proofs for Section 4

Without loss of generality, we always assume that all agents start with the internal memory of null space, i.e., we have $\mathcal{M}_i^0 = \{\mathbf{0}\}$ for any $i \in [n]$.

The main idea in our lower bound analysis splitting the function $g_{T,t} : \mathbb{R}^{Tt} \to \mathbb{R}$ defined in equation (5) by introducing the functions $q_1 : \mathbb{R}^{Tt} \to \mathbb{R}$, $q_2 : \mathbb{R}^{Tt} \to \mathbb{R}$ and $r : \mathbb{R}^{Tt} \to \mathbb{R}$ as

$$
q_1(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{Tt/2} (x_{2i-1} - x_{2i})^2, \tag{6}
$$

$$
q_2(\mathbf{x}) = \frac{1}{2} \sum_{i=0}^{t-1} \left[\left(\frac{7}{8} x_{iT} - x_{iT+1} \right)^2 + \sum_{j=iT/2+1}^{(i+1)T/2-1} (x_{2j} - x_{2j+1})^2 \right], \tag{7}
$$

$$
r(\mathbf{x}) = \sum_{i=1}^{Tt} \psi_{b_i}(b_i - x_i), \tag{8}
$$

where we suppose T is even and let $x_0 = 0$. Then we can verify that the function $g_{T,t}(\cdot)$ can be written as

$$
g_{T,t}(\mathbf{x}) = q_1(\mathbf{b} - \mathbf{x}) + q_2(\mathbf{b} - \mathbf{x}) + r(\mathbf{x}).
$$

We let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be the graph associated to the network of the agents in decentralized optimization, where the node set $\mathcal{V} = \{1, \ldots, n\}$ corresponds to the n agents and the edge set $\mathcal{E} = \{(i, j) : \text{node } i \text{ and node } j \text{ are connected}\}$ describes the topology of the agents network.

For given a subset $\mathcal{C} \subseteq \mathcal{V}$, we define the function $h_i^{\mathcal{C}}(\mathbf{x}) : \mathbb{R}^{Tt} \to \mathbb{R}$ as

$$
h_i^{\mathcal{C}}(\mathbf{x}) = \begin{cases} \frac{r(\mathbf{x})}{n} + \frac{q_1(\mathbf{b} - \mathbf{x})}{|\mathcal{C}|} & i \in \mathcal{C}, \\ \frac{r(\mathbf{x})}{n} + \frac{q_2(\mathbf{b} - \mathbf{x})}{|\mathcal{C}_\sigma|} & i \in \mathcal{C}_\sigma, \\ \frac{r(\mathbf{x})}{n}, & \text{otherwise}, \end{cases} \tag{9}
$$

where $\mathcal{C}_\sigma = \{v \in \mathcal{V} : \mathrm{dis}(\mathcal{C}, v) \geq \sigma\}$ and $\mathrm{dis}(\mathcal{C}, v)$ is the distance between set $\mathcal{C}$ and node v.

We introduce the following property (Yue et al., 2023, Lemma 4) of function $\psi_\theta : \mathbb{R} \to \mathbb{R}$ to analyze the smoothness of $h_i^{\mathcal{C}}(\cdot)$ and the communication complexity for hard instance.

Lemma B.1. *For any $\theta > 0$, the function $\psi_\theta : \mathbb{R} \to \mathbb{R}$ defined in equation (4) is 33-smooth and holds $\psi'_\theta(\theta) = 0$.*

Now we present the proofs for lower bounds in decentralized setting, which is based on our construction (6)–(8).

B.1. Proof of Lemma 4.1

Proof. We can observed that the function $q_1(\mathbf{x})$ is quadratic and holds $q_1(\mathbf{x}) \leq 1$ for any $\mathbf{x} \in \mathbb{R}^{Tt}$, then it is 2-smooth. Similarly, the function $q_2(\mathbf{x})$ is also 2-smooth.

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{Tt}$, we have

$$
\begin{aligned}
\|\nabla r(\mathbf{x}) - \nabla r(\mathbf{y})\|^2 &= \sum_{i=1}^{Tt} \left(v'_{b_i}(b_i - x_i) - v'_{b_i}(b_i - y_i) \right)^2 \\
&\leq \sum_{i=1}^{Tt} \left(33(x_i - y_i) \right)^2 \\
&= 33^2 \|\mathbf{x} - \mathbf{y}\|^2 ,
\end{aligned}
$$

where inequality is based on the Lemma B.1. This implies the function $r : \mathbb{R}^{Tt} \to \mathbb{R}$ is 33-smooth.

Combing above smoothness properties and the definition of $h_i^{\mathcal{C}}(\cdot)$, we conclude each $h_i^{\mathcal{C}}$ is $(33/n + \max\{2/|\mathcal{C}|, 2/|\mathcal{C}_\sigma|\})$-smooth and thus $\{h_i^{\mathcal{C}} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ is $(33/n + \max\{2/|\mathcal{C}|, 2/|\mathcal{C}_\sigma|\})$-mean-squared smooth.

The definition of $h(\cdot)$ and $h_i^{\mathcal{C}}(\cdot)$ implies

$$
h(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n h_i^{\mathcal{C}}(\mathbf{x}) = \frac{q_1(\mathbf{b} - \mathbf{x}) + q_2(\mathbf{b} - \mathbf{x}) + r(\mathbf{x})}{n} = \frac{g_{T,t}(\mathbf{x})}{n}. \tag{10}
$$

Then applying statements (c) and (d) of Lemma 3.1 and Lemma 3.3 finish the proof for the last two statements. □

B.2. Proof of Lemma 4.2

Proof. The definitions $f_i(\mathbf{x}) = \alpha h_i^{\mathcal{C}}(\beta \mathbf{x})$ and $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ and equation (10) implies $f(\mathbf{x}) = \alpha g_{T,t}(\beta \mathbf{x})/n$. Then the statement (e) of Lemma 3.1 implies when $\mathrm{supp}(\mathbf{x}) \subseteq \{1, 2, \cdots, Tt/2\}$, we have

$$
f(\mathbf{x}) - f^* = \frac{\alpha}{n}(g_{T,t}(\beta \mathbf{x}) - g_{T,t}^*) > \frac{3\alpha T\delta}{n}, \tag{11}
$$

since we have assumed $\delta < 0.01$ and $t = 2\lfloor \log_{8/7}(2/3\delta) \rfloor$.

We define

$$
\mathrm{nnz}(s, i) \triangleq \max\{m \in \mathbb{N} : \text{there exists } \mathbf{y} \in \mathcal{M}_i^s \text{ such that } y_m \neq 0\}.
$$

Then statements (a) of Lemma 3.1 implies achieving an $3\alpha T\delta/n$-suboptimal solution requires

$$
\max_{i \in [n]} \{\mathrm{nnz}(s, i)\} \geq \frac{Tt}{2} + 1. \tag{12}
$$

Now we consider how much local computation steps and local communication steps we need to attain the condition (12).

According to Lemma B.1 and equation (8), for any $\mathbf{x} \in \mathcal{M}_i^s$, we have

$$
(\nabla r(\mathbf{x}))_k = 0, \qquad \text{for any} \quad k > \mathrm{nnz}(s, i). \tag{13}
$$

According to equation (6), for any $\mathbf{x} \in \mathcal{M}_i^s$, we have

$$
(\nabla q_1(\mathbf{b} - \mathbf{x}))_k = 0, \qquad \text{for any} \quad k > \mathrm{nnz}(s, i) + \mathbb{I}\{\mathrm{nnz}(s, i) \equiv 1 \pmod 2\}. \tag{14}
$$

According to equation (7), for any $\mathbf{x} \in \mathcal{M}_i^s$, we have

$$
(\nabla q_2(\mathbf{b} - \mathbf{x}))_k = 0, \qquad \text{for any} \quad k > \mathrm{nnz}(s, i) + \mathbb{I}\{\mathrm{nnz}(s, i) \equiv 0 \pmod 2\}. \tag{15}
$$

Combining (9), (13), (14), (15) and Definition 2.10, we know that for any DFO algorithm:

1. If $i \in \mathcal{C}$ and $\mathrm{nnz}(s, i)$ is odd, one step of local computation can increase at most one dimension for memory of node i.

2. If $i \in \mathcal{C}_\sigma$ and $\mathrm{nnz}(s, i)$ is even, one step of local computation can increase at most one dimension for memory of node i.

3. Otherwise, one step of local computation cannot increase the dimension for memory of node i.

In summary, we have

$$\mathrm{nnz}(s+1, i) \leq \begin{cases} \mathrm{nnz}(s, i) + 1, & \text{if } i \in \mathcal{C}, \ \mathrm{nnz}(s, i) \equiv 1 \pmod{2}, \\ \mathrm{nnz}(s, i) + 1, & \text{if } i \in \mathcal{C}_\sigma, \ \mathrm{nnz}(s, i) \equiv 0 \pmod{2}, \\ \mathrm{nnz}(s, i), & \text{otherwise.} \end{cases} \tag{16}$$

We consider the cost to reach the second coordinate from the initial status that $\mathcal{M}_i^0 = \{\mathbf{0}\}$ for all $i \in [n]$. According to equation (16), we need to let a node in $\mathcal{C}$ reach the first coordinate, which requires at least one local computation step on some node in $\mathcal{C}_\sigma$ first. Then, according to definitions of DFO algorithm (Definition 2.10) and $\mathcal{C}_\sigma$, one must perform at least σ local communication steps for a node in $\mathcal{C}$ to receive the information of the first coordinate from some node in $\mathcal{C}_\sigma$. After above steps, we can perform at least 1 computation on nodes in $\mathcal{C}$ to reach the second coordinate. In summary, to reach the second coordinate requires at least 2 local computation step and σ local communication step.

Similarly, to reach the k-th coordinate, a DFO algorithm must perform at least k local computation steps and $(k-1)\sigma$ local communication steps. Thus, to attain the condition (12), one needs at least $Tt/2 + 1$ local computation steps and $Tt\sigma/2$ local communication steps, which corresponds to $Tt\sigma/2$ communications round and $Tt(1 + \sigma\tau)/2$ time cost. □

Remark B.2. Noticing that one computation step corresponds to one unit of time cost. However, some of agents (maybe not all agents) can parallel compute their local gradient, which means the computational time cost may be not proportion to the number of local gradient oracle calls.

B.3. Proof of Theorem 4.3

Proof. We consider the instance graph provided by Scaman et al. (2017), which associated to the specific spectral gap. Concretely, we let $\iota_m = (1 - \cos(\pi/m)) / (1 + \cos(\pi/m))$. For given $\gamma \in (0, 1]$, let $m = \lfloor \pi/\arccos((1 - \gamma)/(1 + \gamma)) \rfloor$, then we have $m \geq 2$ and $\iota_{m+1} < \gamma \leq \iota_m$. We study the cases of $m \geq 3$ and $m = 2$ separately.

We first consider the case of $m \geq 3$. Let the agent number $n = m$, take $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be the undirected linear graph of size n ordered from node 1 to node n such that $\mathcal{V} = \{1, \ldots, n\}$ and $\mathcal{E} = \{(i, j) : |i - j| = 1, i \in \mathcal{V} \text{ and } j \in \mathcal{V}\}$. We define a weighted matrix $\hat{\mathbf{W}}_l \in \mathbb{R}^{n \times n}$ for $\mathcal{G}$ such that $\hat{w}_{i+1,i} = \hat{w}_{i,i+1} = 1 - l\mathbb{I}(i = 1)$ and $\hat{w}_{ij} = 0$ for other entries. Let $\hat{\mathbf{R}}_l \in \mathbb{R}^{n \times n}$ be the Laplacian matrix of graph $\mathcal{G}$ associated to weighted matrix $\hat{\mathbf{W}}_l$. We define $\lambda_1(\hat{\mathbf{R}}_l), \ldots, \lambda_n(\hat{\mathbf{R}}_l)$ be eigenvalues of $\hat{\mathbf{R}}_l$ such that $0 = \lambda_n(\hat{\mathbf{R}}_l) \leq \lambda_{n-1}(\hat{\mathbf{R}}_l) \leq \cdots \leq \lambda_1(\hat{\mathbf{R}}_l)$. A simple calculation gives that

$$\lambda_{n-1}(\hat{\mathbf{R}}_0) = 2(1 - \cos(\pi/m)), \quad \lambda_1(\hat{\mathbf{R}}_0) = 2(1 + \cos(\pi/m)), \quad \lambda_{n-1}(\hat{\mathbf{R}}_1) = 0 \quad \text{and} \quad \lambda_1(\hat{\mathbf{R}}_l) > 0$$

for any $l \in [0, 1]$, then we have $\lambda_{n-1}(\hat{\mathbf{R}}_0)/\lambda_1(\hat{\mathbf{R}}_0) = \iota_m$ and $\lambda_{n-1}(\hat{\mathbf{R}}_1)/\lambda_1(\hat{\mathbf{R}}_1) = 0$. By the continuity of the eigenvalues of a matrix and the fact $0 < \gamma \leq \iota_m$, there exists some $l \in [0, 1)$ such that $\lambda_{n-1}(\hat{\mathbf{R}}_l)/\lambda_1(\hat{\mathbf{R}}_l) = \gamma$. Let $\mathbf{W} = 1 - \hat{\mathbf{R}}_l/\lambda_1(\hat{\mathbf{R}}_l)$, then the spectral gap satisfies $\gamma(\mathbf{W}) = \lambda_{n-1}(\hat{\mathbf{R}}_l)/\lambda_1(\hat{\mathbf{R}}_l) = \gamma$. According to basic properties of Laplacian matrix, the matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a mixing matrix satisfies Assumption 2.6.

We take $\{h_i^\mathcal{C} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ by following equation (9) with

$$\mathcal{C} = \left\{1, \ldots, \left\lceil \frac{n}{32} \right\rceil\right\}, \qquad \sigma = \left\lceil \frac{15n}{16} \right\rceil - 1, \qquad T = 2\left\lfloor \frac{\kappa}{194a} \right\rfloor \qquad \text{and} \qquad t = 2\left\lfloor \log_{\frac{8}{7}} \frac{2\Delta}{3\epsilon} \right\rfloor.$$

Let $h(\cdot) = \frac{1}{n} \sum_{i=1}^n h_i^\mathcal{C}(\cdot)$. Lemma 4.1 means the function set $\{h_i^\mathcal{C} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ is $(33/n + \max\{2/|\mathcal{C}|, 2/|\mathcal{C}_\sigma|\})$-mean-squared smooth, and the function h is $1/(anT)$-PL and satisfies

$$h(\mathbf{0}) - h^* \leq \frac{3T}{n}.$$

18

Let $f_i(\mathbf{x}) = \alpha h_i^{\mathcal{C}}(\beta\mathbf{x})$, $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})/n$ and apply Lemma 3.3 with $\alpha = n\Delta/(3T)$ and $\beta = \sqrt{3LT/(97\Delta)}$, then the function set $\{f_i : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ is $\alpha\beta^2(33/n + \max\{2/|\mathcal{C}|, 2/|\mathcal{C}_\sigma|\})$-mean-squared smooth, and the function $f : \mathbb{R}^{Tt} \to \mathbb{R}$ is $\alpha\beta^2/(anT)$-PL and satisfies

$$f(\mathbf{0}) - f^* \leq \frac{3\alpha T}{n}.$$

The values of α and β means

$$\alpha\beta^2\left(\frac{33}{n} + \frac{2}{|\mathcal{C}_\sigma|}\right) \leq \alpha\beta^2\left(\frac{33}{n} + \frac{2}{|\mathcal{C}|}\right) \leq \frac{nL}{97}\left(\frac{33}{n} + \frac{64}{n}\right) = L,$$

$$\frac{\alpha\beta^2}{anT} = \frac{L}{97aT} \geq \mu \qquad \text{and} \qquad \frac{3\alpha T}{n} = \Delta.$$

According to Lemma 3.3 and Lemma 4.1, we conclude $\{f_i : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ is L-mean-squared smooth and $f : \mathbb{R}^d \to \mathbb{R}$ is μ-PL with $f(\mathbf{0}) - f^* \leq \Delta$.

Applying Lemma 4.2 with $\delta = \epsilon/\Delta < 0.01$, any DFO algorithm needs at least $Tt\sigma/2$ communication steps and $Tt(1+\sigma\tau)/2$ time cost to achieve an ϵ-suboptimal solution since $3\alpha T\delta/n = \epsilon$. The setting $n = m$ implies

$$\frac{2}{(n+1)^2} \leq \iota_{n+1} < \gamma \leq \iota_3 = \frac{1}{3},$$

which means

$$\sigma = \left\lceil\frac{15n}{16}\right\rceil - 1 \geq \frac{15}{16}\left(\sqrt{\frac{2}{\gamma}} - 1\right) - 1 \geq \frac{1}{5\sqrt{\gamma}}.$$

Hence, we achieve the lower bounds for communication complexity and time complexity of $Tt\sigma/2 = \Omega\big(\kappa/\sqrt{\gamma}\log(1/\epsilon)\big)$ and $Tt(1+\sigma\tau)/2 = \Omega\big(\kappa(1 + \tau/\sqrt{\gamma})\log(1/\epsilon)\big)$ respectively.

We then consider the case of $m = 2$. We let the agent number be $n = 3$ and take $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be the totally connected graph of size n such that $\mathcal{V} = \{1, \ldots, n\}$ and $\mathcal{E} = \{(i, j) : i \in \mathcal{V} \text{ and } j \in \mathcal{V}\}$. We define a weighted matrix $\hat{\mathbf{W}}_l \in \mathbb{R}^{n \times n}$ for $\mathcal{G}$ such that $\hat{w}_{1,3} = \hat{w}_{3,1} = l$ and $w_{ij} = 1$ for other entries. Let $\hat{\mathbf{R}}_l \in \mathbb{R}^{n \times n}$ be the Laplacian matrix of graph $\mathcal{G}$ associated to weighted matrix $\hat{\mathbf{W}}_l$. We define $\lambda_1(\hat{\mathbf{R}}_l), \ldots, \lambda_n(\hat{\mathbf{R}}_l)$ be eigenvalues of $\hat{\mathbf{R}}_l$ such that $0 = \lambda_n(\hat{\mathbf{R}}_l) \leq \lambda_{n-1}(\hat{\mathbf{R}}_l) \leq \cdots \leq \lambda_1(\hat{\mathbf{R}}_l)$. A simple calculation gives that

$$\lambda_{n-1}(\hat{\mathbf{R}}_1) = \lambda_1(\hat{\mathbf{R}}_1) = 3, \quad \lambda_{n-1}(\hat{\mathbf{R}}_0) = 2(1 - \cos(\pi/n)), \quad \lambda_1(\hat{\mathbf{R}}_0) = 2(1 + \cos(\pi/n)) \quad \text{and} \quad \lambda_1(\hat{\mathbf{R}}_l) > 0$$

for any $l \in [0, 1]$. Then we have $\lambda_{n-1}(\hat{\mathbf{R}}_0)/\lambda_1(\hat{\mathbf{R}}_0) = \iota_n = \iota_3$ and $\lambda_{n-1}(\hat{\mathbf{R}}_1)/\lambda_1(\hat{\mathbf{R}}_1) = 1$. By continuity of the eigenvalues of a matrix and the fact $\iota_3 < \gamma \leq 1$, there exists some $l \in (0, 1]$ such that $\lambda_{n-1}(\hat{\mathbf{R}}_l)/\lambda_1(\hat{\mathbf{R}}_l) = \gamma$. Let $\mathbf{W} = 1 - \hat{\mathbf{R}}_l/\lambda_1(\hat{\mathbf{R}}_l)$, then the spectral gap satisfies $\gamma(\mathbf{W}) = \lambda_{n-1}(\hat{\mathbf{R}}_l)/\lambda_1(\hat{\mathbf{R}}_l) = \gamma$. According to basic properties of Laplacian matrix, the matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a mixing matrix satisfies Assumption 2.6.

We take $\{h_i^{\mathcal{C}} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ by following equation (9) with

$$\mathcal{C} = \{1\}, \qquad \sigma = 1, \qquad T = 2\left\lfloor\frac{\kappa}{78a}\right\rfloor \qquad \text{and} \qquad t = 2\left\lfloor\log_{\frac{8}{7}}\frac{2\Delta}{3\epsilon}\right\rfloor.$$

Let $h(\cdot) = \frac{1}{n}\sum_{i=1}^n h_i^{\mathcal{C}}(\cdot)$. Lemma 4.1 means the function set $\{h_i^{\mathcal{C}} : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ is $(33/n + \max\{2/|\mathcal{C}|, 2/|\mathcal{C}_\sigma|\})$-mean-squared smooth, h is $1/(anT)$-PL and satisfies

$$h(\mathbf{0}) - h^* \leq 3T/n.$$

Let $f_i(\mathbf{x}) = \alpha h_i^{\mathcal{C}}(\beta\mathbf{x})$, $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})/n$ and apply Lemma 3.3 with $\alpha = n\Delta/(3T)$ and $\beta = \sqrt{LT/(13\Delta)}$, then we conclude $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ is $\alpha\beta^2(33/n + \max\{2/|\mathcal{C}|, 2/|\mathcal{C}_\sigma|\})$-mean-squared smooth and f is $\alpha\beta^2/(anT)$-PL and satisfies

$$f(\mathbf{0}) - f^* \leq \frac{3\alpha T}{n}.$$

The values of α and β means

$$\alpha\beta^2\left(\frac{33}{n}+\frac{2}{|\mathcal{C}_\sigma|}\right) < \alpha\beta^2\left(\frac{33}{n}+\frac{2}{|\mathcal{C}|}\right) = \frac{nL}{39}\left(\frac{33}{n}+\frac{6}{n}\right) = L, \qquad \frac{\alpha\beta^2}{anT} = \frac{L}{39aT} \geq \mu \qquad \text{and} \qquad \frac{3\alpha T}{n} = \Delta.$$

According to Lemma 3.3 and Lemma 4.1, we conclude $\{f_i : \mathbb{R}^{Tt} \to \mathbb{R}\}_{i=1}^n$ is L-mean-squared smooth, and $f : \mathbb{R}^d \to \mathbb{R}$ is μ-PL with $f(\mathbf{0}) - f^* \leq \Delta$.

Applying Lemma 4.2 with $\delta = \epsilon/\Delta < 0.01$, any DFO algorithm needs at least $Tt\sigma/2$ communications and $Tt(1+\sigma\tau)/2$ time to reach an ϵ-suboptimal solution since $3\alpha T\delta/n = \epsilon$. The setting $n = 3$ means $\gamma > \iota_3 = 1/3$ and $\sigma = 1 > 1/\sqrt{3\gamma}$, which results the lower bounds for communication complexity and time complexity of $Tt\sigma/2 = \Omega\left(\kappa/\sqrt{\gamma}\log(1/\epsilon)\right)$ and $Tt(1+\sigma\tau)/2 = \Omega\left(\kappa(1+\tau/\sqrt{\gamma})\log(1/\epsilon)\right)$ respectively.

$\square$

C. The Proofs in Section 5

Recall that we define the Lyapunov function

$$\Phi^t = \mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + \alpha U^t + \beta V^t + LC^t,$$

where $\alpha = 2\eta/p$, $\beta = 8L\rho^2 n\eta^2$,

$$U^t = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\left(\mathbf{g}^t(i) - \nabla f_i(\mathbf{x}^t(i))\right)\right\|^2, \quad V^t = \frac{1}{n}\mathbb{E}\|\mathbf{G}^t - \nabla\mathbf{F}(\mathbf{X}^t)\|^2 \text{ and } C^t = \mathbb{E}\|\mathbf{X}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \eta^2\mathbb{E}\|\mathbf{S}^t - \mathbf{1}\bar{\mathbf{s}}^t\|^2.$$

Compared with the analysis of Luo & Ye (2022); Li et al. (2022a) for general nonconvex case, we introduce the term of f^* into the Lyapunov function to show the linear convergence under the PL condition. Different with previous work (Luo & Ye, 2022; Li et al., 2022a) suppose each $f_i(\cdot)$ is L-smooth, our analysis only require $\{f_i(\cdot)\}_{i=1}^n$ is L-smooth that allows each $f_i(\cdot)$ to be $\sqrt{n}L$-smooth (see Lemma C.3).

The remainder of this section first provide some technical Lemmas, then give the detailed proofs for the results in Section 4.

C.1. Some Technical Lemmas

We introduce some lemmas for our later analysis.

Lemma C.1 (Ye et al. (2023, Proposition 1)). *Under Assumption 2.6, Algorithm 2 holds*

$$\frac{1}{n}\mathbf{1}^\top\mathbf{Y}^K = \bar{\mathbf{y}}^0 \qquad \text{and} \qquad \|\mathbf{Y}^K - \mathbf{1}\bar{\mathbf{y}}^0\| \leq \sqrt{14}\left(1 - \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{1 - \lambda_2(\mathbf{W})}\right)^K\|\mathbf{Y}^0 - \mathbf{1}\bar{\mathbf{y}}^0\|,$$

where $\bar{\mathbf{y}}^0 = \frac{1}{n}\mathbf{1}^\top\mathbf{Y}^0 = \frac{1}{n}\sum_{i=1}^n\mathbf{Y}^0(i)$.

Lemma C.2 ((Ye et al., 2023, Lemma 3)). *For any $\mathbf{X} \in \mathbb{R}^{n\times d}$, we have $\|\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}\| \leq \|\mathbf{X}\|$ where $\bar{\mathbf{X}} = \frac{1}{n}\mathbf{1}^\top\mathbf{X}$.*

Lemma C.3. *Under Assumption 2.2, the function $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ is L-smooth and each $f_i(\cdot)$ is $\sqrt{n}L$-smooth.*

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the mean-square smoothness of $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ implies

$$\begin{aligned}
\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 &= \left\|\frac{1}{n}\sum_{i=1}^n(\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}))\right\|_2^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2^2 \\
&\leq L\|\mathbf{x} - \mathbf{y}\|_2^2,
\end{aligned}$$

which means $f(\cdot)$ is L-smooth.

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i \in [n]$, the mean-square smoothness of $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ implies

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq \sqrt{\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2} \leq \sqrt{nL^2 \|\mathbf{x} - \mathbf{y}\|^2} \leq \sqrt{n}L \|\mathbf{x} - \mathbf{y}\|,$$

which means $f_i(\cdot)$ is $\sqrt{n}L$-smooth. $\qquad\square$

Lemma C.4 (Li et al. (2021, Lemma 2)). *Suppose the function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and the vectors $\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{v}_t \in \mathbb{R}^d$ satisfy $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$ for $\eta > 0$. Then we have*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{\eta}{2} \|\nabla f(\mathbf{x}_t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{\eta}{2} \|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2. \tag{17}$$

We establish the decrease of function value as follows.

Lemma C.5. *Under Assumption 2.1-2.6, Algorithm 4 holds that*

$$\mathbb{E}[f(\bar{\mathbf{x}}^{t+1}) - f^*] \leq (1 - \mu\eta)\mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + \eta U^t + L^2 \eta C^t - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2.$$

Proof. Lemma C.1 and the update $\mathbf{X}^{t+1} = \text{AccGossip}(\mathbf{X}^t - \eta \mathbf{S}^t, \mathbf{W}, K)$ means

$$\begin{aligned}
\bar{\mathbf{x}}^{t+1} &= \frac{1}{n} \mathbf{1}^\top \text{AccGossip}(\mathbf{X}^t - \eta \mathbf{S}^t, \mathbf{W}, K) \\
&= \frac{1}{n} \mathbf{1}^\top (\mathbf{X}^t - \eta \mathbf{S}^t) \\
&= \bar{\mathbf{x}}^t - \eta \bar{\mathbf{s}}^t.
\end{aligned} \tag{18}$$

Lemma C.3 shows the function $f(\cdot)$ is L-smooth, then Lemma C.4 with $\mathbf{x}^t = \bar{\mathbf{x}}^t$, $\mathbf{x}^{t+1} = \bar{\mathbf{x}}^{t+1}$ and $\mathbf{v}^t = \bar{\mathbf{s}}^t$ means

$$f(\bar{\mathbf{x}}^{t+1}) \leq f(\bar{\mathbf{x}}^t) - \frac{\eta}{2} \|\nabla f(\bar{\mathbf{x}}^t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2 + \frac{\eta}{2} \|\bar{\mathbf{s}}^t - \nabla f(\bar{\mathbf{x}}^t)\|^2. \tag{19}$$

We also have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}^t(i)) - \nabla f_i(\bar{\mathbf{x}}^t)\right)\right\|^2 &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\|\nabla f_i(\mathbf{x}^t(i)) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^m nL^2 \mathbb{E}\left\|\mathbf{x}^t(i) - \bar{\mathbf{x}}^t\right\|^2 \\
&= L^2 \mathbb{E}\left\|\mathbf{x}^t - \mathbf{1}\bar{\mathbf{x}}^t\right\|^2 \leq L^2 C^t,
\end{aligned}$$

where we use the inequality $\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{a}_i\right\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|\mathbf{a}_i\|^2$ for $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^d$ and Lemma C.3. Consequently, we have

$$\begin{aligned}
\mathbb{E}\left\|\bar{\mathbf{s}}^t - \nabla f(\bar{\mathbf{x}}^t)\right\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \left(\mathbf{g}^t(i) - \nabla f_i(\bar{\mathbf{x}}^t)\right)\right\|^2 \\
&\leq 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \left(\mathbf{g}^t(i) - \nabla f_i(\mathbf{x}^t(i))\right)\right\|^2 + 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}^t(i)) - \nabla f_i(\bar{\mathbf{x}}^t)\right)\right\|^2 \\
&\leq 2U^t + 2L^2 C^t.
\end{aligned} \tag{20}$$

Combining the results of (19) and (20), we have

$$\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}^{t+1}) - f^*] &\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^t) - f^* - \frac{\eta}{2}\|\nabla f(\bar{\mathbf{x}}^t)\|^2 + \eta U^t + L^2\eta C^t - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2\right] \\
&\leq (1 - \mu\eta)\mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + \eta U^t + L^2\eta C^t - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2,
\end{aligned}$$

where the last step is based on the PL condition in Assumption 2.3. $\qquad\square$

Then we provide the recursion for U_t, V_t and C_t in the following lemma.

Lemma C.6. *Under the setting of Theorem 5.2, we have*

$$C^{t+1} \leq 20\rho^2 n^2 C^t + 4\rho^2 np\eta^2 V^t + 12\rho^2 n^2 \mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2,$$

$$U^{t+1} \leq (1-p)U^t + 4pL^2 C^t + 3pL^2 \mathbb{E}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2,$$

and

$$V^{t+1} \leq (1-p)V^t + 4pnL^2 C^t + 3pnL^2 \mathbb{E}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2,$$

where $\rho = \sqrt{14}\left(1 - \left(1 - 1/\sqrt{2}\right)\sqrt{1 - \lambda_2}\right)^K$.

Proof. The setting Lemma C.1 implies

$$\rho^2 \leq \frac{1}{80n^2} \leq \frac{1}{80}. \tag{21}$$

We first consider $C^{t+1} = \mathbb{E}\|\mathbf{X}^{t+1} - \mathbf{1}\bar{\mathbf{x}}^{t+1}\|^2 + \eta^2 \mathbb{E}\|\mathbf{S}^{t+1} - \mathbf{1}\bar{\mathbf{s}}^{t+1}\|^2$. For the term $\mathbb{E}\|\mathbf{X}^{t+1} - \mathbf{1}\bar{\mathbf{x}}^{t+1}\|^2$, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{X}^{t+1} - \mathbf{1}\bar{\mathbf{x}}^{t+1}\|^2 &\leq \rho^2 \mathbb{E}\|(\mathbf{X}^t - \eta\mathbf{S}^t) - \mathbf{1}(\bar{\mathbf{x}}^t - \eta\bar{\mathbf{s}}^t)\|^2 \\
&\leq 2\rho^2\left(\mathbb{E}\|\mathbf{X}^t - \mathbf{1}\bar{\mathbf{x}}^t\|^2 + \eta^2 \mathbb{E}\|\mathbf{S}^t - \mathbf{1}\bar{\mathbf{s}}^t\|^2\right) \\
&= 2\rho^2 C^t,
\end{aligned} \tag{22}$$

where we use the definition of ρ and Lemma C.1 in the first inequality and Young's inequality in the second inequality.

For the term $\eta^2\|\mathbf{S}^t - \mathbf{1}\bar{\mathbf{s}}^t\|^2$, we have

$$\begin{aligned}
\eta^2 \mathbb{E}\|\mathbf{S}^{t+1} - \mathbf{1}\bar{\mathbf{s}}^{t+1}\|^2 &\leq \rho^2\eta^2 \mathbb{E}\|\mathbf{S}^t + \mathbf{G}^{t+1} - \mathbf{G}^t - \mathbf{1}(\bar{\mathbf{s}}^t + \bar{\mathbf{g}}^{t+1} - \bar{\mathbf{g}}^t)\|^2 \\
&\leq 2\rho^2\eta^2 \mathbb{E}\|\mathbf{S}^t - \mathbf{1}\bar{\mathbf{s}}^t\|^2 + 2\rho^2\eta^2 \mathbb{E}\|\mathbf{G}^{t+1} - \mathbf{G}^t - \mathbf{1}(\bar{\mathbf{g}}^{t+1} - \bar{\mathbf{g}}^t)\|^2 \\
&\leq 2\rho^2\eta^2 \mathbb{E}\|\mathbf{S}^t - \mathbf{1}\bar{\mathbf{s}}^t\|^2 + 2\rho^2\eta^2 \mathbb{E}\|\mathbf{G}^{t+1} - \mathbf{G}^t\|^2 \\
&\leq 2\rho^2 C^t + 2\rho^2\eta^2 \mathbb{E}\|\mathbf{G}^{t+1} - \mathbf{G}^t\|^2,
\end{aligned} \tag{23}$$

where we use the definition of ρ and Lemma C.1 in the first inequality, Young's inequality in the second inequality, Lemma C.2 in the third inequality and definition of C^t in the last inequality.

We bound the term $\|\mathbf{G}^{t+1} - \mathbf{G}^t\|^2$ in inequality (23) as

$$\begin{aligned}
&\mathbb{E}\|\mathbf{G}^{t+1} - \mathbf{G}^t\|^2 \\
=&p\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^{t+1}) - \mathbf{G}^t\|^2 + (1-p)\sum_{i=1}^n \mathbb{E}\left\|\frac{\xi_i^t}{bq}\left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right)\right\|^2 \\
\leq&2p\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^{t+1}) - \nabla\mathbf{F}(\mathbf{X}^t)\|^2 + 2p\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^t) - \mathbf{G}^t\|^2 + (1-p)\sum_{i=1}^n \mathbb{E}\left\|\frac{\xi_i^t}{bq}\left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right)\right\|^2 \\
\leq&2p\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^{t+1}) - \nabla\mathbf{F}(\mathbf{X}^t)\|^2 + 2p\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^t) - \mathbf{G}^t\|^2 + \frac{(1-p)(b+n)}{b}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\|^2 \\
=&2pnV^t + \left(2p + \frac{(1-p)(b+n)}{b}\right)\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^{t+1}) - \nabla\mathbf{F}(\mathbf{X}^t)\|^2 \\
\leq&2pnV^t + \left(2p + p(b+n)\right)\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^{t+1}) - \nabla\mathbf{F}(\mathbf{X}^t)\|^2 \\
\leq&2pnV^t + 4pn\mathbb{E}\|\nabla\mathbf{F}(\mathbf{X}^{t+1}) - \nabla\mathbf{F}(\mathbf{X}^t)\|^2
\end{aligned} \tag{24}$$

where the first inequality is based on Cauchy–Schwarz inequality, the second inequality is based on

$$\mathbb{E}\left\|\frac{\xi_i^t}{bq}\left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right)\right\|^2 \leq \frac{\mathbb{E}[(\xi_i^t)^2]}{b^2 q^2}\mathbb{E}\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\|^2$$

and the fact $\xi_i^t \sim \text{Binomial}(b, q)$ that leads to

$$\frac{\mathbb{E}[(\xi_i^t)^2]}{b^2 q^2} = \frac{b^2 q^2 + bq(1-q)}{b^2 q^2} = \frac{b+n-1}{b} \leq \frac{b+n}{b},$$

the third inequality is based on the fact $b \geq (1-p)/p$ and the last step is based on the fact that $1 \leq b \leq n$.

We bound the term $\mathbb{E}\|\nabla \mathbf{F}(\mathbf{X}^{t+1}) - \nabla \mathbf{F}(\mathbf{X}^t)\|^2$ in inequality (24) as

$$
\begin{aligned}
&\mathbb{E}\|\nabla \mathbf{F}(\mathbf{X}^{t+1}) - \nabla \mathbf{F}(\mathbf{X}^t)\|^2 \\
=& \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i)) \right\|^2 \\
\leq& 3 \sum_{i=1}^{n} \mathbb{E} \big(\left\| \nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\bar{\mathbf{x}}^{t+1}) \right\|^2 + \left\| \nabla f_i(\bar{\mathbf{x}}^{t+1}) - \nabla f_i(\bar{\mathbf{x}}^t) \right\|^2 + \left\| \nabla f_i(\bar{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}^t(i)) \right\|^2 \big) \\
\leq& 3(\sqrt{n}L)^2 \sum_{i=1}^{n} \mathbb{E} \left\| \mathbf{x}^{t+1}(i) - \bar{\mathbf{x}}^{t+1} \right\|^2 + 3nL^2 \mathbb{E} \left\| \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t \right\|^2 + 3(\sqrt{n}L)^2 \sum_{i=1}^{n} \mathbb{E} \left\| \mathbf{x}^t(i) - \bar{\mathbf{x}}^t \right\|^2 \\
=& 3nL^2 \big(\mathbb{E} \left\| \mathbf{X}^{t+1} - \mathbf{1}\bar{\mathbf{x}}^{t+1} \right\|^2 + \mathbb{E} \left\| \mathbf{X}^t - \mathbf{1}\bar{\mathbf{x}}^t \right\|^2 + \mathbb{E} \left\| \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t \right\|^2 \big) \\
\leq& 3nL^2 \big(2\rho^2 C^t + C^t + \mathbb{E} \left\| \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t \right\|^2 \big) \\
\leq& 4nL^2 C^t + 3nL^2 \mathbb{E} \left\| \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t \right\|^2,
\end{aligned}
\tag{25}
$$

where the the second inequality is based on Lemma C.3, the third inequality is based on inequality (22) and last step is based on the the fact $\rho^2 \leq 1/6$.

Combining inequalities (22) – (25), we achieve

$$C^{t+1} \leq 20\rho^2 n^2 C^t + 4\rho^2 np\eta^2 V^t + 12\rho^2 n^2 \mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2.$$

We then consider $U^{t+1} = \mathbb{E}\big\| \frac{1}{n} \sum_{i=1}^{n} \big(\mathbf{g}^{t+1}(i) - \nabla f_i(\mathbf{x}^{t+1}(i)) \big) \big\|^2$. We let ξ be the random variable satisfying $P(\xi = i) = q$. The update rule for $\mathbf{g}_{t+1}(i)$ implies

$$
\begin{aligned}
U^{t+1} =& (1-p)\mathbb{E}\Big\| \frac{1}{n} \sum_{i=1}^{n} \Big(\mathbf{g}^t(i) + \frac{\xi_i^t}{bq} \big(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i)) \big) - \nabla f_i(\mathbf{x}^{t+1}(i)) \Big) \Big\|^2 \\
=& (1-p)U^t + (1-p)\mathbb{E}\Big\| \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{\xi_i^t}{bq} \big(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i)) \big) - \big(\nabla f_i(\mathbf{x}^{t+1}(i)) \big) - \nabla f_i(\mathbf{x}^t(i)) \big) \Big) \Big\|^2 \\
=& (1-p)U^t + (1-p)\mathbb{E}\Big\| \sum_{i=1}^{n} \Big(\frac{\xi_i^t}{b} \big(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i)) \big) - \frac{1}{n} \big(\nabla f_i(\mathbf{x}^{t+1}(i)) \big) - \nabla f_i(\mathbf{x}^t(i)) \big) \Big) \Big\|^2 \\
=& (1-p)U^t + \frac{1-p}{b}\mathbb{E}\Big\| \big(\nabla f_\xi(\mathbf{x}^{t+1}(\xi)) - \nabla f_\xi(\mathbf{x}^t(\xi)) \big) - \frac{1}{n} \sum_{i=1}^{n} \big(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i)) \big) \Big\|^2 \\
\leq& (1-p)U^t + \frac{1-p}{b}\mathbb{E}\Big\| \big(\nabla f_\xi(\mathbf{x}^{t+1}(\xi)) - \nabla f_\xi(\mathbf{x}^t(\xi)) \big) \Big\|^2 \\
=& (1-p)U^t + \frac{1-p}{nb} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i)) \right\|^2 \\
\leq& (1-p)U^t + \frac{p}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i)) \right\|^2 \\
\leq& (1-p)U^t + 4pL^2 C^t + 3pL^2 \mathbb{E} \left\| \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t \right\|^2,
\end{aligned}
$$

where the second equality is based on the property of Martingale (Fang et al., 2018, Proposition 1), the third equality is

based on the choice of q, the fourth equality is based on the fact $[\xi_1^t, \cdots, \xi_n^t]^\top \sim \text{Multinomial}(b, \frac{1}{n}\mathbf{1})$ and

$$\mathbb{E}\left[\nabla f_\xi(\mathbf{x}^{t+1}(\xi)) - \nabla f_\xi(\mathbf{x}^t(\xi))\right] = \frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right), \tag{26}$$

the first inequality is also based on (26), the second inequality is based on the setting of $b \geq (1-p)/p$ and the last step is because of the following inequality.

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right\|^2$$

$$\leq \frac{3}{n}\sum_{i=1}^n \left(\mathbb{E}\left\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\bar{\mathbf{x}}^{t+1})\right\|^2 + \mathbb{E}\left\|\nabla f_i(\bar{\mathbf{x}}^{t+1}) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + \mathbb{E}\left\|\nabla f_i(\bar{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}^t(i))\right\|^2\right)$$

$$\leq \frac{3}{n}\sum_{i=1}^n \left(nL^2\mathbb{E}\left\|\mathbf{x}^{t+1}(i) - \bar{\mathbf{x}}^{t+1}\right\|^2 + \mathbb{E}\left\|\nabla f_i(\bar{\mathbf{x}}^{t+1}) - \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + nL^2\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}^t(i)\right\|^2\right) \tag{27}$$

$$\leq 3L^2\mathbb{E}\left\|\mathbf{X}^{t+1} - \mathbf{1}\bar{\mathbf{x}}^{t+1}\right\|^2 + 3L^2\mathbb{E}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2 + 3L^2\mathbb{E}\left\|\mathbf{X}^t - \mathbf{1}\bar{\mathbf{x}}^t\right\|^2$$

$$\leq 4L^2C^t + 3L^2\mathbb{E}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2,$$

where the second inequality is based on Lemma C.3, the third inequality is based on Assumption 2.2 and the last step is based on inequalities (21) and (22).

We finally consider $V^{t+1} = \frac{1}{n}\mathbb{E}\|\mathbf{G}^{t+1} - \nabla\mathbf{F}(\mathbf{X}^{t+1})\|^2$. The update rule for $\mathbf{g}_{t+1}(i)$ implies

$$V^{t+1} = \frac{1-p}{n}\sum_{i=1}^n \mathbb{E}\left\|\mathbf{g}^t(i) + \frac{\xi_i^t}{bq}\left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right) - \nabla f_i(\mathbf{x}^{t+1}(i))\right\|^2$$

$$= (1-p)V^t + \frac{1-p}{n}\sum_{i=1}^n \mathbb{E}\left\|\frac{\xi_i^t}{bq}\left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right) - \left(\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\right)\right\|^2$$

$$\leq (1-p)V^t + \frac{1-p}{n}\sum_{i=1}^n \frac{\text{Var}[\xi_i^t]}{b^2q^2}\mathbb{E}\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\|^2$$

$$= (1-p)V^t + \frac{1-p}{n}\sum_{i=1}^n \frac{bq(1-q)}{b^2q^2}\mathbb{E}\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\|^2$$

$$\leq (1-p)V^t + \frac{1-p}{b}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\|^2$$

$$\leq (1-p)V^t + p\sum_{i=1}^n \mathbb{E}\|\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))\|^2$$

$$\leq (1-p)V^t + 4pnL^2C^t + 3pnL^2\mathbb{E}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2,$$

where the second equality is based on the property of Martingale (Fang et al., 2018, Proposition 1), the first inequality is based on the fact $\xi_i^t \sim \text{Binomial}(b, q)$ and the independence between ξ_i^t and $\nabla f_i(\mathbf{x}^{t+1}(i)) - \nabla f_i(\mathbf{x}^t(i))$, the second and third and inequalities are based on the settings $q = 1/n$ and $b \geq (1-p)/p$, and the last step is based on the inequality (27). $\square$

C.2. The Proof of Theorem 5.2

Proof. Recall that Lemma C.5 says

$$\mathbb{E}[f(\bar{\mathbf{x}}^{t+1}) - f^*] \leq (1-\mu\eta)\mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + \eta U^t + L^2\eta C^t - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2.$$

Combining above inequality with Lemma C.6, we have

$$
\begin{aligned}
\Phi^{t+1} =& \mathbb{E}[f(\bar{\mathbf{x}}^{t+1}) - f^*] + \alpha U^{t+1} + \beta V^{t+1} + LC^{t+1} \\
\leq& (1 - \mu\eta)\mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + \eta U^t + L^2\eta C^t - (\frac{1}{2\eta} - \frac{L}{2})\mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2 \\
& + \alpha(1-p)U^t + 4\alpha pL^2 C^t + 3\alpha pL^2\mathbb{E}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2 \\
& + \beta(1-p)V^t + 4\beta pnL^2 C^t + 3\beta pnL^2\mathbb{E}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2 \\
& + 20L\rho^2 n^2 C^t + 4L\rho^2 np\eta^2 V^t + 12L\rho^2 n^2 \mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2 \\
\leq& (1-\mu\eta)\mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + \big(\eta + \alpha(1-p)\big)U^t + \big(\beta(1-p) + 4L\rho^2 np\eta^2\big)V^t \\
& + (L^2\eta + 4\alpha pL^2 + 4\beta pnL^2 + 20L\rho^2 n^2)C^t \\
& - \left(\frac{1}{2\eta} - \frac{L}{2} - 3\alpha pL^2 - 3\beta pnL^2 - 12L\rho^2 n^2\right)\mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2.
\end{aligned}
$$

The setting of p, η, b, ρ, α and β implies

$$
\begin{aligned}
\eta + \alpha(1-p) &\leq \alpha(1 - \mu\eta), \\
\beta(1-p) + 4L\rho^2 np\eta^2 &\leq \beta(1 - \mu\eta), \\
L^2\eta + 4\alpha pL^2 + 4\beta pnL^2 + 20L\rho^2 n^2 &\leq L(1 - \mu\eta),
\end{aligned}
$$

and

$$
\frac{1}{2\eta} - \frac{L}{2} - 3\alpha pL^2 - 3\beta pnL^2 - 12L\rho^2 n^2 \geq 0.
$$

Therefore, we have

$$
\Phi^{t+1} \leq (1 - \mu\eta)\Phi^t,
$$

and the setting

$$
T \geq \left\lceil \frac{1}{\mu\eta} \log \frac{\Phi^0}{\epsilon} \right\rceil
$$

leads to

$$
\Phi^T \leq (1 - \mu\eta)^T \Phi^0 \leq \epsilon. \tag{28}
$$

The output $\mathbf{x}^{\text{out}}$ holds that

$$
\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{\text{out}}) - f^*] &= \mathbb{E}[f(\bar{\mathbf{x}}^T) - f^*] + \mathbb{E}[f(\mathbf{x}^{\text{out}}) - f(\bar{\mathbf{x}}^T)] \\
&= \mathbb{E}[f(\bar{\mathbf{x}}^T) - f^*] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(\mathbf{x}^T(i)) - f(\bar{\mathbf{x}}^T)] \\
&\leq \mathbb{E}[f(\bar{\mathbf{x}}^T) - f^*] + \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}\left[\langle \nabla f(\bar{\mathbf{x}}^T), \mathbf{x}^T(i) - \bar{\mathbf{x}}^T\rangle + \frac{L}{2}\mathbb{E}\left\|\mathbf{x}^T(i) - \bar{\mathbf{x}}^T\right\|^2\right]\right) \\
&= \mathbb{E}[f(\bar{\mathbf{x}}^T) - f^*] + \frac{L}{2n}\sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}^T(i) - \bar{\mathbf{x}}^T\right\|^2 \\
&\leq \mathbb{E}[f(\bar{\mathbf{x}}^T) - f^*] + \frac{L}{2n}C^T \\
&\leq \Phi^T \leq \epsilon,
\end{aligned}
$$

where the first inequality is based on Lemma C.3, the second inequality is based on the definition of C^t, the third inequality is based on the definition of Φ^t and the last step is based on inequality (28) $\qquad\square$

Algorithm 5 Centralized GD

1: **Input:** initial point $\tilde{\mathbf{x}}^0 \in \mathbb{R}^d$, iteration number T and stepsize $\eta > 0$

2: **for** $t = 0, 1 \cdots T - 1$ **do**

3:    **for** $i = 1, \ldots, n$ **do in parallel**

4:       i-th client:

5:          receive $\tilde{\mathbf{x}}^t$

6:          compute $\mathbf{g}^t(i) = \nabla f_i(\mathbf{x}^t)$

7:          send $\mathbf{g}^t(i)$

8:    **end for**

9:    server:

10:       receive $\mathbf{g}^t(1), \ldots, \mathbf{g}^t(n)$

11:       compute $\tilde{\mathbf{x}}^{t+1} = \tilde{\mathbf{x}}^t - \eta \cdot \frac{1}{n} \sum_{i=1}^n \mathbf{g}^t(i)$

12:       broadcast $\tilde{\mathbf{x}}^{t+1}$

13: **end for**

14: **Output:** $\tilde{\mathbf{x}}^T$

C.3. The Proof of Corollary 5.3

Proof. The parameters setting in this corollary means

$$p = \Theta\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{1}{\kappa}\right\}\right), \qquad b = \Theta\left(\min\{\sqrt{n}, \kappa\}\right), \qquad \eta = \Theta\left(\frac{1}{L}\right),$$

$$K = \Theta\left(\frac{\sqrt{2}\,(4 + \log n)}{(\sqrt{2} - 1)\sqrt{\gamma}}\right) \qquad \text{and} \qquad T = \Theta\left(\kappa \log\left(\frac{1}{\epsilon}\right)\right).$$

At each iteration, the algorithm takes $pn + (1 - p)b = \mathcal{O}\left(n/\min\{\kappa, \sqrt{n}\}\right)$ LFO calls (in expectation), K communication rounds and $1 + K\tau$ time cost. Multiplying the overall iteration numbers T on these quantities completes the proof. $\qquad\square$

Remark C.7. We omit the detailed proof of Theorem 5.1, since it can be easily achieved by following Theorem 5.2 and Corollary 5.3 with $p = 1$, $U^t = V^t = 0$ and showing the linear convergence of $\mathbb{E}[f(\bar{\mathbf{x}}^t) - f^*] + LC^t$.

D. More Details for Experiments

All of our experiments are performed on PC with Intel(R) Core(TM) i7-8550U CPU@1.80GHz processor and we implement the algorithms by MPI for Python 3.9.

We formally present the details of centralized gradient descent (CGD) in Algorithm 5. The network in CGD has sever-client architecture, which allows the server to communicate with all clients. The convergence of CGD can be described by gradient descent step

$$\tilde{\mathbf{x}}^{t+1} = \tilde{\mathbf{x}}^t - \eta \nabla f(\tilde{\mathbf{x}}^t),$$

which can find an ϵ-suboptimal solution within $\mathcal{O}(\kappa \log(1/\epsilon))$ iterations (Karimi et al., 2016). Therefore, it requires the LFO complexity of $\mathcal{O}(\kappa n \log(1/\epsilon))$. Since the sever-client architecture does not suffer from consensus error, it has communication complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$ and time complexity of $\mathcal{O}(\kappa(1 + \tau) \log(1/\epsilon))$.

The objective functions for linear regression and logistic regression are not strongly convex when $d > m$, while they satisfy the PL conditions. Please refer to Section 3.2 of Karimi et al. (2016).