Decentralized Convex Finite-Sum Optimization with Better Dependence on Condition Numbers

Yuxing Liu [1] Lesi Chen [2] Luo Luo [1 3]

Abstract

This paper studies decentralized optimization problem, where the local objective on each node is an average of a finite set of convex functions and the global function is strongly convex. We propose an efficient stochastic variance reduced first-order method that allows the different nodes to establish their stochastic local gradient estimator with different mini-batch sizes per iteration. We prove the upper bound on the computation time of the proposed method contains the dependence on the global condition number, which is sharper than the previous results that only depend on the local condition numbers. Compared with the state-of-the-art methods, we also show that our method requires less local incremental first-order oracle calls and comparable communication cost. We further perform numerical experiments to validate the advantage of our method.

1. Introduction

We study the distributed optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x) \tag{1}$$

over a connected and undirected network with m nodes, where the global objective $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex, and every local function $f_i : \mathbb{R}^d \to \mathbb{R}$ on node i has the form of

$$f_i(x) \triangleq \frac{1}{n}\sum_{j=1}^{n} f_{i,j}(x) \tag{2}$$

where each component function $f_{i,j} : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex, and n is number of components on every node. We focus on first-order decentralized optimization methods that desire all the m nodes to minimize the global objective cooperatively, which allows each node to access its local incremental first-order oracle (LIFO) and communicate with its neighbors. This problem setting is very popular in training machine learning models with large amounts of data samples.

First-order methods for decentralized convex optimization have been extensively studied in recent years. The decentralized gradient descent (DGD) method (Yuan et al., 2016) incorporates the communication steps into the full-batch gradient descent with diminishing stepsizes, leading to the sublinear convergence rates. The linear convergent methods can be achieved by introducing the gradient tracking step (Nedic & Ozdaglar, 2009; Qu & Li, 2017; Shi et al., 2015), which maintains the gradient estimator of the global objective function and iterates with the fixed stepsize. The seminal? work of Scaman et al. (2017) provided the lower bounds for the running time and communication rounds of the full-batch first-order methods for decentralized strongly convex optimization. Scaman et al. (2017) also proposed the multi-step dual accelerated (MSDA) method by applying Chebyshev acceleration (Arioli & Scott, 2014) in dual formulation, which results the optimal time complexity and communication complexity in terms of the accuracy, the maximum condition number of local functions and the spectral gap of the network. However, MSDA requires accessing the dual gradients of the local functions, which is potentially expensive. Later, Kovalev et al. (2020b); Li & Lin (2021); Song et al. (2023) developed dual-free methods that match the upper complexity bounds of MSDA but only require accessing the gradients of primal local functions. Recently, Ye et al. (2023) proposed the multi-consensus decentralized accelerated gradient descent (Mudag) to further improve the condition number dependence in the upper complexity bounds from the local functions to the global objective.

Stochastic first-order methods are widely used to speed up training large-scale machine learning models, which can take advantage of the finite-sum structure in the objective to establish efficient iteration schemes. The stochastic variance

[1] School of Data Science, Fudan University, Shanghai, China [2] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China [3] Shanghai Key Laboratory for Contemporary Applied Mathematics, Shanghai, China. Correspondence to: Luo Luo <luoluo@fudan.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

reduced methods (Defazio et al., 2014; Johnson & Zhang, 2013; Kovalev et al., 2020a; Schmidt et al., 2017; Zhang et al., 2013) establish the gradient estimator by involving the exact first-order information at the snapshot point. which leads to optimal incremental first-order oracle complexity (Agarwal & Bottou, 2015; Woodworth & Srebro, 2016) by integrating with the negative momentum (Allen-Zhu, 2017; Kovalev et al., 2020a; Qian et al., 2021). For decentralized finite-sum problem (1), Li et al. (2020); Mokhtari & Ribeiro (2016); Xin et al. (2020); Ye et al. (2020) designed variance reduced methods to improve the computation efficiency for large n, while their dependence on condition number and spectral gap do not match the full-batch methods (Kovalev et al., 2020b; Scaman et al., 2017; Song et al., 2023; Ye et al., 2023). Later, Hendrikx et al. (2020) applied Catalyst acceleration (Lin et al., 2018) to improve the condition number and spectral gap dependence. Consequently, Hendrikx et al. (2021) provided lower bounds for computation time complexity and communication complexity, and proposed a dual-based method to match their lower bounds. More recently, Li et al. (2022b) applied Katyusha acceleration (Allen-Zhu, 2017) to achieve dual-free methods which match the upper complexity bounds of Hendrikx et al. (2021). However, the condition number dependence in the analysis of existing decentralized stochastic first-order methods are based on the local functions. The potential tighter complexity bounds by considering the condition number dependence of the global function have not been studied.

In this paper, we propose computation efficient stochastic decentralized algorithm (CESAR), which establishes the local stochastic variance-reduced gradient estimators with non-uniform sampling based on the heterogeneity of the individual functions. In contrast to existing decentralized stochastic methods that fix the mini-batch size for all nodes (Hendrikx et al., 2021; Li et al., 2020; 2022b; Mokhtari & Ribeiro, 2016; Xin et al., 2020; Ye et al., 2020), the mechanism of proposed CESAR allows different nodes to access their stochastic local gradients with different mini-batch sizes in per iteration. Our theoretical analysis proves such strategy leads both computation time complexity and communication complexity to contain the dependence on global condition number, resulting tighter upper bounds than previous stochastic methods that only depend on local condition numbers (Hendrikx et al., 2021; Li et al., 2020; 2022b; Mokhtari & Ribeiro, 2016; Xin et al., 2020; Ye et al., 2020). We observed that these upper bounds of CESAR match the corresponding lower bounds if we take the global condition number into consideration. We also show CESAR enjoys sharper upper bound on LIFO complexity. Note that the computation time and the LIFO complexity of CESAR do not simply correspond to each other, since the mini-batch sizes on different nodes may not be identical. We also show the superiority of CESAR through experiments.

2. Preliminaries and Related Work

This section first formally introduces notations and settings for our problem, then provides a review of related work.

2.1. Preliminaries

We use $\|\cdot\|$ to present the Euclidean norm of a vector and the Frobenius norm of a matrix. We denote the aggregated variable as

$$\mathbf{x} = [x_1, \cdots, x_m]^\top \in \mathbb{R}^{m\times d} \quad \text{and} \quad \bar{x} = \frac{1}{m}\mathbf{1}^\top \mathbf{x} \in \mathbb{R}^{1\times d},$$

where $x_i \in \mathbb{R}^d$ is the local variable on node i and $\mathbf{1}$ is the vector of all one entries. We allow the input of functions to be presented as either column vector or row vector, e.g., $f_i(x_i)$ and $f(\bar{x})$.

We introduce the following assumptions on the decentralized finite-sum optimization problem (1).

Assumption 2.1. We assume each component function $f_{i,j}(\cdot)$ is $L_{i,j}$-smooth, each local function $f_i(\cdot)$ is L_i-smooth and the global function $f(\cdot)$ is L-smooth, i.e., there exist constants $L_{i,j}, L_i, L > 0$ such that

$$\begin{aligned} f_{i,j}(y) - f_{i,j}(x) \leq& \langle \nabla f_{i,j}(x), y - x\rangle + \frac{L_{i,j}}{2}\|y-x\|^2 \\ f_i(y) - f_i(x) \leq& \langle \nabla f_i(x), y - x\rangle + \frac{L_i}{2}\|y-x\|^2 \\ f(y) - f(x) \leq& \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|^2 \end{aligned}$$

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

Assumption 2.2. We assume each component function $f_{i,j}(\cdot)$ is convex, i.e., it holds that

$$f_{i,j}(y) - f_{i,j}(x) \geq \langle \nabla f_{i,j}(x), y - x\rangle$$

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

Assumption 2.3. We assume the global function $f(\cdot)$ is μ-strongly convex, i.e., there exists constant $\mu > 0$ such that

$$f(y) - f(x) \geq \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y-x\|^2$$

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

The strongly convex assumption ensures that problem (1) has the unique minimizer $x^* \in \mathbb{R}^d$. We say an aggregate variable $\hat{\mathbf{x}} = [\hat{x}_1, \ldots, \hat{x}_m]^\top \in \mathbb{R}^{m\times d}$ is an ϵ-suboptimal solution if it satisfies $\|\hat{\mathbf{x}} - \mathbf{1}\bar{x}^*\| \leq \epsilon$, where $\bar{x}^* = (x^*)^\top$.

Besides the parameter $L > 0$ for the smoothness of global objective $f(\cdot)$, Assumption 2.1 implies we can also define another two smoothness parameters for problem (1), i.e.,

$$\bar{L} \triangleq \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} L_{i,j} \quad \text{and} \quad \bar{L}_{\max} \triangleq \max_{i\in[m]} \frac{1}{n}\sum_{j=1}^{n} L_{i,j}.$$

Based on the definitions of L, $\bar{L}_{\max}$ and $\bar{L}$, we define three corresponding condition numbers for problem (1) as

$$\kappa \triangleq \frac{L}{\mu} \qquad \bar{\kappa} \triangleq \frac{\bar{L}}{\mu}, \qquad \text{and} \qquad \bar{\kappa}_{\max} \triangleq \frac{\bar{L}_{\max}}{\mu}$$

which is used to describe the complexity of our methods which is introduced in later sections. Note that definitions of κ, $\bar{\kappa}$ and $\bar{\kappa}_{\max}$ only require the strong convexity of global objective $f(\cdot)$. In contrast, the design and the analysis of existing decentralized accelerated stochastic methods (Hendrikx et al., 2021; Li et al., 2022b) depend on the strong convexity of every local function $f_i(\cdot)$, and their complexity bounds depend on the condition numbers

$$\kappa_{\max} \triangleq \max_{i\in[m]} \frac{L_i}{\mu_i} \quad \text{and} \quad \bar{\kappa}'_{\max} \triangleq \max_{i\in[m]} \frac{1}{n}\sum_{j=1}^{n} \frac{L_{i,j}}{\mu_i}.$$

Based on above definitions, we refer to $\bar{\kappa}_{\max}$, $\kappa_{\max}$ and $\bar{\kappa}'_{\max}$ as the local condition numbers and refer to κ and $\bar{\kappa}$ as the global condition numbers.

We can establish relationships $\kappa \leq \bar{\kappa} \leq \bar{\kappa}_{\max} \leq m\bar{\kappa}'_{\max}$,

$$\kappa \leq \kappa_{\max} \leq \bar{\kappa}'_{\max} \quad \text{and} \quad \bar{\kappa}_{\max} \leq mn\kappa.$$

In fact, the magnitude of these condition numbers may be quite different when the local data is heterogeneous. Please see the example in the following remark.

Remark 2.4. We consider the functions

$$f_{i,j}(x) = \frac{1}{2}x_{i,j}^{\top} H_{i,j} x_{i,j} + \frac{\mu_i}{2}\|x\|_2^2,$$

for $x = [x_1; \dots; x_m] \in \mathbb{R}^{2mn}$, $i \in [m]$ and $j \in [n]$, where $x_i = [x_{i,1}; \dots; x_{i,n}] \in \mathbb{R}^{2n}$, $x_{i,j} \in \mathbb{R}^2$ contains the $(2m(i-1)+2j-1)$-th and the $(2m(i-1)+2j)$-th coordinates of x, and $H_{i,j} = \operatorname{diag}(mn(L-\mu), 0) \in \mathbb{R}^{2\times 2}$ with $\mu_i = 2i\mu/(m+1)$ for some $L, \mu > 0$ such that $L \gg \mu$. Then the condition numbers hold that

$$\kappa = \frac{L}{\mu}, \quad \bar{\kappa} = \Theta\left(\frac{mnL}{\mu}\right), \quad \kappa_{\max} = \Theta\left(\frac{m^2 L}{\mu}\right),$$
$$\bar{\kappa}_{\max} = \Theta\left(\frac{mnL}{\mu}\right) \quad \text{and} \quad \bar{\kappa}'_{\max} = \Theta\left(\frac{m^2 nL}{\mu}\right).$$

This example implies different types of condition numbers may be quite different for large m and n. We show detailed expressions for these condition numbers in Appendix A.2.

For decentralized optimization, we denote $W \in \mathbb{R}^{m\times m}$ be the mixing matrix associated to the network of m nodes. We impose the following assumption on matrix W.

Assumption 2.5. For the mixing matrix $W \in \mathbb{R}^{m\times m}$, we assume (i) W is symmetric and its entry satisfies $w_{i,j} \neq 0$ if and only if node i and node j are connected in the network. (ii) $\mathbf{0} \preceq W \preceq I, W\mathbf{1} = \mathbf{1}$ and $\operatorname{null}(I - W) = \operatorname{span}(\mathbf{1})$. (iii) There exists some $\gamma \in (0, 1]$ such that $1 - \lambda_2(W) \geq \gamma$, where $\lambda_2(W)$ is the second largest eigenvalue of W.

Algorithm 1 FastMix $(\mathbf{v}_0, K)$

```
Initialize: $\mathbf{v}^{-1} = \mathbf{v}^0$, $\beta = \frac{1-\sqrt{1-\lambda_2^2(W)}}{1+\sqrt{1-\lambda_2^2(W)}}$
for $k = 0, \dots, K$
    $\mathbf{v}^{k+1} = (1+\beta)W\mathbf{v}^k - \beta\mathbf{v}^{k-1}$
end for
Output: $\mathbf{v}^K$
```

We present communication among nodes as multiplication with matrix W on aggregate variables. We can apply the multi-consensus step with Chebyshev acceleration to reduce the consensus error in decentralized optimization, which is described in Algorithm 1 (Arioli & Scott, 2014; Liu & Morse, 2011; Saad, 1984; Scaman et al., 2017). Under Assumption 2.5, it holds the following convergence result (Song et al., 2023; Ye et al., 2023).

Proposition 2.6. *Let $\bar{v} = \frac{1}{m}\mathbf{1}^{\top}\mathbf{v}^0$ for $\mathbf{v}^0 \in \mathbb{R}^{m\times d}$, then the output of Algorithm 1 holds that $\frac{1}{m}\mathbf{1}^{\top}\mathbf{v}^K = \bar{v}$ and*

$$\left\|\mathbf{v}^K - \mathbf{1}\bar{v}\right\| \leq \sqrt{14}\left(1 - c_1\sqrt{1-\lambda_2(W)}\right)^K \left\|\mathbf{v}^0 - \mathbf{1}\bar{v}\right\|,$$

where $c_1 = 1 - 1/\sqrt{2}$.

2.2. Related Work

The design and the analysis of most existing decentralized first-order methods only focus on the complexity of local condition numbers. For example, Kovalev et al. (2020a); Li & Lin (2021); Scaman et al. (2017); Song et al. (2023) proposed the full-batch methods with computation time complexity of $\mathcal{O}\big(n\sqrt{\bar{\kappa}'_{\max}}\log(1/\varepsilon)\big)$ and communication complexity of $\mathcal{O}\big(\sqrt{\kappa_{\max}/\gamma}\log(1/\varepsilon)\big)$, which are optimal with respect to local condition numbers $\kappa'_{\max}$ and $\kappa_{\max}$ (Scaman et al., 2017). In a recent work, Ye et al. (2023) provided the tighter upper bounds of $\mathcal{O}\big(n\sqrt{\kappa}\log(1/\varepsilon)\big)$ and $\tilde{\mathcal{O}}\big(\sqrt{\kappa/\gamma}\log(1/\varepsilon)\big)$ for computation time complexity and communication complexity, which are near-optimal with respect to the global condition number κ. However, the global condition numbers dependence in the complexity for decentralized finite-sum optimization has not been explored. The best-known decentralized stochastic first-order methods proposed by Hendrikx et al. (2020); Li et al. (2022b) require computation time complexity $\mathcal{O}\big((\sqrt{n\bar{\kappa}'_{\max}} + n)\log(1/\varepsilon)\big)$ and communication complexity $\mathcal{O}\big(\sqrt{\kappa_{\max}/\gamma}\log(1/\varepsilon)\big)$. Additionally, Li et al. (2022b) considered the heterogeneity of the individual functions and established the local gradient estimators by importance sampling, however, their methods enforce all nodes access their stochastic local gradients with identical mini-batch sizes. Intuitively, this may affect the use of global properties of our problem, since the importance of local functions on different nodes can be quite different. We present the upper complexity bounds of existing methods and compare them with our results in Table 1

Table 1. We summarize the upper complexity bounds on computation time, communication rounds and LIFO calls for proposed CESAR and existing methods. We use notations $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide the logarithmic factors of m, n and condition numbers.

Methods	Computation Time	# Communication	# LIFO
decentralized full-batch algorithms			
MSDA+CA (Scaman et al., 2017)	$\mathcal{O}\Big(n\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\sqrt{\frac{\kappa_{\max}}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(mn\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$
OPAPC (Kovalev et al., 2020b)	$\mathcal{O}\Big(n\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\sqrt{\frac{\kappa_{\max}}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(mn\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$
ACC-GT+CA (Li & Lin, 2021)	$\mathcal{O}\Big(n\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\sqrt{\frac{\kappa_{\max}}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(mn\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$
OGT (Song et al., 2023)	$\mathcal{O}\Big(n\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\sqrt{\frac{\kappa_{\max}}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(mn\sqrt{\bar\kappa'_{\max}}\log\left(\frac{1}{\varepsilon}\right)\Big)$
Mudag (Ye et al., 2023)	$\mathcal{O}\Big(n\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\tilde{\mathcal{O}}\Big(\sqrt{\frac{\kappa}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(mn\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)\Big)$
Decentralized stochastic methods			
DSA (Mokhtari & Ribeiro, 2016)	$\mathcal{O}\Big(\big(n\bar\kappa'_{\max}+\frac{(\bar\kappa'_{\max})^4}{\gamma}\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\frac{(\bar\kappa'_{\max})^4}{\gamma}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\big(\frac{m(\bar\kappa'_{\max})^2}{\gamma^2}+mn\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$
GT-SVRG / GT-SAGA (Xin et al., 2020)	$\mathcal{O}\Big(\big(\frac{\bar\kappa^2_{\max}}{\gamma^2}+n\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\big(\frac{\bar\kappa^2_{\max}}{\gamma^2}+n\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\big(\frac{m\bar\kappa^2_{\max}}{\gamma^2}+mn\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$
DVR+Catalyst (Hendrikx et al., 2020)	$\tilde{\mathcal{O}}\Big(\big(\sqrt{n\bar\kappa'_{\max}}+n\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\tilde{\mathcal{O}}\Big(\sqrt{\frac{n\kappa^2_{\max}}{\bar\kappa'_{\max}\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\tilde{\mathcal{O}}\Big(\big(m\sqrt{n\bar\kappa'_{\max}}+mn\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$
PMGT-LSVRG (Ye et al., 2020)	$\mathcal{O}\Big(\big(\bar\kappa_{\max}+n\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\tilde{\mathcal{O}}\Big(\frac{\bar\kappa_{\max}+n}{\sqrt{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\big(m\bar\kappa_{\max}+mn\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$
ADFS (Hendrikx et al., 2021)	$\mathcal{O}\Big(\big(\sqrt{n\bar\kappa'_{\max}}+n\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\sqrt{\frac{\kappa_{\max}}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\big(m\sqrt{n\bar\kappa'_{\max}}+mn\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$
Acc-VR-EXTRA+CA (Li et al., 2022b)	$\mathcal{O}\Big(\big(\sqrt{n\bar\kappa'_{\max}}+n\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\sqrt{\frac{\kappa_{\max}}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\big(m\sqrt{n\bar\kappa'_{\max}}+mn\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$
CESAR Theorem 3.7	$\tilde{\mathcal{O}}\Big(\big(\sqrt{\frac{n\bar\kappa_{\max}}{m}}+\sqrt{\kappa}+n\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\tilde{\mathcal{O}}\Big(\sqrt{\frac{\kappa}{\gamma}}\log\left(\frac{1}{\varepsilon}\right)\Big)$	$\mathcal{O}\Big(\big(\sqrt{mn\bar\kappa_{\max}}+mn\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$

3. The Algorithm and Main Results

This section proposes the computation efficient stochastic decentralized algorithm (CESAR) and provides theoretical analysis for its upper complexity bounds.

3.1. The Algorithm

We present the details of CESAR in Algorithm 2, which extends the techniques of variance reduction (Johnson & Zhang, 2013; Zhang et al., 2013) and negative momentum (Allen-Zhu, 2017; Kovalev et al., 2020a; Qian et al., 2021) to decentralized optimization.

The computational efficiency of CESAR mainly comes from our local gradient estimator, i.e.,

$$v_i^t = u_i^t + \sum_{j=1}^{n} \frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t) - \nabla f_{i,j}(w_i^t)\right), \tag{3}$$

where u_i^t is the local gradient estimator for node i at the snapshot point w_i^t and $\xi_{i,j}^{t+1}$ is distributed to

$$\xi_{i,j}^{t+1} \sim \text{Bernoulli}(q_{i,j}), \tag{4}$$

with

$$q_{i,j} = \min\left(1, \frac{bL_{i,j}}{mn\bar{L}_{\max}}\right) \quad \text{and} \quad b = \sqrt{\frac{mn\bar\kappa_{\max}}{\kappa}}. \tag{5}$$

At the t-th iteration, node i establishes the estimator v_i^t by accessing local stochastic gradients $\nabla_{i,j} f(x_i^t)$ and $\nabla_{i,j} f(w_i^t)$ for all $j \in [n]$ such that $\xi_{i,j}^{t+1} = 1$. This implies the number of LIFO calls on node i can be written as

$$Y_i^t = 2\sum_{j=1}^{n} \xi_{i,j}^{t+1}, \tag{6}$$

which is a random variable. Therefore, our sample sizes for different nodes are not fixed. If all $L_{i,1}, \ldots, L_{i,n}$ are small, the node i may even skip any LIFO computation at some iteration. On the other hand, the nodes contain individual functions with large smoothness parameters $L_{i,j}$ will tend to perform more LIFO computation, which indicates our strategy indeed uses potential heterogeneity among nodes and finally leads to better complexity dependence.

Recall that all nodes perform the computation in parallel, then the total computation time for achieving all of the local estimators $v_1^t, \dots, v_m^t$ relies on the node that requires the maximum number of LIFO calls at the t-th iteration, which takes the computation time complexity of

$$\max_{i\in[m]} Y_i^t = 2\max_{i\in[m]} \sum_{j=1}^{n} \xi_{i,j}^{t+1}. \tag{7}$$

Based on the specifically designed variance-reduction scheme, we employ the Katyusha-like acceleration (Allen-Zhu, 2017) in to achieve better convergence. We also incorporate the multi-consensus steps (Ye et al., 2023) with gradient tracking (Nedic & Ozdaglar, 2009; Qu & Li, 2017; Shi et al., 2015) and Chebyshev acceleration (Arioli & Scott, 2014; Liu & Morse, 2011; Saad, 1984; Scaman et al., 2017) into our algorithm, which results the global condition number dependence in convergence rate.

3.2. The Complexity Analysis

In this subsection, we upper bound the expectation of the random variable $\max_{i\in[m]} Y_i^t$ defined in equation (7) and provide the convergence analysis for CESAR (Algorithm 2), which indicates the superiority of our method in theoretical.

3.2.1. Computation Time of Per Iteration

We start our analysis from the Chernoff bound for Bernoulli variables (Motwani & Raghavan, 1995; Zhang, 2023).

Lemma 3.1. *Suppose random variables $X_1, ..., X_n$ are independent and each of X_i is distributed to $\mathrm{Bernoulli}(p_i)$ for some $p_i \in [0,1]$. We let $X = \sum_{j=1}^n X_j$ and $\nu = \mathbb{E}[X]$. Then for any $\delta > 0$, it holds*

$$\mathbb{P}\left(X \geq (1+\delta)\nu\right) \leq \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{\nu}. \tag{8}$$

Based on Lemma 3.1, we achieve the following upper bound for the sum of Bernoulli variables with high probability.

Lemma 3.2. *Suppose random variables $Z_1, \dots, Z_m$ are distributed to $Z_i = \sum_{j=1}^n X_{i,j}$ for all $i \in [m]$, where the random variables $X_{1,1}, \dots X_{m,n}$ are mutually independent and each of $X_{i,j}$ is distributed to $\mathrm{Bernoulli}(p_{i,j})$ for some $p_{i,j} \in [0,1]$. Then it holds*

$$\mathbb{P}\left(\exists i \in [m], Z_i \geq 2\mathrm{e}\max\left\{\mathbb{E}[Z_i], (\ln mn)^2\right\}\right) \leq \frac{1}{mn}.$$

Recall that it always holds $Y_i^t \leq 2n$ for all $i \in [m]$, then applying Lemma 3.2 results the following upper bound of the expectation of $\max_{i\in[m]} Y_i^t$.

Algorithm 2 CESAR

```
Input: the initial point $\bar{w}^0$, probabilities $p$ and $q_{i,j}$ for all $i \in [m]$ and $j \in [n]$, numbers of consensus steps $K$ and $K_{\mathrm{out}}$, total iterations number $T$, and parameters $\eta$, $\theta_1$, $\theta_2$ and $\sigma$.
$\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{w}^0 = \mathbf{1}\bar{w}^0, \quad \mathbf{v}^{-1} = \mathbf{s}^{-1} = \mathbf{0}$
$\mathbf{g}^0 = \mathbf{u}^0 = [\nabla f_1(\bar{w}^0); \nabla f_2(\bar{w}^0); \cdots; \nabla f_m(\bar{w}^0)]$
for $t = 0, \dots, T$
    $\mathbf{x}^t = \theta_1\mathbf{z}^t + \theta_2\mathbf{w}^t + (1-\theta_1-\theta_2)\mathbf{y}^t$
    parallel for $i = 1, \dots, m$ do
        $\xi_{i,j}^{t+1} \sim \mathrm{Bernoulli}(q_{i,j})$
        $v_i^t = u_i^t + \sum_{j=1}^{n} \frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t) - \nabla f_{i,j}(w_i^t)\right)$
    end parallel for
    $\mathbf{s}^t = \mathrm{FastMix}\left(\mathbf{s}^{t-1} + \mathbf{v}^t - \mathbf{v}^{t-1}, K\right)$
    $\mathbf{z}^{t+1} = \mathrm{FastMix}\left(\frac{1}{1+\eta\sigma}\left(\eta\sigma\mathbf{x}^t + \mathbf{z}^t - \frac{\eta}{L}\mathbf{s}^t\right), K\right)$
    $\mathbf{y}^{t+1} = \mathrm{FastMix}\left(\mathbf{x}^t + \theta_1(\mathbf{z}^{t+1} - \mathbf{z}^t), K\right)$
    $\zeta^{t+1} \sim \mathrm{Bernoulli}(p)$
    parallel for $i = 1, \dots, m$ do
        $w_i^{t+1} = \begin{cases} y_i^t, & \text{if } \zeta^{t+1} = 1 \\ w_i^t, & \text{otherwise} \end{cases}$
        $g_i^{t+1} = \begin{cases} \nabla f_i(w_i^{t+1}), & \text{if } \zeta^{t+1} = 1 \\ g_i^t, & \text{otherwise} \end{cases}$
    end parallel for
    $\mathbf{u}^{t+1} = \mathrm{FastMix}\left(\mathbf{u}^t + \mathbf{g}^{t+1} - \mathbf{g}^t, K\right)$
end for
Output: $\mathbf{y}_{\mathrm{out}} = \mathrm{FastMix}(\mathbf{y}_T, K_{\mathrm{out}})$ .
```

Theorem 3.3. *Following notations of (4)–(6), we have*

$$\mathbb{E}\left[\max_{i\in[m]} Y_i^t\right] \leq \mathcal{O}\left(\sqrt{\frac{n\bar{\kappa}_{\max}}{m\kappa}} + (\ln mn)^2\right).$$

Theorem 3.3 implies that CESAR takes the computation time of $\tilde{\mathcal{O}}\big(\big(\sqrt{n\bar{\kappa}_{\max}/m} + n\big)/\sqrt{\kappa}\big)$ to achieve $v_1^t, \dots, v_m^t$. Additionally, the step

$$g_i^{t+1} = \begin{cases} \nabla f_i(w_i^{t+1}), & \text{if } \zeta^{t+1} = 1, \\ g_i^t, & \text{otherwise}, \end{cases}$$

with $\zeta^{t+1} \sim \mathrm{Bernoulli}(p)$ requires the computation time of $\mathcal{O}(np)$ in expectation. Therefore, the overall expected computation time per iteration can be upper bounded by

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{n\bar{\kappa}_{\max}}{m\kappa}}+np\right).$$

We provide detailed proofs for Lemma 3.2 and Theorem 3.3 in Appendix F, which cannot be achieved through simple calculations like the analysis of fixed mini-batch size.

3.2.2. Convergence Analysis

The convergence analysis of CESAR (Algorithm 2) is based on considering the mean vectors

$$\bar{z}^t=\frac{1}{m}\sum_{i=1}^m z_i^t,\quad \bar{w}^t=\frac{1}{m}\sum_{i=1}^m w_i^t \quad\text{and}\quad \bar{y}^t=\frac{1}{m}\sum_{i=1}^m y_i^t.$$

We define the corresponding Lyapunov function as follows

$$V^t \triangleq \mathcal{Z}^t+\mathcal{Y}^t+\mathcal{W}^t, \tag{9}$$

where

$$\mathcal{Z}^t \triangleq \frac{L(1+\eta\sigma)}{2\eta}\left\|\bar{z}^t-\bar{x}^*\right\|^2,\quad \mathcal{Y}^t \triangleq \frac{1}{\theta_1}(f(\bar{y}^t)-f(\bar{x}^*))$$
$$\mathcal{W}^t \triangleq \frac{\theta_2}{p\lambda\theta_1}(f(\bar{w}^t)-f(\bar{x}^*)) \quad\text{and}\quad \eta,\sigma,\lambda,\theta_1,\theta_2>0.$$

Note that here λ is an constant that only appears in our analysis. Besides the Lyapunov function V^t, we also need to consider the consensus error aroused from the decentralized setting, which leads to more complicated analysis than accelerated variance-reduced methods on a single machine (Allen-Zhu, 2017; Kovalev et al., 2020a; Qian et al., 2021).

We first provide the recursion for V^t by involving consensus error $\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2$ and $\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2$.

Lemma 3.4. *Under Assumption 2.1, 2.2, 2.3 and 2.5, we run Algorithm 2 by taking q_{ij} as (5), $\eta=1/(13\theta_1)$, $\sigma=\mu/L$, $\lambda\in[1/2,1)$ and $\theta_1,\theta_2\in(0,1/2)$. Then it holds*

$$\mathbb{E}\left[V^{t+1}\right]\le\beta V^t+\sqrt{\frac{2\eta\hat{L}^2V^t}{(1+\eta\sigma)mL}}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|$$
$$+\left(\frac{12L\hat{L}}{b}+\frac{2\hat{L}^2}{b}\right)\cdot\left(\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\right),$$

where we denote $\hat{L}=\max_{i\in[m],j\in[n]}L_{i,j}$ and

$$\beta=\max\left\{\frac{1}{1+\eta\sigma},1-\left(\theta_1+\theta_2-\frac{\theta_2}{\lambda}\right),1-p(1-\lambda)\right\}.$$

We bound the consensus error by introducing the vector

$$r^t=\frac{L}{m}\Big[\frac{1}{L^2}\|\mathbf{u}^t-\mathbf{1}\bar{u}^t\|^2,\ \frac{1}{L^2}\left\|\mathbf{s}^t-\mathbf{1}\bar{s}^t\right\|^2,$$
$$\|\mathbf{z}^t-\mathbf{1}\bar{z}^t\|^2,\ \|\mathbf{y}^t-\mathbf{1}\bar{y}^t\|^2\Big]^\top\in\mathbb{R}^4.$$

We apply Proposition 2.6 to analyze the communication steps in CESAR, which leads to the following results on consensus error.

Lemma 3.5. *Under the settings of Lemma 3.4, we run Algorithm 2 by specifically taking $K=\left\lceil(\log(1/\rho))/\sqrt{\gamma}\right\rceil$ with $1/\rho=\mathcal{O}\left(\mathrm{poly}(m,n,\kappa)\right)$. Then it holds*

$$\mathbb{E}\left[r^{t+1}\right]\le\rho^2\left(Ar^t+h^t\right)$$

for some matrix $A\in\mathbb{R}^{4\times4}$ and vector $h^t\in\mathbb{R}^4$ such that

$$\|A\|\le\frac{40m^3n^3}{b}$$

and

$$\left\|h^t\right\|\le\frac{48m^3n^3}{b}\max\left\{\frac{2}{13\theta_1},\frac{65\theta_1}{6\theta_2}\right\}\cdot\left(V^{t+1}+V^t\right)$$
$$+\frac{(66+324\rho^2)m^3n^3}{b}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2$$
$$+\frac{324\rho^2m^3n^3}{b}\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2.$$

Remark 3.6. The notation of "$\le$" between vectors means that each corresponding scalar entry has less than or equal to relationship. The explicit expressions of A and h^t are so complicated and we present them in Appendix C.2.

By connecting the above two lemmas, we obtain the main convergence results for CESAR.

Theorem 3.7. *Under Assumption 2.1, 2.2, 2.3 and 2.5, we run Algorithm 2 by taking*

$$p=\max\left\{\frac{1}{2\sqrt{\kappa}},\frac{\bar{\kappa}_{\max}}{2\kappa b}\right\},\qquad b=\sqrt{\frac{mn\bar{\kappa}_{\max}}{\kappa}},$$
$$q_{i,j}=\min\left\{1,\frac{bL_{i,j}}{mn\bar{L}_{\max}}\right\},\qquad \eta=\frac{1}{13\theta_1},$$
$$\sigma=\frac{\mu}{L},\quad \lambda\in\left[\frac{2}{3},1\right),\quad \theta_1=\frac{1}{2\sqrt{\kappa}},\quad \theta_2=\frac{\bar{\kappa}_{\max}}{2\kappa b}$$

and setting K by following Lemma 3.5. Then it holds

$$\mathbb{E}\left[V^t+\left\|r^t\right\|\right]\le2\alpha^t\left(V^0+\left\|r^0\right\|\right),$$

where

$$\begin{aligned}\alpha=&1-\min\left\{\frac{\eta}{\kappa},\frac{\theta_1+\theta_2-\theta_2/\lambda}{2},\frac{p(1-\lambda)}{2}\right\}\\=&1-\Theta\left(\frac{1}{\sqrt{\kappa}}\right).\end{aligned}$$

3.2.3. Upper Complexity Bounds

Based on Theorem 3.3 and 3.7, we achieve upper complexity bounds for CESAR as follows.

Table 2. We summarize the lower complexity bounds on computation time, communication rounds and LIFO calls. Note that we present the lower bounds by considering different types of condition numbers.

Condition Numbers	Computation Time	# Communication	# LIFO
$\kappa_{\max}$ and $\bar{\kappa}'_{\max}$ (Hendrikx et al., 2021)	$\Omega\Big(\big(\sqrt{n\bar{\kappa}'_{\max}}+n\big)\log\big(\frac{1}{\varepsilon}\big)\Big)$	$\Omega\Big(\sqrt{\frac{\kappa_{\max}}{\gamma}}\log\big(\frac{1}{\varepsilon}\big)\Big)$	–
κ, $\bar{\kappa}_{\max}$ and $\bar{\kappa}$ Theorem 4.1, 4.2 and 4.3	$\Omega\Big(\big(\sqrt{\frac{n\bar{\kappa}_{\max}}{m}}+\sqrt{\kappa}+n\big)\log\big(\frac{1}{\varepsilon}\big)\Big)$	$\Omega\Big(\sqrt{\frac{\kappa}{\gamma}}\log\big(\frac{1}{\varepsilon}\big)\Big)$	$\Omega\Big(\big(\sqrt{mn\bar{\kappa}}+mn\big)\log\big(\frac{1}{\varepsilon}\big)\Big)$

Corollary 3.8. *Under the assumptions and the settings of Theorem 3.7, we can achieve an ϵ-suboptimal solution of Problem (1) by Algorithm 2 with $T=\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ and $K_{\rm out}=\tilde{\mathcal{O}}(\sqrt{1/\gamma})$, which takes computation time complexity of $\tilde{\mathcal{O}}\big(\big(\sqrt{n\bar{\kappa}_{\max}/m}+\sqrt{\kappa}+n\,\big)\log(1/\epsilon)\big)$, communication complexity of $\tilde{\mathcal{O}}(\sqrt{\kappa/\gamma}\log(1/\epsilon))$ and LIFO complexity of $\mathcal{O}\left(\left(\sqrt{mn\bar{\kappa}_{\max}}+mn\,\right)\log(1/\epsilon)\right)$ in expectation.*

All of the complexity bounds in Corollary 3.8 are tighter (no worse) than the results of the state-of-the-art decentralized first-order methods (Table 1), which can be verified by the relationships among condition numbers shown in Section 2:

- Compared with best-known stochastic methods ADFS (Hendrikx et al., 2021) and ACC-VR-EXTRA+CA (Li et al., 2022b), all of the computation time complexity, the communication complexity and the LIFO complexity of CESAR are no worse, since it holds that $\kappa\leq\kappa_{\max}$ and $\kappa_{\max}\leq\bar{\kappa}'_{\max}$. Furthermore, our complexity bounds may be much tighter when the data is heterogeneous, such as the example shown in Remark 2.4.
- Compared with the best-known full-batch methods Mudag (Ye et al., 2023), our method has the no worse computation time complexity and LIFO complexity, and the comparable communication complexity, since it holds that $\bar{\kappa}_{\max}\leq mn\kappa$. In the case of $\bar{\kappa}_{\max}\approx\kappa$ and $m,n\gg 1$, our computation time complexity and LIFO complexity are much tighter.

4. The Lower Bounds and Discussion

We show the lower complexity bounds on communication and computation time in the following theorems.

Theorem 4.1. *Let $\gamma\in(0,1]$, $\kappa\geq 1$ and $n\in\mathbb{N}$. There exist a mixing matrix $W\in\mathbb{R}^{m\times m}$ with $m\geq\sqrt{3/\gamma}$ and $m\times n$ functions $f_{i,j}:\ell_2\to\mathbb{R}$ such that each $f_{i,j}$ is convex and smooth, function $f\triangleq 1/(mn)\sum_{i=1}^{m}\sum_{j=1}^{n}f_{i,j}$ is μ-strongly convex and L-smooth such that $\kappa\geq L/\mu$. Then any black-box procedure for achieving an ϵ-suboptimal solution of Problem (1) needs at least the communication rounds of $\Omega\big(\sqrt{\kappa/\gamma}\log(1/\varepsilon)\big)$.*

Theorem 4.2. *Let $\kappa,\bar{\kappa}_{\max}\geq 1$ and $m,n\in\mathbb{N}$. There exist a mixing matrix $W\in\mathbb{R}^{m\times m}$ and $m\times n$ functions $f_{i,j}:\ell_2\to\mathbb{R}$ such that each $f_{i,j}$ is convex and smooth, function $f\triangleq 1/(mn)\sum_{i=1}^{m}\sum_{j=1}^{n}f_{i,j}$ is μ-strongly convex and L-smooth such that $\bar{\kappa}_{\max}\geq\max_{i\in[m]}\sum_{j=1}^{n}L_{i,j}/(n\mu)$, where $L_{i,j}$ is the smooth parameter of $f_{i,j}$. Then any black-box procedure for achieving an ϵ-suboptimal solution of Problem (1) needs at least the computation steps of $\Omega\big(\big(n+\sqrt{n\bar{\kappa}_{\max}/m}+\sqrt{\kappa}\,\big)\log(1/\varepsilon)\big)$.*

The black-box procedure in the above theorems is followed by the definition of Hendrikx et al. (2021). Due to the space limitation, we present its details in Appendix D. Note that the computation step in our description is the procedure where several nodes access their own local LIFO in parallel, which corresponds to the computation time complexity and is different from the overall LIFO complexity.

We then provide the lower bound on LIFO complexity by following the analysis in non-distributed settings (Agarwal & Bottou, 2015; Woodworth & Srebro, 2016).

Theorem 4.3. *Let $\gamma\in(0,1]$, $\bar{\kappa}\geq 1$ and $m,n\in\mathbb{N}$. There exist a mixing matrix $W\in\mathbb{R}^{m\times m}$ and $m\times n$ functions $f_{i,j}:\ell_2\to\mathbb{R}$ such that each $f_{i,j}$ is convex and smooth, function $f\triangleq 1/(mn)\sum_{i=1}^{m}\sum_{j=1}^{n}f_{i,j}$ is μ-strongly convex and L-smooth such that $\bar{\kappa}\geq 1/(mn)\sum_{i=1}^{m}\sum_{j=1}^{n}L_{i,j}/\mu$, where $L_{i,j}$ is the smooth parameter of $f_{i,j}$. Then any black-box procedure for achieving an ϵ-suboptimal solution of Problem (1) needs at least $\Omega((mn+\sqrt{mn\bar{\kappa}}\,)\log(1/\varepsilon))$ LIFO calls.*

We compare our lower complexity bounds with related work in Table 2. The results in Theorem 4.1 and 4.2 nearly match the corresponding upper bounds in Corollary 3.8. However, lower bound on LIFO complexity in Theorem 4.3 depends on $\bar{\kappa}$, while the upper bound in CESAR 3.8 depends on $\bar{\kappa}_{\max}$. This implies the results in this work cannot lead to the optimality of LIFO complexity. How to fill the gap between $\bar{\kappa}$ and $\bar{\kappa}_{\max}$ is still an open problem.

We consider the case of $n=1$, which ignores the finite-sum structure in local functions. Then CESAR has communication complexity of $\tilde{\mathcal{O}}\left(\sqrt{\kappa}\log(1/\epsilon)\right)$ and computation time complexity of $\tilde{\mathcal{O}}\left(\sqrt{\kappa}\log(1/\epsilon)\right)$, which nearly match the corresponding upper bounds of near-optimal full-batch first-order methods Mudag (Ye et al., 2023). In this case, CESAR has the local first-order oracle complexity

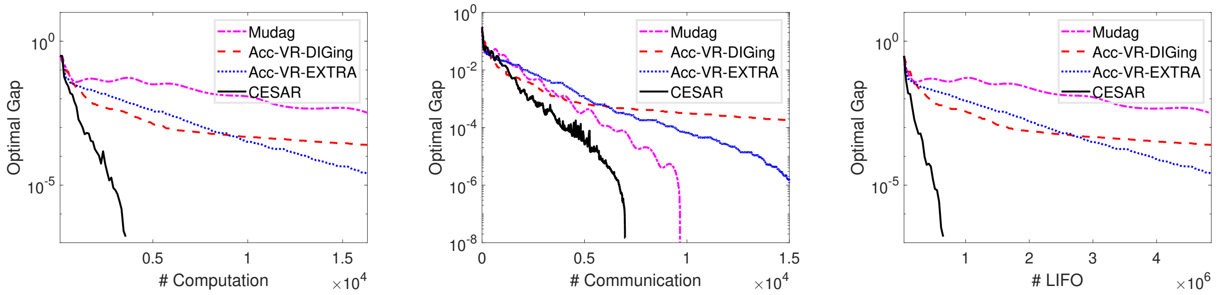

Figure 1. The empirical results for logistic regression on dataset “a9a”.

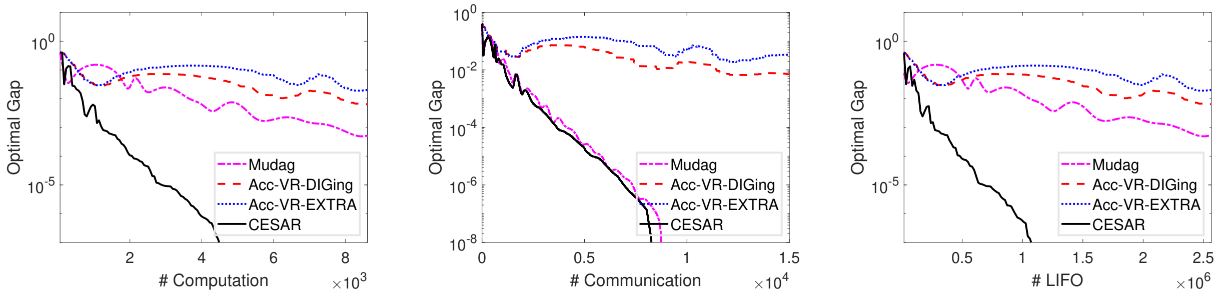

Figure 2. The empirical results for logistic regression on dataset “w6a”.

Table 3. The condition numbers of the problem in experiments.

	κ	$\kappa_{\max}$	$\bar{\kappa}_{\max}$	$\bar{\kappa}'_{\max}$
a9a	1.58×10^4	1.70×10^4	3.50×10^4	3.50×10^4
w6a	6.59×10^3	1.46×10^5	2.06×10^5	2.06×10^5

of $\mathcal{O}((\sqrt{m\bar{\kappa}_{\max}} + m)\log(1/\epsilon))$, which is sharper than the complexity of $\mathcal{O}(m\sqrt{\kappa}\log(1/\epsilon))$ in Mudag when condition numbers satisfy $\bar{\kappa}_{\max} \le m\kappa$. Intuitively, CESAR encourages most nodes to completely skip local gradient computation in some iterations, however, Mudag always requires all of the nodes to perform the local gradient computation in every iteration. In this view, CESAR takes less overall energy consumption in computation, which is more friendly to applications in networks within the limited computational resources, such as wireless sensors (Rabbat & Nowak, 2004), mobile devices (Wang et al., 2020) and smart home appliances (Joo & Choi, 2017). We then consider the case of $m = 1$, which is the finite-sum optimization on a single machine. Then CESAR has the incremental first-order oracle complexity of $\mathcal{O}\big((\sqrt{mn\bar{\kappa}} + mn)\log(1/\varepsilon)\big)$, matching the result of near-optimal first-order method Katyusha (Allen-Zhu, 2017) in non-distributed setting.

5. Numerical Experiments

In this section, we provide the numerical experiments to compare the performance of CESAR with baseline methods Mudag (Ye et al., 2023), Acc-VR-EXTRA and Acc-VR-DIGING (Li et al., 2022a). We consider the problem of ℓ_2-regularized logistic regression, which is formulated by

$$\min_{x\in\mathbb{R}^d} f(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x) \quad \text{with} \quad f_i(x) = \frac{1}{n}\sum_{j=1}^{n} f_{i,j}(x)$$
$$\text{and} \quad f_{i,j}(x) = \log\left(1 + \exp(-b_{i,j}a_{i,j}^\top x)\right) + \frac{\mu}{2}\|x\|_2^2,$$

where $a_{i,j} \in \mathbb{R}^d$ is the feature vector of the j-th sample on the i-th node, $b_{i,j} \in \{-1, 1\}$ is the corresponding label and $\mu > 0$ is the hyperparamter.

We conduct our experiments on datasets “a9a” and “w6a” (Chang & Lin, 2011) and let $\mu = 10^{-4}$ and $m = 300$. We set the mixing matrix W to be associated with a random graph that each edge is connected with probability $1/30$, which leads to $1 - \lambda_2(W) \approx 0.0382$. The condition numbers in our problem are listed in Table 3.

We present the experimental results in Figure 1 and 2, where the optimal gap is defined as $\frac{1}{m}\sum_{i=1}^{m} f_i(x_i) - f^*$. The number of computation corresponds to $\sum_{t=0}^{T-1}\max_{i\in[m]} Y_i^t$ for CESAR and Tb' for other methods, where Y_i^t is defined in equation (6) and b' is the batch-size in baseline methods.

We can observe that our CESAR always outperforms the stochastic methods Acc-VR-DIGing and Acc-VR-EXTRA in all measures. CESAR also performs better than Mudag on the complexity of computation and LIFO, and it has a comparable communication cost to Mudag. All of these results validate our theoretical analysis. Concretely, the gap of communication complexity between CESAR and baseline algorithms Acc-VR-DIGing and Acc-VR-EXTRA on "w6a" is much larger than the one in "a9a", since the ratio between $\kappa_{\max}$ and κ in "w6a" is larger than the one in "a9a". Additionally, the computation complexity and LIFO complexity of CESAR are much better than baselines, since both datasets hold $\bar{\kappa}_{\max} \approx \bar{\kappa}'_{\max}$, and $\bar{\kappa}_{\max}$ is much smaller than $m\bar{\kappa}'_{\max}$.

6. Conclusion

This paper has studied decentralized convex finite-sum optimization. We have proposed an accelerated stochastic variance-reduced first-order algorithm with non-uniform sampling, which leads to complexity bounds with better dependence on condition numbers. We have validated our theory by numerical experiments. In future work, we are interested in studying decentralized nonconvex optimization by considering different kinds of condition numbers.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62206058), Shanghai Sailing Program (22YF1402900), and Shanghai Basic Research Program (23JC1401000).

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.

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *STOC*, 2017.

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

Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Hendrikx, H., Bach, F., and Massoulié, L. Dual-free stochastic decentralized optimization with variance reduction. In *NeurIPS*, 2020.

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

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

Joo, I.-Y. and Choi, D.-H. Distributed optimization framework for energy management of multiple smart homes with distributed energy resources. *IEEE Access*, 5:15551–15560, 2017.

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.

Li, B., Cen, S., Chen, Y., and Chi, Y. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *Journal of Machine Learning Research*, 21(1):7331–7381, 2020.

Li, H. and Lin, Z. Accelerated gradient tracking over time-varying graphs for decentralized optimization. *arXiv preprint arXiv:2104.02596*, 2021.

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, 2022a.

Li, H., Lin, Z., and Fang, Y. Variance reduced extra and diging and their optimal acceleration for strongly convex decentralized optimization. *The Journal of Machine Learning Research*, 23(1):10057–10097, 2022b.

Lin, H., Mairal, J., and Harchaoui, Z. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(1): 7854–7907, 2018.

Liu, J. and Morse, A. S. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2): 160–165, 2011.

Mokhtari, A. and Ribeiro, A. DSA: Decentralized double stochastic averaging gradient algorithm. *Journal of Machine Learning Research*, 17(1):2165–2199, 2016.

Motwani, R. and Raghavan, P. *Randomized algorithms*. Cambridge university press, 1995.

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

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

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.

Rabbat, M. G. and Nowak, R. D. Decentralized source localization and tracking [wireless sensor networks]. In *ICASSP*, 2004.

Saad, Y. Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems. *Mathematics of Computation*, 42(166):567–588, 1984.

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*, 162:83–112, 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.

Wang, M., Xu, C., Chen, X., Zhong, L., Wu, Z., and Wu, D. O. Bc-mobile device cloud: A blockchain-based decentralized truthful framework for mobile device cloud. *IEEE Transactions on Industrial Informatics*, 17(2):1208–1219, 2020.

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

Xin, R., Khan, U. A., and Kar, S. Variance-reduced decentralized stochastic optimization with accelerated convergence. *IEEE Transactions on Signal Processing*, 68: 6255–6271, 2020.

Ye, H., Xiong, W., and Zhang, T. PMGT-VR: A decentralized proximal-gradient algorithmic framework with variance reduction. *arXiv preprint arXiv:2012.15010*, 2020.

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., Ling, Q., and Yin, W. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

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

Zhang, T. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023. doi: 10.1017/9781009093057.

Appendix Outlines

In Appendix A, we provide the proofs for relationships for condition numbers and detailed calculation for Remark 2.4.

In Appendix B, C and D, we provide proofs for the complexity analysis of CESAR in Section 3.2, which is organized as follows:

- In Appendix B, we consider the computation time per iteration discussed in Section 3.2.1.
- In Appendix C.1, we provide the proof of Lemma 3.4, which gives the recursion of Lyapunov function V^t.
- In Appendix C.2, we provide the proof Lemma 3.5, which gives the recursion for vector

$$r^t = \frac{L}{m}\Big[\frac{1}{L^2}\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2, \frac{1}{L^2}\left\|\mathbf{s}^t - \mathbf{1}\bar{s}^t\right\|^2, \left\|\mathbf{z}^t - \mathbf{1}\bar{z}^t\right\|^2, \left\|\mathbf{y}^t - \mathbf{1}\bar{y}^t\right\|^2\Big]^\top.$$

  Additionally, we present the expression of A and e_t in the statement of Lemma C.15.

- In Appendix C.3, we provide the proof of Theorem 3.7 by applying Lemma 3.4 and Lemma C.15.
- In Appendix C.4, we provide the proof of Corollary 3.8 by applying Theorem 3.3 and 3.7.
- In Appendix D, we provide the proofs for the lower bounds in Section 4.

A. Relationships for the Condition Numbers

Recall that we have defined smoothness parameters

$$\bar{L} \triangleq \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n L_{i,j} \qquad \text{and} \qquad \bar{L}_{\max} \triangleq \max_{i\in[m]}\frac{1}{n}\sum_{j=1}^n L_{i,j},$$

and the condition numbers

$$\kappa \triangleq \frac{L}{\mu}, \qquad \bar{\kappa} \triangleq \frac{\bar{L}}{\mu}, \qquad \bar{\kappa}_{\max} \triangleq \frac{\bar{L}_{\max}}{\mu}, \qquad \kappa_{\max} \triangleq \max_{i\in[m]}\frac{L_i}{\mu_i} \qquad \text{and} \qquad \bar{\kappa}'_{\max} \triangleq \max_{i\in[m]}\frac{1}{n}\sum_{j=1}^n\frac{L_{i,j}}{\mu_i},$$

where L, L_i and $L_{i,j}$ are the smoothness parameters of $f(\cdot)$, $f_i(\cdot)$ and $f_{i,j}(\cdot)$ respectively.

Now, we prove and verify the relationship among these condition numbers and verify the example in Remark 2.4.

A.1. The Inequalities of the Condition Numbers

Based on the definitions, we have the following proposition.

Proposition A.1. *Assume that each $f_{i,j}(\cdot)$ is convex and smooth for $i \in [m]$ and $j \in [n]$. Then the condition numbers hold the relationships*

$$\kappa \leq \kappa_{\max} \leq \bar{\kappa}'_{\max}, \qquad \kappa \leq \bar{\kappa} \leq \bar{\kappa}_{\max} \leq m\bar{\kappa}'_{\max} \qquad \textit{and} \qquad \bar{\kappa}_{\max} \leq mn\kappa.$$

Proof. For the inequality $\kappa \leq \kappa_{\max}$, we first verify the inequalities

$$L \leq \frac{1}{m}\sum_{i=1}^m L_i \qquad \text{and} \qquad L_i \leq \frac{1}{n}\sum_{j=1}^n L_{i,j}. \tag{10}$$

We can prove $L \leq \sum_{i=1}^m L_i/m$ by triangle inequality, i.e., it holds

$$\left\|\nabla f(x) - \nabla f(y)\right\| = \left\|\frac{1}{m}\sum_{i=1}^m(\nabla f_i(x) - \nabla f_i(y))\right\| \leq \frac{1}{m}\sum_{i=1}^m\left\|\nabla f_i(x) - \nabla f_i(y)\right\| \leq \frac{1}{m}\sum_{i=1}^m L_i\left\|x-y\right\|.$$

for any $x, y \in \mathbb{R}^d$, which suggests that $f(\cdot)$ is $\sum_{i=1}^m L_i/m$-smooth and it implies that $L \leq \sum_{i=1}^m L_i/m$. Since each $f_i(\cdot)$ is μ_i-strongly convex, which means

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{\mu_i}{2} \|y - x\|^2$$

for any $x, y \in \mathbb{R}^d$. Therefore, by summing up above inequalities for $i = 1, \ldots, m$, we have

$$\begin{aligned} f(y) =& \frac{1}{m}\sum_{i=1}^m f_i(y) \geq \frac{1}{m}\sum_{i=1}^m \left(f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{\mu_i}{2} \|y - x\|^2 \right) \\ =& f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\sum_{i=1}^m \frac{\mu_i}{m} \|y - x\|^2 . \end{aligned} \tag{11}$$

Therefore, we conclude $f(\cdot)$ is $\sum_{i=1}^m \mu_i/m$-strongly convex, which implies that $\mu \geq \sum_{i=1}^m \mu_i/m$. By Hölder's inequality, we have

$$\sum_{i=1}^m L_i = \sum_{i=1}^m \frac{L_i}{\mu_i} \cdot \mu_i \leq \max_{i \in [m]} \frac{L_i}{\mu_i} \cdot \sum_{i=1}^m \mu_i.$$

After rearrangement, we obtain

$$\kappa = \frac{L}{\mu} \leq \frac{\sum_{i=1}^m L_i}{\sum_{i=1}^m \mu_i} \leq \max_{i \in [m]} \frac{L_i}{\mu_i} = \kappa_{\max}.$$

For the inequality $\kappa_{\max} \leq \bar{\kappa}'_{\max}$, we can obtain that $L_i \leq \sum_{j=1}^n L_{i,j}/n$ for all $i \in [m]$ by the result of (10). Then the definitions of $\kappa_{\max}$ and $\bar{\kappa}'_{\max}$ leads to $\kappa_{\max} \leq \bar{\kappa}'_{\max}$.

For the inequality $\kappa \leq \bar{\kappa}$, we connect two inequalities in that (10) to obtain $L \leq 1/(mn)\sum_{i=1}^m \sum_{j=1}^n L_{i,j}$. Then the definitions of κ and $\bar{\kappa}$ leads to $\kappa \leq \bar{\kappa}$.

For the inequality $\bar{\kappa} \leq \bar{\kappa}_{\max}$, we directly follow the definitions of $\bar{\kappa}$ and $\bar{\kappa}_{\max}$ to achieve $\sum_{i=1}^m \sum_{j=1}^n L_{i,j}/(mn) \leq \max_{i \in [m]} \sum_{j=1}^n L_{i,j}/n$. This inequality is equivalent to $\bar{\kappa} \leq \bar{\kappa}_{\max}$.

For the inequality $\bar{\kappa}_{\max} \leq m\bar{\kappa}'_{\max}$, from (11) we know that f is $\sum_{i=1}^m \mu_i/m$-strongly convex. Thus we can obtain that $\mu_i \leq \sum_{i=1}^m \mu_i \leq m\mu$ for all $i \in [m]$, which results

$$\bar{\kappa}_{\max} = \frac{\max_{i \in [m]} \sum_{j=1}^n L_{i,j}}{n\mu} = \frac{\sum_{j=1}^n L_{i^*,j}}{n\mu} \leq \frac{\sum_{j=1}^n L_{i^*,j}}{n\mu_{i^*}/m} \leq \max_{i \in [m]} \frac{m\sum_{j=1}^n L_{i,j}}{n\mu_i} = m\bar{\kappa}'_{\max}.$$

where we define $i^* = \arg\max_{i \in [m]} \sum_{j=1}^n L_{i,j}$.

For the inequality $\bar{\kappa}_{\max} \leq mn\kappa$, it follows from the fact that

$$0 \leq f_{i,j}(y) - f_{i,j}(x) - \langle \nabla f_{i,j}(x), y - x \rangle \leq mn\left(f(y) - f(x) - \langle \nabla f(x), y - x \rangle\right) \leq mn \cdot \frac{L}{2} \|y - x\|^2$$

for all $x, y \in \mathbb{R}^d$ and $i \in [m]$, $j \in [n]$, where the first inequality is based on the fact that each $f_{i,j}(\cdot)$ is convex. This implies that each $f_{i,j}$ is mnL-smooth and $L_{i,j} \leq mnL$. Thus we have

$$\bar{\kappa}_{\max} = \max_{i \in [m]} \frac{\sum_{j=1}^n L_{i,j}}{n\mu} \leq \frac{mnL}{\mu} = mn\kappa.$$

Now, we have finished the proof.

□

A.2. The Example in Remark 2.4

In Remark 2.4, we have defined functions

$$f_{i,j}(x) = \frac{1}{2}x_{i,j}^\top H_{i,j}x_{i,j} + \frac{\mu_i}{2}\left\|x\right\|_2^2,$$

for $x = [x_1; \dots; x_m] \in \mathbb{R}^{2mn}$, $i \in [m]$ and $j \in [n]$, where $x_i = [x_{i,1}; \dots; x_{i,n}] \in \mathbb{R}^{2n}$, $x_{i,j} \in \mathbb{R}^2$ contains the $(2m(i-1)+2j-1)$-th and the $(2m(i-1)+2j)$-th coordinates of x, and $H_{i,j} = \mathrm{diag}(mn(L-\mu), 0) \in \mathbb{R}^{2\times 2}$ with $\mu_i = 2i\mu/(m+1)$ for some $L > 0$ and $\mu > 0$ such that $L/\mu \gg 1$.

Note that each $f_{i,j}(\cdot)$ is quadratic function, which means their Hessian are fixed. Therefore, we can calculate the condition numbers based on the Hessians of $f_{i,j}(\cdot)$, $f_i(\cdot)$ and $f(\cdot)$. We provide the details as follows.

The Hessian of $f_{i,j}(\cdot)$ has the form of

$$\nabla^2 f_{i,j}(\cdot) = \mathrm{diag}(\mu_i, \dots, \mu_i, mn(L-\mu)+\mu_i, \mu_i \dots, \mu_i),$$

where only the $2n(i-1)+2j-1$-th diagonal entry is $mn(L-\mu)+\mu_i$ and the others are μ_i. This implies $f_{i,j}(\cdot)$ is $L_{i,j}$ smooth with $L_{i,j} = mn(L-\mu)+\mu_i$.

Then we have

$$\nabla^2 f_i(\cdot) = \frac{1}{n}\sum_{j=1}^n \nabla^2 f_{i,j}(\cdot) = \mathrm{diag}(\mu_i, \dots, \mu_i, m(L-\mu)+\mu_i, \mu_i \dots, m(L-\mu)+\mu_i, \mu_i, \dots, \mu_i),$$

where only $(2n(i-1)+2j-1)$-th diagonal entry with $j \in [n]$ is $m(L-\mu)+\mu_i$ and the others are μ_i. This implies $f_i(\cdot)$ is L_i-smooth with $L_i = m(L-\mu)+\mu_i$ and μ_i-strongly convex.

We also have

$$\begin{aligned}\nabla^2 f(\cdot) =& \frac{1}{m}\sum_{i=1}^m \nabla^2 f_i(\cdot) = \mathrm{diag}\left((L-\mu) + \frac{1}{m}\sum_{i=1}^m \mu_i, \frac{1}{m}\sum_{i=1}^m \mu_i, \dots, (L-\mu) + \frac{1}{m}\sum_{i=1}^m \mu_i, \frac{1}{m}\sum_{i=1}^m \mu_i\right)\\ =& \,\mathrm{diag}\left(L, \mu, \dots, L, \mu\right),\end{aligned}$$

where the last step is based on the fact

$$\frac{1}{m}\sum_{i=1}^m \mu_i = \frac{1}{m}\sum_{i=1}^m \frac{2\mu i}{m+1} = \frac{2\mu}{m(m+1)}\sum_{i=1}^m i = \frac{2\mu}{m(m+1)}\cdot\frac{m(m+1)}{2} = \mu.$$

This implies $f(\cdot)$ is L-smooth and μ-strongly convex.

Noticing that the maximum of L_i/μ_i and $\sum_{j=1}^n L_{i,j}/\mu_i$ is obtained by $i=1$. Now we achieve the condition numbers

$$\begin{aligned}&\kappa = \frac{L}{\mu}, \quad \bar{\kappa} = \frac{mn(L-\mu)+\mu}{\mu} = \frac{mnL}{\mu} - mn + 1 = \Theta\left(\frac{mnL}{\mu}\right),\\ &\kappa_{\max} = \frac{L_1}{\mu_1} = \frac{mL}{\mu_1} = \frac{m(m+1)L}{2\mu} = \Theta\left(\frac{m^2 L}{\mu}\right),\\ &\bar{\kappa}_{\max} = \frac{mn(L-\mu)+\mu_m}{\mu} = \frac{mnL}{\mu} - mn + \frac{2m}{m+1} = \Theta\left(\frac{mnL}{\mu}\right),\\ &\bar{\kappa}'_{\max} = \frac{\sum_{j=1}^n L_{1,j}}{\mu_1} = \frac{m(m+1)nL}{2\mu} = \Theta\left(\frac{m^2 nL}{\mu}\right).\end{aligned}$$

B. The Proofs in Section 3.2.1

We first prove Lemma 3.2 by applying Lemma 3.1, then we prove Theorem 3.3 by using Lemma 3.2.

B.1. The Proof of Lemma 3.2

Proof. We first consider the random variable $Z=\sum_{j=1}^n X_j$, where $X_1,\dots,X_n$ are are mutually independent and each of X_j is distributed to $\mathrm{Bernoulli}(p_j)$ for some $p_j\in[0,1]$. We denote $\nu=\mathbb{E}[Z]=\sum_{j=1}^n p_j$ and $a\in\mathbb{R}$ be a constant such that $a\geq \mathrm{e}^2$. We consider the cases of Z as follows:

(a) If $\nu\geq\ln a$, we apply Lemma 3.1 for $t\geq 2\mathrm{e}\nu$, it holds that

$$\mathbb{P}(Z\geq t)\leq 2^{-t}.$$

This implies

$$\mathbb{P}(Z\geq 2\mathrm{e}\nu)\leq 2^{-2\mathrm{e}\nu}\leq 2^{-2\mathrm{e}\ln a}\leq\frac{1}{a^2}.$$

(b) If $1\leq\nu\leq\ln a$, it holds

$$\begin{aligned}\mathbb{P}\left(Z\geq t\ln a\right)\leq&\mathbb{P}\left(Z\geq t\nu\right)\overset{t\triangleq(1+\delta)}{=}\mathbb{P}\left(Z\geq(1+\delta)\nu\right)\\ &\overset{(8)}{\leq}\left(\frac{\mathrm{e}^\delta}{(1+\delta)^{1+\delta}}\right)^\nu=\left(\frac{\mathrm{e}^t}{\mathrm{e}t^t}\right)^\nu\leq\frac{1}{a^2},\end{aligned}$$

where we set $t\geq 2\mathrm{e}\ln a$. Thus, in this case, the extra term is of order $\ln a$ compared to case (a).

(c) If $\nu<1$, we let $X_1'\sim\mathrm{Bernoulli}(p_1')$ with $p_1'=1-\nu+\mathbb{E}[X_1]$ and $Z'=\sum_{j=2}^n X_j+X_1'$. It is clear that for any $t\geq 0$, it holds

$$\mathbb{P}(Z\geq t)\leq\mathbb{P}(Z'\geq t).$$

As we have $\mathbb{E}[Z']=1$, it holds that

$$\mathbb{P}\left(Z\geq t\ln a\right)\leq\mathbb{P}\left(Z'\geq t\ln a\right)\leq\frac{1}{a^2},$$

where the last inequality is based on case (b) by taking $t\geq 2\mathrm{e}\ln a$.

Since each Z_i follows the same distribution as Z, combining above three cases and Boole's inequality leads to

$$\begin{aligned}\mathbb{P}\left(\exists i\in[m],Z_i\geq\max\left\{2\mathrm{e}\mathbb{E}[Z_i],2\mathrm{e}(\ln a)^2\right\}\right)\leq&\sum_{i=1}^m\mathbb{P}\left(Z_i\geq\max\left\{2\mathrm{e}\mathbb{E}[Z_i],2\mathrm{e}(\ln a)^2\right\}\right)\\ =&\sum_{j=1}^m\min\left\{\mathbb{P}\left(Z_j\geq 2\mathrm{e}\mathbb{E}[Z_j]\right),\mathbb{P}\left(Z_j\geq 2\mathrm{e}(\ln a)^2\right)\right\}\\ \leq&\sum_{j=1}^m\frac{1}{a^2}=\frac{m}{a^2}.\end{aligned}$$

We complete the proof by taking $a=mn$. □

B.2. The Proof of Theorem 3.3

Proof. Recall that $Y_i^t=2\sum_{j=1}^n\xi_{i,j}^{t+1}$, where $\xi_{i,j}^{t+1}\sim\mathrm{Bernoulli}(q_{i,j})$. Then using Lemma 3.2, we obtain

$$\mathbb{P}\left(\max_{i\in[m]}Y_i^t\geq 4\mathrm{e}\max\left\{\max_{i\in[m]}\mathbb{E}[Y_i^t],(\ln mn)^2\right\}\right)\leq\frac{1}{mn}.$$

Since each Y_i is upper bounded by $2n$, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{i\in[m]} Y_i^t\right] \leq & 4\mathrm{e}\max\left\{\max_{i\in[m]} \mathbb{E}[Y_i^t], (\ln mn)^2\right\} + \frac{1}{mn}\cdot 2n \\
= & 4\mathrm{e}\max\left\{\max_{i\in[m]} \sum_{j=1}^n q_{i,j}, (\ln mn)^2\right\} + \frac{2}{m} \\
= & 4\mathrm{e}\max\left\{\max_{i\in[m]} \sum_{j=1}^n q_{i,j}, (\ln mn)^2\right\} + \frac{2}{m} \\
= & 4\mathrm{e}\max\left\{\sqrt{\frac{n\bar{\kappa}_{\max}}{m\kappa}}, (\ln mn)^2\right\} + \frac{2}{m} \\
= & \mathcal{O}\left(\sqrt{\frac{n\bar{\kappa}_{\max}}{m\kappa}} + (\ln mn)^2\right),
\end{aligned}$$

which concludes the proof. □

C. Proof of Section 3.2.2

In this section, we focus on analyzing Lyapunov function

$$V^t \triangleq \mathcal{Z}^t + \mathcal{Y}^t + \mathcal{W}^t,$$

where

$$\mathcal{Z}^t \triangleq \frac{L(1+\eta\sigma)}{2\eta}\left\|\bar{z}^t - x^*\right\|^2, \quad \mathcal{Y}^t \triangleq \frac{1}{\theta_1}(f(\bar{y}^t) - f(x^*)) \quad \text{and} \quad \mathcal{W}^t \triangleq \frac{\theta_2}{p\lambda\theta_1}(f(\bar{w}^t) - f(x^*)).$$

Our convergence analysis is more complicated than the counterpart of L-Katyusha (Kovalev et al., 2020a; Qian et al., 2021), since we have to address the additional consensus error aroused from the decentralized setting.

C.1. Proof of Lemma 3.4

We first provide some useful lemmas.

Lemma C.1 (Nesterov (2018))**.** *Under Assumption 2.1 and 2.3, it holds that*

$$\frac{1}{2L_{i,j}}\left\|\nabla f_{i,j}(x) - \nabla f_{i,j}(y)\right\|^2 \leq f_{i,j}(x) - f_{i,j}(y) - \langle \nabla f_{i,j}(y), x - y\rangle, \tag{12}$$

for all $i \in [m], j \in [m]$ and $x, y \in \mathbb{R}^d$.

Lemma C.2. *Given n independent random vectors $x_1, \dots, x_n$ such that $\mathbb{E}[x_i] = 0$ for all $i \in [n]$, then it holds*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n x_i\right\|^2\right] = \sum_{i=1}^n \mathbb{E}\left[\|x_i\|^2\right] \tag{13}$$

Proof. It holds that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^n \mathbf{x}_i\right\|^2\right] &= \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{x}_i\|^2\right] + \sum_{i\neq j}\mathbb{E}\left[\langle \mathbf{x}_i, \mathbf{x}_j\rangle\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{x}_i\|^2\right],
\end{aligned}$$

where the last equality is because that $x_1, \dots, x_n$ are independent and $\mathbb{E}[x_i] = 0$. □

Lemma C.3. *The vectors $\bar{s}^t$, $\bar{u}^t$ and $\bar{g}^t$ in Algorithm 2 satisfy*

$$\bar{s}^t = \frac{1}{m}\sum_{i}^{m} v_i^t \qquad \textit{with} \qquad \mathbb{E}_{\xi_{i,j}^{t+1}}[\bar{s}^t] = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_i^t) \tag{14}$$

and

$$\bar{u}^t = \bar{g}^t = \frac{1}{m}\sum_{i}^{m} \nabla f_i(w_i^t). \tag{15}$$

Proof. Applying Proposition 2.6 to the update rules of Algorithm 2 directly finishes the the proof. □

Lemma C.4. *Under the settings of Lemma 3.4, Algorithm 2 holds that*

$$\left\|\nabla f(\bar{x}^t) - \mathbb{E}_{\xi_{i,j}^{t+1}}[\bar{s}^t]\right\| \leq \frac{\max_{i\in[m]} L_i}{\sqrt{m}} \left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|, \tag{16}$$

where L_i is the smoothness parameter of $f_i(\cdot)$.

Proof. We have

$$\begin{aligned}
\left\|\nabla f(\bar{x}^t) - \mathbb{E}_{\xi_{i,j}^{t+1}}[\bar{s}^t]\right\|^2 =& \left\|\frac{1}{m}\sum_{i=1}^{m}\left(\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\bar{x}^t)\right)\right\|^2 \\
\leq& \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla f_i(x_i^t) - \nabla f_i(\bar{x}^t)\right\|^2 \\
\leq& \frac{1}{m}\sum_{i=1}^{m} L_i^2 \left\|x_i^t - \bar{x}^t\right\|^2 \\
=& \frac{\max_{i\in[m]} L_i^2}{m}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2,
\end{aligned}$$

where the first equality is based on Lemma C.3. □

Then we provide some lemmas for the mean vectors. The following Lemma C.5 is important to our analysis, which guarantees the appropriate smoothness dependence for sample complexity.

Lemma C.5. *Under the settings of Lemma 3.4, it holds that*

$$\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\bar{s}^t - \nabla f(\bar{x}^t)\right\|^2\right] \leq \frac{12\bar{L}_{\max}}{b}\left(f(\bar{w}^t) - f(\bar{x}^t) - \langle\nabla f(\bar{x}^t), \bar{w}^t - \bar{x}^t\rangle\right) + C_1\cdot\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + C_2\cdot\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2, \tag{17}$$

where

$$C_1 = \frac{12\bar{L}\hat{L}}{b} + \frac{2\hat{L}^2}{m}, \qquad C_2 = \frac{12\bar{L}\hat{L}}{b} \qquad \textit{and} \qquad \hat{L} = \max_{i\in[m], j\in[n]} L_{i,j}. \tag{18}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\bar{s}^t - \nabla f(\bar{x}^t)^2\right\|^2\right] \\
=&\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\bar{u}^t + \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t) - \nabla f_{i,j}(w_i^t)\right) - \nabla f(\bar{x}^t)\right\|^2\right] \\
\overset{(15)}{=}&\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(w_i^t) + \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\xi_{i,j}^{t+1}}{q_{i,j}}\left(\nabla f_{i,j}(x_i^t) - \nabla f_{i,j}(w_i^t)\right) - \nabla f(\bar{x}^t)\right\|^2\right]
\end{aligned}$$

$$\overset{(14)}{=}\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\xi_{i,j}^{t+1}}{q_{i,j}}(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t))-(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t))\right]+(\mathbb{E}_{\xi_{i,j}^{t+1}}[\bar{s}^t]-\nabla f(\bar{x}^t))\right\|^2\right]$$

$$\leq 2\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\xi_{i,j}^{t+1}}{q_{i,j}}(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t))-(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t)\right]\right\|^2\right]+2\left\|\mathbb{E}_{\xi_{i,j}^{t+1}}[\bar{s}^t]-\nabla f(\bar{x}^t)\right\|^2,$$

where the inequality is based on Young's inequality. Then we consider the two terms on the right hand side separately. First, based on Lemma C.4, we have

$$2\left\|\mathbb{E}_{\xi_{i,j}^{t+1}}[\bar{s}^t]-\nabla f(\bar{x}^t))\right\|^2\overset{(16)}{\leq}\frac{2\max_{i\in[m]}L_i^2}{m}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2\leq\frac{2\hat{L}^2}{m}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2.$$

Then we consider the first term.

$$\begin{aligned}
&2\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\xi_{i,j}^{t+1}}{q_{i,j}}(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t))-(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t)\right]\right\|^2\right]\\
\overset{(13)}{=}&\frac{2}{m^2n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\frac{\xi_{i,j}^{t+1}}{q_{i,j}}(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t))-(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t))\right\|^2\right]\\
=&\frac{2}{m^2n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1-q_{i,j}}{q_{i,j}}\left\|\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t)\right\|^2\\
\leq&\frac{6}{m^2n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1-q_{i,j}}{q_{i,j}}\left[\left\|\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(\bar{x}^t)\right\|^2+\left\|\nabla f_{i,j}(\bar{x}^t)-\nabla f_{i,j}(\bar{w}^t)\right\|^2+\left\|\nabla f_{i,j}(\bar{w}^t)-\nabla f_{i,j}(w_i^t)\right\|^2\right]\\
\overset{(12)}{\leq}&\frac{12}{m^2n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{1-q_{i,j}}{q_{i,j}}L_{i,j}\left(f_{i,j}(\bar{w}^t)-f_{i,j}(\bar{x}^t)-\left\langle\nabla f_{i,j}(\bar{x}^t),\bar{w}^t-\bar{x}^t\right\rangle\right)+L_{i,j}^2\left(\left\|x_i^t-\bar{x}^t\right\|^2+\left\|w_i^t-\bar{w}^t\right\|^2\right)\right)\\
\leq&\frac{12}{m^2n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1-q_{i,j}}{q_{i,j}}L_{i,j}\left(f_{i,j}(\bar{w}^t)-f_{i,j}(\bar{x}^t)-\left\langle\nabla f_{i,j}(\bar{x}^t),\bar{w}^t-\bar{x}^t\right\rangle\right)+C_2\cdot\left(\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2+\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2\right)\\
\leq&\frac{12}{mnb}\sum_{i=1}^{m}\sum_{j=1}^{n}(1-q_{i,j})\bar{L}\left(f_{i,j}(\bar{w}^t)-f_{i,j}(\bar{x}^t)-\left\langle\nabla f_{i,j}(\bar{x}^t),\bar{w}^t-\bar{x}^t\right\rangle\right)+C_2\cdot\left(\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2+\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2\right)\\
\leq&\frac{12\bar{L}_{\max}}{b}\left(f(\bar{w}^t)-f(\bar{x}^t)-\left\langle\nabla f(\bar{x}^t),\bar{w}^t-\bar{x}^t\right\rangle\right)+C_2\cdot\left(\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2+\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2\right),
\end{aligned}$$

where the second inequality is based on the fact that $\hat{L}\geq L_{i,j}$ for all $i\in[m]$ and $j\in[n]$ and the second last inequality holds because of we take $q_{i,j}=\min\{1,bL_{i,j}/(mn\bar{L}_{\max})\}$. Other inequalities are based on Young's inequality. Then by combining the three inequalities we have

$$\begin{aligned}
\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\bar{s}^t-\nabla f(\bar{x}^t)^2\right\|^2\right]\leq&\frac{12\bar{L}_{\max}}{b}\left(f(\bar{w}^t)-f(\bar{x}^t)-\left\langle\nabla f(\bar{x}^t),\bar{w}^t-\bar{x}^t\right\rangle\right)+C_2\cdot\left(\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2+\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2\right)\\
&+\frac{2\hat{L}^2}{m}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2,
\end{aligned}$$

which concludes the proof. □

Next, we provide some lemmas by following the analysis on non-distributed methods (Kovalev et al., 2020a; Qian et al., 2021). For the completeness, we also give their detailed proofs.

Lemma C.6. *Under the settings of Lemma 3.4, we have*

$$\left\langle\bar{s}^t,x^*-\bar{z}^{t+1}\right\rangle+\frac{\mu}{2}\left\|\bar{x}^t-x^*\right\|^2\geq\frac{L}{2\eta}\left\|\bar{z}^t-\bar{z}^{t+1}\right\|^2+\mathcal{Z}^{t+1}-\frac{1}{1+\eta\sigma}\mathcal{Z}^t.\tag{19}$$

Proof. Based on the definition of $\mathbf{z}^{t+1}$, we have

$$\mathbf{z}^{t+1} \overset{\text{Alg. 2}}{=} \frac{1}{1+\eta\sigma}\left(\eta\sigma x^t + \mathbf{z}^t - \frac{\eta}{L}\mathbf{s}^t\right),$$

which means

$$\frac{\eta}{L}\bar{s}^t = \eta\sigma(\bar{x}^t - \bar{z}^{t+1}) + (\bar{z}^t - \bar{z}^{t+1}).$$

It further implies that

$$\begin{aligned}\left\langle \bar{s}^t, \bar{z}^{t+1} - x^* \right\rangle =& \mu \left\langle \bar{x}^t - \bar{z}^{t+1}, \bar{z}^{t+1} - x^* \right\rangle + \frac{L}{\eta}\left\langle \bar{z}^t - \bar{z}^{t+1}, \bar{z}^{t+1} - x^* \right\rangle \\ =& \frac{\mu}{2}\left(\left\|\bar{x}^t - x^*\right\|^2 - \left\|\bar{x}^t - \bar{z}^{t+1}\right\|^2 - \left\|\bar{z}^{t+1} - x^*\right\|^2\right) \\ &+ \frac{L}{2\eta}\left(\left\|\bar{z}^t - x^*\right\|^2 - \left\|\bar{z}^t - \bar{z}^{t+1}\right\|^2 - \left\|\bar{z}^{t+1} - x^*\right\|^2\right) \\ \leq& \frac{\mu}{2}\left\|\bar{x}^t - x^*\right\|^2 + \frac{L}{2\eta}\left(\left\|\bar{z}^t - x^*\right\|^2 - (1+\eta\sigma)\left\|\bar{z}^{t+1} - x^*\right\|^2\right) - \frac{L}{2\eta}\left\|\bar{z}^t - \bar{z}^{t+1}\right\|^2,\end{aligned}$$

which concludes the proof. □

Lemma C.7. *Under the settings of Lemma 3.4, we have*

$$\frac{1}{\theta_1}\left(f(\bar{y}^{t+1}) - f(\bar{x}^t)\right) - \frac{1}{24L\theta_1}\left\|\bar{s}^t - \nabla f(\bar{x}^t)\right\|^2 \leq \frac{L}{2\eta}\left\|\bar{z}^{t+1} - \bar{z}^t\right\|^2 + \left\langle \bar{s}^t, \bar{z}^{t+1} - \bar{z}^t \right\rangle. \tag{20}$$

Proof. We have

$$\begin{aligned}&\frac{L}{2\eta}\left\|\bar{z}^{t+1} - \bar{z}^t\right\|^2 + \langle \bar{s}^t, \bar{z}^{t+1} - \bar{z}^t\rangle \\ =& \frac{1}{\theta_1}\left(\frac{L}{2\eta\theta_1}\left\|\theta_1(\bar{z}^{t+1} - \bar{z}^t)\right\|^2 + \langle \bar{s}^t, \theta_1(\bar{z}^{t+1} - \bar{z}^t)\rangle\right) \\ =& \frac{1}{\theta_1}\left(\frac{L}{2\eta\theta_1}\left\|\bar{y}^{t+1} - \bar{x}^t\right\|^2 + \left\langle \bar{s}^t, \bar{y}^{t+1} - \bar{x}^t\right\rangle\right) \\ =& \frac{1}{\theta_1}\left(\frac{L}{2\eta\theta_1}\left\|\bar{y}^{t+1} - \bar{x}^t\right\|^2 + \left\langle \nabla f(\bar{x}^t), \bar{y}^{t+1} - \bar{x}^t\right\rangle + \left\langle \bar{s}^t - \nabla f(\bar{x}^t), \bar{y}^{t+1} - \bar{x}^t\right\rangle\right) \\ =& \frac{1}{\theta_1}\left(\frac{L}{2}\left\|\bar{y}^{t+1} - \bar{x}^t\right\|^2 + \left\langle \nabla f(\bar{x}^t), \bar{y}^{t+1} - \bar{x}^t\right\rangle + \frac{L}{2}\left(\frac{1}{\eta\theta_1} - 1\right)\left\|\bar{y}^{t+1} - \bar{x}^t\right\|^2 + \left\langle \bar{s}^t - \nabla f(\bar{x}^t), \bar{y}^{t+1} - \bar{x}^t\right\rangle\right) \\ \geq& \frac{1}{\theta_1}\left(f(\bar{y}^{t+1}) - f(\bar{x}^t) + \frac{L}{2}\left(\frac{1}{\eta\theta_1} - 1\right)\left\|\bar{y}^{t+1} - \bar{x}^t\right\|^2 + \left\langle \bar{s}^t - \nabla f(\bar{x}^t), \bar{y}^{t+1} - \bar{x}^t\right\rangle\right) \\ \geq& \frac{1}{\theta_1}\left(f(\bar{y}^{t+1}) - f(\bar{x}^t) - \frac{\eta\theta_1}{2L(1-\eta\theta_1)}\left\|\bar{s}^t - \nabla f(\bar{x}^t)\right\|^2\right) \\ =& \frac{1}{\theta_1}\left(f(\bar{y}^{t+1}) - f(\bar{x}^t) - \frac{1}{24L}\left\|\bar{s}^t - \nabla f(\bar{x}^t)\right\|^2\right),\end{aligned}$$

where the first inequality is because of Assumption 2.1, the last inequality uses Young's inequality in the form of

$$\langle a, b\rangle \geq -\frac{\|a\|^2}{2\beta} - \frac{\beta\|b\|^2}{2} \qquad \text{with } \beta = \frac{\eta\theta_1}{L(1-\eta\theta_1)}$$

and the last equality is because of the setting $\eta = 1/(13\theta_1)$. □

Lemma C.8. *Under the settings of Lemma 3.4, we have*

$$\mathbb{E}\left[\mathcal{W}^{t+1}\right] = (1-p)\mathcal{W}^t + \frac{\theta_2}{\lambda}\mathcal{Y}^t.$$

Proof. From Algorithm 2, we know that

$$\mathbb{E}\left[f(\bar{w}^{t+1})\right] = (1-p)f(\bar{w}^t) + pf(\bar{y}^t).$$

Then from the definition of $\mathcal{W}^t$ and $\mathcal{Y}^t$, we directly finish the proof. □

Using the above lemmas, we prove Lemma 3.4 as follows.

Proof of Lemma 3.4. Combining Lemma C.4, C.5, C.6, C.7 and C.8, we obtain

$$\begin{aligned}
f(x^*) \geq & f(\bar{x}^t) + \left\langle \nabla f(\bar{x}^t), x^* - \bar{x}^t \right\rangle + \frac{\mu}{2}\left\|\bar{x}^t - x^*\right\|^2 \\
= & f(\bar{x}^t) + \frac{\mu}{2}\left\|\bar{x}^t - x^*\right\|^2 + \left\langle \nabla f(\bar{x}^t), x^* - \bar{z}^t + \bar{z}^t - \bar{x}^t \right\rangle \\
= & f(\bar{x}^t) + \frac{\mu}{2}\left\|\bar{x}^t - x^*\right\|^2 + \left\langle \nabla f(\bar{x}^t), x^* - \bar{z}^t \right\rangle + \frac{\theta_2}{\theta_1}\left\langle \nabla f(\bar{x}^t), \bar{x}^t - \bar{w}^t \right\rangle \\
& + \frac{(1-\theta_1-\theta_2)}{\theta_1}\left\langle \nabla f(\bar{x}^t), \bar{x}^t - \bar{y}^t \right\rangle \\
= & f(\bar{x}^t) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(\bar{x}^t), \bar{x}^t - \bar{w}^t \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(\bar{x}^t) - f(\bar{y}^t)) \\
& + \mathbb{E}\left[\frac{\mu}{2}\left\|\bar{x}^t - x^*\right\|^2 + \left\langle \bar{s}^t, x^* - \bar{z}^{t+1} \right\rangle + \left\langle \bar{s}^t, \bar{z}^{t+1} - \bar{z}^t \right\rangle\right] + \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle \\
\overset{(19)}{\geq} & f(\bar{x}^t) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(\bar{x}^t), \bar{x}^t - \bar{w}^t \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(\bar{x}^t) - f(\bar{y}^t)) \\
& + \mathbb{E}\left[\mathcal{Z}^{t+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^t\right] + \mathbb{E}\left[\left\langle \bar{s}^t, \bar{z}^{t+1} - \bar{z}^t \right\rangle + \frac{L}{2\eta}\left\|\bar{z}^t - \bar{z}^{t+1}\right\|^2\right] \\
& + \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle \\
\overset{(20)}{\geq} & f(\bar{x}^t) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(\bar{x}^t), \bar{x}^t - \bar{w}^t \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(\bar{x}^t) - f(\bar{y}^t)) \\
& + \mathbb{E}\left[\mathcal{Z}^{t+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^t\right] + \mathbb{E}\left[\frac{1}{\theta_1}\left(f(\bar{y}^{t+1}) - f(\bar{x}^t)\right) - \frac{1}{24L\theta_1}\left\|\bar{s}^t - \nabla f(\bar{x}^t)\right\|^2\right] \\
& + \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle \\
\overset{(17)}{\geq} & f(\bar{x}^t) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(\bar{x}^t), \bar{x}^t - \bar{w}^t \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(\bar{x}^t) - f(\bar{y}^t)) + \mathbb{E}\left[\mathcal{Z}^{t+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^t\right] \\
& + \mathbb{E}\left[\frac{1}{\theta_1}\left(f(\bar{y}^{t+1}) - f(\bar{x}^t)\right) - \frac{\theta_2}{\theta_1}\left(f(\bar{w}^t) - f(\bar{x}^t) - \left\langle \nabla f(\bar{x}^t), \bar{w}^t - \bar{x}^t \right\rangle\right)\right] \\
& + \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle - \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 - \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
= & f(\bar{x}^t) + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(\bar{x}^t) - f(\bar{y}^t)) - \frac{1}{1+\eta\sigma}\mathcal{Z}^t - \frac{\theta_2}{\theta_1}(f(\bar{w}^t) - f(\bar{x}^t)) \\
& + \mathbb{E}\left[\mathcal{Z}^{t+1} + \frac{1}{\theta_1}\left(f(\bar{y}^{t+1}) - f(\bar{x}^t)\right)\right] \\
& + \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle - \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 - \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2,
\end{aligned}$$

where the first inequality is because of Assumption 2.3, the second inequality uses the convexity of $f(\cdot)$, and the last inequality is obtained by $\theta_2 = \bar{L}_{\max}/(2Lb)$. Note that the notations of C_1 and C_2 follow the definitions in (18). Furthermore, the procedure of Algorithm 2 and Proposition 2.6 implies

$$\bar{x}^t = \theta_1\bar{z}^t + \theta_2\bar{w}^t + (1-\theta_1-\theta_2)\bar{y}^t \quad \text{and} \quad \bar{z}^t - \bar{x}^t = \frac{\theta_2}{\theta_1}(\bar{x}^t - \bar{w}^t) + \frac{1-\theta_1-\theta_2}{\theta_1}(\bar{x}^t - \bar{y}^t).$$

Combining all the above results, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{Z}^{t+1}+\mathcal{Y}^{t+1}\right] \leq & \frac{1}{1+\eta\sigma}\mathcal{Z}^t + (1-\theta_1-\theta_2)\mathcal{Y}^t + \frac{\theta_2}{\theta_1}(f(\bar{w}^t)-f^*) \\
& - \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle + \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2.
\end{aligned}
$$

Based on the definition of $\mathcal{W}^t$, we achieve

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{Z}^{t+1}+\mathcal{Y}^{t+1}\right] \leq & \frac{1}{1+\eta\sigma}\mathcal{Z}^t + (1-\theta_1-\theta_2)\mathcal{Y}^t + p\lambda\mathcal{W}^t \\
& + \left\|\nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t]\right\| \left\|x^* - \bar{z}^t\right\| + \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2.
\end{aligned}
$$

Finally, we use Lemma C.8 to achieve

$$
\begin{aligned}
& \mathbb{E}[V^{t+1}] = \mathbb{E}\left[\mathcal{Z}^{t+1}+\mathcal{Y}^{t+1}+\mathcal{W}^{t+1}\right] \\
\leq & \frac{1}{1+\eta\sigma}\mathcal{Z}^t + (1-\theta_1-\theta_2)\mathcal{Y}^t + p\lambda\mathcal{W}^t + (1-p)\mathcal{W}^t + \frac{\theta_2}{\lambda}\mathcal{Y}^t \\
& + \left\|\nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t]\right\| \left\|x^* - \bar{z}^t\right\| + \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
= & \frac{1}{1+\eta\sigma}\mathcal{Z}^t + \left(1 - \left(\theta_1+\theta_2-\frac{\theta_2}{\lambda}\right)\right)\mathcal{Y}^t + \left(1-p(1-\lambda)\right)\mathcal{W}^t \\
& + \left\|\nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t]\right\| \left\|x^* - \bar{z}^t\right\| + \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
\leq & \frac{1}{1+\eta\sigma}\mathcal{Z}^t + \left(1 - \left(\theta_1+\theta_2-\frac{\theta_2}{\lambda}\right)\right)\mathcal{Y}^t + \left(1-p(1-\lambda)\right)\mathcal{W}^t \\
& + \left\|\nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t]\right\| \left\|x^* - \bar{z}^t\right\| + \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
\overset{(16)}{\leq} & \underbrace{\max\left\{\frac{1}{1+\eta\sigma}, \left(1 - \left(\theta_1+\theta_2-\frac{\theta_2}{\lambda}\right)\right), \left(1-p(1-\lambda)\right)\right\}}_{\beta} V^t + \sqrt{\frac{2\eta\hat{L}^2V^t}{(1+\eta\sigma)mL}}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\| \\
& + \left(\frac{12L\hat{L}}{b} + \frac{2\hat{L}^2}{b}\right)\cdot\left[\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2\right],
\end{aligned}
$$

where the last inequality is obtained by the definition of V_t. □

C.2. Proof of Lemma 3.5

The main idea for analyzing the consensus error is establishing the recursion for

$$
r^t = \frac{L}{m}\left[\frac{\eta^2}{L^2}\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2, \frac{\eta^2}{L^2}\left\|\mathbf{s}^t - \mathbf{1}\bar{s}^t\right\|, \left\|\mathbf{z}^t - \mathbf{1}\bar{z}^t\right\|^2, \left\|\mathbf{y}^t - \mathbf{1}\bar{y}^t\right\|^2\right]^\top,
$$

There are actually extra consensus error terms $\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\|^2$ and $\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\|^2$. These terms can be bounded by the recursion of r^t and we present the process in proof of Theorem 3.7. Recall that we have defined

$$
\hat{L} \triangleq \max_{i\in[m], j\in[n]} L_{i,j},
$$

which satisfies

$$
mnL \geq \hat{L} \geq \bar{L}_{\max} \geq L. \tag{21}
$$

We first provide lemmas for the proof of Lemma 3.5.

Lemma C.9. *Under the notations and settings of Lemma 3.5, we have*

$$\begin{aligned}
\left\|\mathbf{x}^{t+1}-\mathbf{1}\bar{x}^{t+1}\right\|^2 \leq & 3\theta_1^2\left\|\mathbf{z}^{t+1}-\mathbf{1}\bar{z}^{t+1}\right\|^2+3\theta_2^2\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2+3(1-\theta_1-\theta_2)^2\left\|\mathbf{y}^{t+1}-\mathbf{1}\bar{y}^{t+1}\right\|^2, \\
\left\|\mathbf{u}^{t+1}-\mathbf{1}\bar{u}^{t+1}\right\|^2 \leq & 2\rho^2\left\|\mathbf{u}^{t}-\mathbf{1}\bar{u}^{t}\right\|^2+2\rho^2\left\|\mathbf{g}^{t+1}-\mathbf{g}^{t}\right\|^2, \\
\left\|\mathbf{s}^{t+1}-\mathbf{1}\bar{s}^{t+1}\right\|^2 \leq & 2\rho^2\left\|\mathbf{s}^{t}-\mathbf{1}\bar{s}^{t}\right\|^2+2\rho^2\left\|\mathbf{v}^{t+1}-\mathbf{v}^{t}\right\|^2, \\
\left\|\mathbf{z}^{t+1}-\mathbf{1}\bar{z}^{t+1}\right\|^2 \leq & \frac{3\rho^2\eta^2\sigma^2}{(1+\eta\sigma)^2}\left\|\mathbf{x}^{t}-\mathbf{1}\bar{x}^{t}\right\|^2+3\rho^2\frac{1}{(1+\eta\sigma)^2}\left\|\mathbf{z}^{t}-\mathbf{1}\bar{z}^{t}\right\|^2+\frac{3\rho^2\eta^2}{(1+\eta\sigma)^2L^2}\left\|\mathbf{s}^{t}-\mathbf{1}\bar{s}^{t}\right\|^2, \\
\left\|\mathbf{y}^{t+1}-\mathbf{1}\bar{y}^{t+1}\right\|^2 \leq & 3\rho^2\left\|\mathbf{x}^{t}-\mathbf{1}\bar{x}^{t}\right\|^2+3\rho^2\theta_1^2\left\|\mathbf{z}^{t+1}-\mathbf{1}\bar{z}^{t+1}\right\|^2+3\rho^2\theta_1^2\left\|\mathbf{z}^{t}-\mathbf{1}\bar{z}^{t}\right\|^2.
\end{aligned}$$

Proof. The upper bounds of $\left\|\mathbf{x}^{t+1}-\mathbf{1}\bar{x}^{t+1}\right\|$, $\left\|\mathbf{z}^{t+1}-\mathbf{1}\bar{z}^{t+1}\right\|$ and $\left\|\mathbf{y}^{t+1}-\mathbf{1}\bar{y}^{t+1}\right\|$ hold by combining Young's inequality and Proposition 2.6. The upper bounds of $\left\|\mathbf{u}^{t+1}-\mathbf{1}\bar{u}^{t+1}\right\|$ and $\left\|\mathbf{s}^{t+1}-\mathbf{1}\bar{s}^{t+1}\right\|$ can be obtained in the same way. We only provide the details for $\left\|\mathbf{u}^{t+1}-\mathbf{1}\bar{u}^{t+1}\right\|$ as follows. We have

$$\begin{aligned}
\left\|\mathbf{u}^{t+1}-\bar{u}^{t+1}\right\|^2 = & \sum_{i=1}^m\left\|u_i^{t+1}-\bar{u}^{t+1}\right\|^2 \\
\leq & \rho^2\sum_{i=1}^m\left\|(u_i^t-\bar{u}^t)+\left(g_i^{t+1}-g_i^t-\left(\bar{g}^{t+1}-\bar{g}^t\right)\right)\right\|^2 \\
\leq & 2\rho^2\sum_{i=1}^m\left\|u_i^t-\bar{u}^t\right\|^2+2\rho^2\sum_{i=1}^m\left\|\left(g_i^{t+1}-g_i^t-\left(\bar{g}^{t+1}-\bar{g}^t\right)\right)\right\|^2 \\
\leq & 2\rho^2\sum_{i=1}^m\left\|u_i^t-\bar{u}^t\right\|^2+2\rho^2\sum_{i=1}^m\left\|g_i^{t+1}-g_i^t\right\|^2 \\
= & 2\rho^2\left\|\mathbf{u}^t-\mathbf{1}\bar{u}^t\right\|^2+2\rho^2\left\|\mathbf{g}^{t+1}-\mathbf{g}^t\right\|^2,
\end{aligned}$$

where the first inequality is because of Proposition 2.6 and the last inequality is based on the fact

$$\sum_{i=1}^m\|a_i-\bar{a}\|_2^2\leq\sum_{i=1}^m\|a_i\|_2^2.$$

for $\{a_i\in\mathbb{R}^{1\times d}\}_{i=1}^m$ and $\bar{a}=\sum_{i=1}^m a_i/m$. □

Lemma C.10 (Expected consensus error of $\mathbf{w}^t$)**.** *Under the settings of Lemma 3.5, it holds that*

$$\mathbb{E}\left[\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2\right]=\sum_{s=1}^t p(1-p)^{t-s}\mathbb{E}\left[\|\mathbf{y}^s-\mathbf{1}\bar{y}^s\|^2\right]. \tag{22}$$

Proof. At the t-th iteration where $t=0,\dots,$ we have

$$\mathbb{E}_{\zeta^t}\left[\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2\right]=(1-p)\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2+p\left\|\mathbf{y}^t-\mathbf{1}\bar{y}^t\right\|^2.$$

We can obtain from this recursion that

$$\mathbb{E}\left[\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2\right]=\sum_{s=0}^t p(1-p)^{t-s}\mathbb{E}\left[\|\mathbf{y}^s-\mathbf{1}\bar{y}^s\|^2\right]+(1-p)^{t+1}\left\|\mathbf{w}^0-\mathbf{1}\bar{w}^0\right\|^2.$$

Then as initialization implies $\left\|\mathbf{w}^0-\mathbf{1}\bar{w}^0\right\|^2=\left\|\mathbf{y}^0-\mathbf{1}\bar{y}^0\right\|^2=0$, this leads to the desired result. □

Lemma C.11. *Under the settings of Lemma 3.5, it holds that*

$$\left\|\mathbf{g}^{t+1}-\mathbf{g}^t\right\|^2\leq 4\hat{L}^2\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2+\frac{8\hat{L}mp\lambda\theta_1}{\theta_2}\mathcal{W}^{t+1}+4\hat{L}^2\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2+\frac{8\hat{L}mp\lambda\theta_1}{\theta_2}\mathcal{W}^t.$$

Proof. We have

$$\left\|\mathbf{g}^{t+1} - \mathbf{g}^t\right\|^2 = \sum_{i=1}^m \left\|g_i^{t+1} - g_i^t\right\|^2 \le 2\sum_{i=1}^m \left\|\nabla f_i(w_i^{t+1}) - \nabla f_i(x^*)\right\|^2 + 2\sum_{i=1}^m \left\|\nabla f_i(w_i^t) - \nabla f_i(x^*)\right\|^2.$$

We also have

$$\begin{aligned}\left\|\nabla f_i(w_i^t) - \nabla f_i(x^*)\right\|^2 &= \left\|\nabla f_i(w_i^t) - \nabla f_i(\bar{w}^t) + \nabla f_i(\bar{w}^t) - \nabla f_i(x^*)\right\|^2 \\ &\le 2\left\|\nabla f_i(w_i^t) - \nabla f_i(\bar{w}^t)\right\|^2 + 2\left\|\nabla f_i(\bar{w}^t) - \nabla f_i(x^*)\right\|^2 \\ &\le 2\hat{L}^2\left\|w_i^t - \bar{w}^t\right\|^2 + 2\left\|\nabla f_i(\bar{w}^t) - \nabla f_i(x^*)\right\|^2 \\ &\overset{(12)}{\le} 2\hat{L}^2\left\|w_i^t - \bar{w}^t\right\|^2 + 4\hat{L}(f_i(\bar{w}^t) - f_i(x^*)).\end{aligned}$$

Combining the above results, we achieve

$$\begin{aligned}\left\|\mathbf{g}^{t+1} - \mathbf{g}^t\right\|^2 \le& 2\sum_{i=1}^m \left\|\nabla f_i(w_i^{t+1}) - \nabla f_i(x^*)\right\|^2 + 2\sum_{i=1}^m \left\|\nabla f_i(w_i^t) - \nabla f_i(x^*)\right\|^2 \\ \le& \sum_{i=1}^m 4\hat{L}^2\left\|w_i^{t+1} - \bar{w}^{t+1}\right\|^2 + 8\hat{L}(f_i(\bar{w}^{t+1}) - f_i(x^*)) + \sum_{i=1}^m 4\hat{L}^2\left\|w_i^t - \bar{w}^t\right\|^2 + 8\hat{L}(f_i(\bar{w}^t) - f_i(x^*)) \\ =& 4\hat{L}^2\left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2 + 8\hat{L}m(f(\bar{w}^{t+1}) - f(x^*)) + 4\hat{L}^2\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 + 8\hat{L}m(f(\bar{w}^t) - f(x^*)) \\ =& 4\hat{L}^2\left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2 + \frac{8\hat{L}mp\lambda\theta_1}{\theta_2}\mathcal{W}^{t+1} + 4\hat{L}^2\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 + \frac{8\hat{L}mp\lambda\theta_1}{\theta_2}\mathcal{W}^t.\end{aligned}$$

□

Next, we target to bound $\left\|\mathbf{v}^t - \mathbf{1}\bar{v}^t\right\|^2$, which is decomposed into two parts.

Lemma C.12. *Under the settings of Lemma 3.5, it holds that*

$$\left\|\mathbf{u}^t\right\|^2 = \sum_{i=1}^m \left\|u_i^t\right\|^2 \le 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + 3\hat{L}^2\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 + \frac{6\hat{L}mp\lambda\theta_1}{\theta_2}\mathcal{W}^t.$$

Proof. We have

$$\begin{aligned}\sum_{i=1}^m \left\|u_i^t\right\|^2 &= \sum_{i=1}^m \left\|(u_i^t - \bar{u}^t) + (\bar{u}^t - \nabla f(\bar{w}^t)) + (\nabla f(\bar{w}^t) - \nabla f(x^*))\right\|^2 \\ &\le \sum_{i=1}^m \left(3\left\|u_i^t - \bar{u}^t\right\|^2 + 3\left\|\bar{u}^t - \nabla f(\bar{w}^t)\right\|^2 + 3\left\|\nabla f(\bar{w}^t) - \nabla f(x^*)\right\|^2\right) \\ &\overset{(12)}{\le} 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + 3\sum_{i=1}^m \left\|\frac{1}{m}\sum_{j=1}^m(\nabla f_j(w_j^t) - \nabla f_j(\bar{w}^t))\right\|^2 + 3\sum_{i=1}^m(2\hat{L}(f_i(\bar{w}^t) - f_i(x^*)) \\ &\le 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{3}{m}\sum_{i=1}^m\sum_{j=1}^m \left\|\nabla f_j(w_j^t) - \nabla f_j(\bar{w}^t)\right\|^2 + 3\sum_{i=1}^m(2L(f_i(\bar{w}^t) - f_i(x^*)) \\ &\overset{(21)}{\le} 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{3\hat{L}^2}{m}\sum_{i=1}^m\sum_{j=1}^m \left\|w_j^t - \bar{w}^t\right\|^2 + 3\sum_{i=1}^m(2L(f(\bar{w}^t) - f_i(x^*)) \\ &= 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + 3\hat{L}^2\left\|w^t - \mathbf{1}\bar{w}^t\right\|^2 + \frac{6\hat{L}mp\lambda\theta_1}{\theta_2}\mathcal{W}^t,\end{aligned}$$

where the first and third inequalities are based on Young's inequality. □

Lemma C.13. *Under the settings of Lemma 3.5, it holds that*

$$
\begin{aligned}
&\sum_{i=1}^{m}\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\sum_{j=1}^{n}\frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t)\right)\right\|^2\right]\\
\le&\frac{3mn\hat{L}^2}{b}\left(\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\right)+\frac{12m^2n\hat{L}}{b}\max\left\{\frac{2}{13\theta_1},\frac{2\theta_1}{\theta_2}\right\}V^t.
\end{aligned}
$$

Proof. It holds that

$$
\begin{aligned}
&\sum_{i=1}^{m}\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\sum_{j=1}^{n}\frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t)\right)\right\|^2\right]\\
\le&\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\frac{\xi_{i,j}^{t+1}}{q_{i,j}}\left(\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t)\right)\right\|^2\right]=\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{q_{i,j}}\left\|\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(w_i^t)\right\|^2\\
\le&\frac{3}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{q_{i,j}}\left(\left\|\nabla f_{i,j}(x_i^t)-\nabla f_{i,j}(\bar{x}^t)\right\|^2+\left\|\nabla f_{i,j}(\bar{x}^t)-\nabla f_{i,j}(\bar{w}^t)\right\|^2+\left\|\nabla f_{i,j}(w_i^t)-\nabla f_{i,j}(\bar{w}^t)\right\|^2\right)\\
\le&\frac{3}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{q_{i,j}}\left(L_{i,j}^2\left\|x_i^t-\bar{x}^t\right\|^2+L_{i,j}^2\left\|w_i^t-\bar{w}^t\right\|^2+\left\|\nabla f_{i,j}(\bar{x}^t)-\nabla f_{i,j}(\bar{w}^t)\right\|^2\right),
\end{aligned}
$$

where the first and second inequality are based on Young's inequality. In the third inequality, we use smoothness of $f_{i,j}(\cdot)$. In the second last inequality, we use the convexity of $f(\cdot)$. Then we consider the terms separately. First, we can obtain the consensus error terms as

$$
\begin{aligned}
&\frac{3}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{q_{i,j}}\left(L_{i,j}^2\left\|x_i^t-\bar{x}^t\right\|^2+L_{i,j}^2\left\|w_i^t-\bar{w}^t\right\|^2\right)\\
=&\frac{3}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{L_{i,j}^2}{q_{i,j}}\left(\left\|x_i^t-\bar{x}^t\right\|^2+\left\|w_i^t-\bar{w}^t\right\|^2\right)\\
\overset{(a)}{\le}&\frac{3m\hat{L}^2}{b}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\left\|x_i^t-\bar{x}^t\right\|^2+\left\|w_i^t-\bar{w}^t\right\|^2\right)\\
=&\frac{3mn\hat{L}^2}{b}\left(\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\right),
\end{aligned}
$$

where the step (a) is based on the setting $q_{i,j}=\min\left\{1, bL_{i,j}/(mn\bar{L}_{\max})\right\}$ and facts $mn\ge b$ and $\hat{L}\ge\bar{L}_{\max}$ that leads to

$$
\frac{L_{i,j}}{q_{i,j}}=L_{i,j}\cdot\max\left\{1,\frac{mn\bar{L}_{\max}}{bL_{i,j}}\right\}=\max\left\{L_{i,j},\frac{mn\bar{L}_{\max}}{b}\right\}\le\frac{mn\hat{L}}{b}. \tag{23}
$$

Then we consider that

$$
\begin{aligned}
&\frac{3}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{q_{i,j}}\left\|\nabla f_{i,j}(\bar{x}^t)-\nabla f_{i,j}(\bar{w}^t)\right\|^2\\
\le&\frac{3}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{q_{i,j}}\left(2\left\|\nabla f_{i,j}(\bar{x}^t)-\nabla f_{i,j}(x^*)\right\|^2+2\left\|\nabla f_{i,j}(\bar{w}^t)-\nabla f_{i,j}(x^*)\right\|^2\right)\\
\overset{(12)}{\le}&\frac{12}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{L_{i,j}}{q_{i,j}}\left(\left(f_{i,j}(\bar{x}^t)-f_{i,j}(x^*)\right)+\left(f_{i,j}(\bar{w}^t)-f_{i,j}(x^*)\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(23)}{\leq} \frac{12m\hat{L}}{b}\sum_{i=1}^{m}\sum_{j=1}^{n}\big((f_{i,j}(\bar{x}^t) - f_{i,j}(x^*)) + (f_{i,j}(\bar{w}^t) - f_{i,j}(x^*))\big) \\
&= \frac{12m^2n\hat{L}}{b}\big((f(\bar{x}^t) - f(x^*)) + (f(\bar{w}^t) - f(x^*))\big) \\
&\leq \frac{12m^2n\hat{L}}{b}\big((f(\bar{z}^t) - f(x^*)) + (f(\bar{y}^t) - f(x^*)) + 2(f(\bar{w}^t) - f(x^*))\big) \\
&\leq \frac{12m^2n\hat{L}}{b}\left(\frac{L}{2}\left\|\bar{z}^t - x^*\right\|^2 + (f(\bar{y}^t) - f(x^*)) + 2(f(\bar{w}^t) - f(x^*))\right) \\
&= \frac{12m^2n\hat{L}}{b}\left(\frac{2\eta}{1+\eta\sigma}\mathcal{Z}^t + \theta_1\mathcal{Y}^t + \frac{2p\lambda\theta_1}{\theta_2}\mathcal{W}^t\right) \\
&\leq \frac{12m^2n\hat{L}}{b}\max\left\{\frac{2}{13\theta_1}, \frac{2\theta_1}{\theta_2}\right\}V^t,
\end{aligned}
$$

where the fourth inequality is based on the fact that $f(\cdot)$ is convex. Then by combining the above results, we can obtain that

$$
\begin{aligned}
&\sum_{i=1}^{m}\mathbb{E}_{\xi_{i,j}^{t+1}}\left[\left\|\sum_{j=1}^{n}\frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t) - \nabla f_{i,j}(w_i^t)\right)\right\|^2\right] \\
\leq& \frac{3mn\hat{L}^2}{b}\left(\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2\right) + \frac{12m^2n\hat{L}}{b}\max\left\{\frac{2}{13\theta_1}, \frac{2\theta_1}{\theta_2}\right\}V^t,
\end{aligned}
$$

which concludes the proof. □

Lemma C.14. *Under the settings of Lemma 3.5, it holds that*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{v}^{t+1} - \mathbf{v}^t\right\|^2\right] \leq& 12\left\|\mathbf{u}^{t+1} - \mathbf{1}\bar{u}^{t+1}\right\|^2 + \frac{12mn\hat{L}^2}{b}\left\|\mathbf{x}^{t+1} - \mathbf{1}\bar{x}^{t+1}\right\|^2 + \frac{24mn\hat{L}^2}{b}\left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2 \\
&+ 12\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{12mn\hat{L}^2}{b}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{24mn\hat{L}^2}{b}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
&+ \frac{24m^2n\hat{L}}{b}\max\left\{\frac{2}{13\theta_1}, \frac{5\theta_1}{2\theta_2}\right\}\cdot(V^t + V^{t+1}).
\end{aligned}
$$

Proof. It holds that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2\right] =& \sum_{i=1}^{m}\mathbb{E}\left[\left\|u_i^t + \sum_{j=1}^{n}\frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t) - \nabla f_{i,j}(w_i^t)\right)\right\|^2\right] \\
\leq& \sum_{i=1}^{m}\mathbb{E}\left[2\left\|u_i^t\right\|^2 + 2\left\|\sum_{j=1}^{n}\frac{\xi_{i,j}^{t+1}}{nq_{i,j}}\left(\nabla f_{i,j}(x_i^t) - \nabla f_{i,j}(w_i^t)\right)\right\|^2\right] \\
\leq& 6\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{6mn\hat{L}^2}{b}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{12mn\hat{L}^2}{b}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
&+ \frac{12m^2n\hat{L}}{b}\max\left\{\frac{2}{13\theta_1}, \frac{5\theta_1}{2\theta_2}\right\}\cdot V^t,
\end{aligned}
$$

where the last inequality is because of Lemma C.12 and C.13. Then we can obtain the desired result by using the fact use the fact that $\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2 \leq 2\|\mathbf{v}^{t+1}\|^2 + 2\left\|\mathbf{v}^t\right\|^2$. □

Substituting the result of Lemma C.11 and Lemma C.14 into Lemma C.9, we obtain the result of Lemma 3.5. Here, we rewrite Lemma 3.5 by including the detailed expressions of A and h^t.

Lemma C.15. *Under the settings of Lemma 3.4, we run Algorithm 2 by taking*

$$K = \left\lceil \frac{\log(1/\rho)}{\sqrt{\gamma}} \right\rceil \qquad \text{with} \qquad 1/\rho = \mathcal{O}\left(\mathrm{poly}(m, n, \kappa)\right).$$

Then it holds

$$\mathbb{E}\left[r^{t+1}\right] \leq \rho^2\big(A \cdot r^t + h^t\big), \tag{24}$$

where

$$A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 24(1+2\rho^2) & a_{22} & a_{23} & a_{24} \\ 0 & \frac{3\eta^2}{(1+\eta\sigma)^2} & \frac{3+9\theta_1^2\eta^2\sigma^2}{(1+\eta\sigma)^2} & \frac{9(1-\theta_1-\theta_2)^2\eta^2\sigma^2}{(1+\eta\sigma)^2} \\ 0 & \frac{9\rho^2\eta^2\theta_1^2}{(1+\eta\sigma)^2} & 6\theta_1^2 + \frac{9\rho^2\theta_1^2+27\rho^2\theta_1^4\eta^2\sigma^2}{(1+\eta\sigma)^2} & 3(1-\theta_1-\theta_2)^2 + \frac{27\rho^2\theta_1^2(1-\theta_1-\theta_2)^2\eta^2\sigma^2}{(1+\eta\sigma)^2} \end{bmatrix}$$

and

$$h^t = \begin{bmatrix} \frac{8\hat{L}^2}{Lm}\left(\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 + \left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2\right) + \frac{16\hat{L}\theta_1 p\lambda}{L\theta_2}(\mathcal{W}^{t+1} + \mathcal{W}^t) \\ h_2 \\ \frac{9\eta^2\sigma^2\theta_2^2 L}{(1+\eta\sigma)^2 m}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\ \left(9\theta_2^2 + \frac{27\rho^2\eta^2\sigma^2\theta_2^2\theta_1^2}{(1+\eta\sigma)^2}\right)\frac{L}{m}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \end{bmatrix},$$

with

$$\begin{aligned} a_{22} =& 2 + \frac{9\rho^2\eta^2\theta_1^2}{(1+\eta\sigma)^2} + \frac{27\rho^4\eta^2\theta_1^2(1-\theta_1-\theta_2)^2}{(1+\eta\sigma)^2} \\ a_{23} =& \frac{72mn\hat{L}^2\theta_1^2}{bL^2}\left(1 + \frac{\rho^2(3+9\theta_1^2\eta^2\sigma^2)}{(1+\eta\sigma)^2} + \rho^2(1-\theta_1-\theta_2)^2\left(6 + \frac{9\rho^2 + 27\rho^2\theta_1^2\eta^2\sigma^2}{(1+\eta\sigma)^2}\right)\right) \\ a_{24} =& \frac{72mn\hat{L}^2(1-\theta_1-\theta_2)^2}{bL^2}\left(1 + \frac{9\rho^2\theta_1^2\eta^2\sigma^2}{(1+\eta\sigma)^2} + \rho^2(1-\theta_1-\theta_2)^2\left(3 + \frac{27\rho^2\theta_1^2\eta^2\sigma^2}{(1+\eta\sigma)^2}\right)\right) \\ h_2 =& \frac{48(1+\frac{3}{2}\theta_2^2)n\hat{L}^2}{bL}\left(\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 + \left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2\right) + \frac{48m^2n\hat{L}}{bL}\max\left\{\frac{2}{13\theta_1}, \frac{21\theta_1}{2\theta_2}\right\}\cdot(V^{t+1} + V^t) \\ &+ \frac{648\rho^2n\hat{L}^2\theta_2^2}{bL}\left(\frac{\eta^2\sigma^2\theta_1^2}{(1+\eta\sigma)^2} + (1-\theta_1-\theta_2)^2 + \frac{3\rho^2\eta^2\sigma^2(1-\theta_1-\theta_2)^2\theta_1^2}{(1+\eta\sigma)^2}\right)\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2. \end{aligned}$$

Additionally, we have

$$\|A\| \leq \frac{91m^3n^3}{b} \tag{25}$$

and

$$\begin{aligned} \left\|h^t\right\| <& \frac{48m^3n^3}{b}\max\left\{\frac{2}{13\theta_1}, \frac{65\theta_1}{6\theta_2}\right\}(V^{t+1} + V^t) + \frac{(66+324\rho^2)n\hat{L}^2}{bL}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 + \frac{324\rho^2n\hat{L}^2}{bL}\left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2 \\ \leq& \frac{48m^3n^3}{b}\max\left\{\frac{2}{13\theta_1}, \frac{65\theta_1}{6\theta_2}\right\}(V^{t+1} + V^t) + \frac{(66+324\rho^2)m^2n^3L}{b}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\ &+ \frac{324\rho^2m^2n^3L}{b}\left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2. \end{aligned} \tag{26}$$

C.3. Proof of Theorem 3.7

Proof. We prove this theorem by induction. We assume that

$$\mathbb{E}\left[V^t\right] \leq \Bigg(\underbrace{\max\left(1 - \frac{1}{2}\eta\sigma, 1 - \frac{1}{2}\left(\theta_1 + \theta_2 - \frac{\theta_2}{\lambda}\right), 1 - \frac{1}{2}p(1-\lambda)\right)}_{\alpha}\Bigg)^t \left(V^0 + \left\|r^0\right\|\right) \tag{27}$$

and

$$\mathbb{E}\left[\left\|r^t\right\|\right] \leq \alpha^t\left(V^0+\left\|r^0\right\|\right) \tag{28}$$

holds when $t \leq k$ and are going to prove it holds for $t = k+1$. It should be straightforward that the two assumptions hold when $t = 0$. We use the notation A and h^t by following Lemma C.15.

Based on Lemma C.15 and Lemma 3.4, we need to first deal with additional consensus error $\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2$ and $\left\|\mathbf{x}^{t+1}-\mathbf{1}\bar{x}^{t+1}\right\|^2$ for $t = 0, \cdots, k$. From the simple observation that $1 - p/2 \leq \alpha$, we can first obtain an estimation that for $t = 0, \cdots, k$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2\right] \overset{(22)}{=} & \sum_{s=1}^{t} p(1-p)^{t-s}\mathbb{E}\left[\left\|\mathbf{y}^s-\mathbf{1}\bar{y}^s\right\|^2\right] \\ \leq & \frac{m}{L}\sum_{s=1}^{t} p(1-p)^{t-s}\mathbb{E}\left[\left\|r^s\right\|\right] \overset{(28)}{\leq} \frac{m}{L}\sum_{s=1}^{t} p(1-p)^{t-s}\alpha^s\left(V^0+\left\|r^0\right\|\right).\end{aligned}$$

We can further obtain that

$$\sum_{s=1}^{t} p(1-p)^{t-s}\alpha^s = p\sum_{s=1}^{t}\left(\frac{1-p}{\alpha}\right)^{t-s}\alpha^t \leq p\sum_{s=1}^{t}\left(1-\frac{2}{3}p\right)^{t-s}\alpha^t \leq \frac{3}{2}\alpha^t, \tag{29}$$

where the first inequality is based on the fact that $\alpha \leq 1 - p/2$ and $p \leq 1/2$. Therefore, by plugging in, we can obtain that for $t = 0, \ldots, k$,

$$\mathbb{E}\left[\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2\right] \overset{(29)}{\leq} \frac{3m}{2L}\alpha^t\left(V^0+\left\|r^0\right\|\right), \tag{30}$$

As equation (30) might be loose, we then seek an even tighter bound for $\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2$. By Lemma C.15, we can obtain that for all $t = 1, \cdots, k+1$, it holds that

$$\begin{aligned} & \mathbb{E}\left[\left\|r^t\right\|\right] \\ \overset{(24)}{\leq} & \mathbb{E}\left[\rho^2\left\|A \cdot r^{t-1}+h^{t-1}\right\|\right] \leq \mathbb{E}\left[\rho^2\left\|A\right\|\left\|r^{t-1}\right\|+\rho^2\left\|h^{t-1}\right\|\right] \\ \overset{(25),(26),(27)}{\leq} & \rho^2\mathbb{E}\left[\frac{91m^3n^3}{b}\alpha^{t-1}\left(V^0+\left\|r^0\right\|\right)+\frac{48m^3n^3}{b}\max\left\{\frac{2}{13\theta_1}, \frac{65\theta_1}{6\theta_2}\right\}\left(V^t+V^{t-1}\right)\right] \\ & +\rho^2\mathbb{E}\left[\frac{(66+324\rho^2)m^2n^3L}{b}\left\|\mathbf{w}^{t-1}-\mathbf{1}\bar{w}^{t-1}\right\|^2+\frac{324\rho^2m^2n^3L}{b}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\right] \\ \overset{(30)}{\leq} & \rho^2\left[\frac{91m^3n^3}{b}\alpha^{t-1}+\frac{96m^3n^3\alpha^{t-1}}{b}\max\left\{\frac{2}{13\theta_1}, \frac{65\theta_1}{6\theta_2}\right\}+\frac{(66+648\rho^2)m^3n^3}{b}\cdot\frac{3}{2}\alpha^{t-1}\right]\cdot\left(V^0+\left\|r^0\right\|\right) \\ \leq & \frac{2750m^3n^3\rho^2\alpha^t}{b}\cdot\left(V^0+\left\|r^0\right\|\right),\end{aligned} \tag{31}$$

where the last inequality holds by the setting $\alpha > 1/2$, $b = \sqrt{mn\bar{\kappa}_{\max}/\kappa}$, $\theta_1 = 1/(2\sqrt{\kappa})$, $\theta_2 = \bar{\kappa}_{\max}/(2b\kappa)$ and $1/\rho \geq 18\sqrt{2}$. Thus we obtain a sharper upper bound of $\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2$ for $t = 1, \cdots, k$, that is

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2\right] \overset{(22)}{=} & \sum_{s=1}^{t} p(1-p)^{t-s}\mathbb{E}\left[\left\|\mathbf{y}^s-\mathbf{1}\bar{y}^s\right\|^2\right] \\ \leq & \frac{m}{L}\sum_{s=1}^{t} p(1-p)^{t-s}\mathbb{E}\left[\left\|r^s\right\|\right] \\ \overset{(31)}{\leq} & \sum_{s=1}^{t} p(1-p)^{t-s}\alpha^s\cdot\frac{2750m^4n^3\rho^2}{bL}\cdot\left(V^0+\left\|r^0\right\|\right) \\ \overset{(29)}{\leq} & \frac{4125m^4n^3\rho^2\alpha^t}{bL}\cdot\left(V^0+\left\|r^0\right\|\right).\end{aligned} \tag{32}$$

Next, we bound $\left\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\right\|^2$ by

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\right\|^2\right] &\leq\ 3\theta_1^2\mathbb{E}\left[\left\|\mathbf{z}^k - \mathbf{1}\bar{z}^k\right\|^2\right] + 3\theta_2^2\mathbb{E}\left[\left\|\mathbf{w}^k - \mathbf{1}\bar{w}^k\right\|^2\right] + 3(1-\theta_1-\theta_2)^2\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{1}\bar{y}^k\right\|^2\right] \\ &\leq\ \frac{3m}{L}\mathbb{E}\left[\left\|r^k\right\|\right] + \mathbb{E}\left[\left\|\mathbf{w}^k - \mathbf{1}\bar{w}^k\right\|^2\right] \\ &\overset{(31),(32)}{\leq}\ \frac{12375m^4n^3\rho^2\alpha^k}{bL}\cdot\left(V^0 + \left\|r^0\right\|\right).\end{aligned}\tag{33}$$

Furthermore, Cauchy–Schwarz inequality implies

$$\mathbb{E}\left[\sqrt{\frac{2\eta\hat{L}^2V^k}{(1+\eta\sigma)mL}}\left\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\right\|\right] \leq \sqrt{\frac{2\eta\hat{L}^2\mathbb{E}[V^k]}{(1+\eta\sigma)mL}}\cdot\sqrt{\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\right\|^2\right]}.\tag{34}$$

Now we finish bounding all the pieces of consensus error and notice that each piece is multiplied by at least ρ. Denote

$$\beta = \max\left\{\frac{1}{1+\eta\sigma}, 1-\left(\theta_1+\theta_2-\frac{\theta_2}{\lambda}\right), 1-p(1-\lambda)\right\},$$

and by substituting the estimation of the pieces above, from Lemma 3.4, we have

$$\begin{aligned}&\mathbb{E}[V^{k+1}] \\ \overset{(27)}{\leq}\quad &\mathbb{E}\left[\beta V^k + \sqrt{\frac{2\eta\hat{L}^2V^k}{(1+\eta\sigma)mL}}\left\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\right\| + \frac{C_1}{24L\theta_1}\left\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\right\|^2 + \frac{C_2}{24L\theta_1}\left\|\mathbf{w}^k - \mathbf{1}\bar{w}^k\right\|^2\right] \\ \overset{(31),(34),(33)}{\leq}\quad &\mathbb{E}\left[\beta V^k + \sqrt{\frac{4\eta\hat{L}^2}{(1+\eta\sigma)mL}}\sqrt{\frac{12375m^4n^3\rho^2}{bL}}\cdot\alpha^k\left(V^0 + \left\|r^0\right\|\right) + \frac{C_1\left\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\right\|^2 + C_2\left\|\mathbf{w}^k - \mathbf{1}\bar{w}^k\right\|^2}{24L\theta_1}\right] \\ \overset{(32),(33)}{\leq}\quad &\alpha^k\left(\beta + \rho\cdot C_3 + \rho^2\cdot C_4\right)(V^0 + \left\|r^0\right\|),\end{aligned}$$

where

$$C_3 = \sqrt{\frac{4\eta\hat{L}^2}{(1+\eta\sigma)mL^2}\cdot\frac{12375m^3n^3}{b}} \qquad \text{and} \qquad C_4 = \frac{C_1}{24L\theta_1}\frac{12375m^3n^3}{b} + \frac{C_2}{24L\theta_1}\frac{4125m^3n^3}{b},$$

and C_1 and C_2 are defined at (18). Therefore, above result and the fact $\alpha - \beta = \Theta(1/\sqrt{\kappa}\,)$ means have

$$\mathbb{E}\left[V^{k+1}\right] \leq \alpha^k(V^0 + \left\|r^0\right\|)\cdot\left(\beta + \left(1 - \frac{1}{2}\beta\right)\right) \leq \alpha^{k+1}(V^0 + \left\|r^0\right\|),$$

and

$$\mathbb{E}\left[\left\|r^{k+1}\right\|\right] \leq \alpha^{k+1}\cdot\left(V^0 + \left\|r^0\right\|\right),$$

by taking $\rho > 0$ such that $1/\rho = \mathcal{O}\left(\mathrm{poly}(m,n,\kappa)\right)$, which finishes the induction. Hence, we have

$$\mathbb{E}\left[V^t + \left\|r^t\right\|\right] \leq 2\left(\underbrace{\max\left(1-\frac{1}{2}\eta\sigma, 1-\frac{1}{2}\left(\theta_1+\theta_2-\frac{\theta_2}{\lambda}\right), 1-\frac{1}{2}p(1-\lambda)\right)}_{\alpha}\right)^t\left(V^0 + \left\|r^0\right\|\right).$$

for all t. The condition of ρ can be satisfied by taking

$$K = \frac{\log(1/\rho)}{\sqrt{\gamma}} = \mathcal{O}\left(\frac{\log(\mathrm{poly}\,(mn\kappa))}{\sqrt{\gamma}}\right).$$

Combining the above results, we finish the proof.

□

C.4. Proof of Corollary 3.8

Proof. We first prove that CESAR (Algorithm 2) can find an ϵ-suboptimal solution in expectation. We run Algorithm 2 with the setting of Theorem 3.7 and let

$$T = \mathcal{O}\left(\left(\frac{\kappa}{\eta} + \frac{2}{\theta_1 + \theta_2 - \frac{\theta_2}{\lambda}} + \frac{2}{p(1-\lambda)}\right)\log\frac{1}{\epsilon}\right). \tag{35}$$

Then Theorem 3.7 means

$$\mathbb{E}[f(\bar{y}^T) - f(x^*)] \le \frac{\epsilon}{2},$$

and $\mathbb{E}\left[\left\|y_i^T - \bar{y}^T\right\|^2\right] \le m\epsilon/L$ for each $i \in [m]$. Moreover, Proposition 2.6 means step $\mathbf{y}_{\rm out} = {\rm FastMix}(\mathbf{y}_T, K_{\rm out})$ with $K_{\rm out} = \mathcal{O}\big(\sqrt{1/\alpha}\log m\big)$ (line 20 of Algorithm 2) leads to $\bar{y}^{\rm out} = \bar{y}^T$ and

$$L\mathbb{E}\left[\left\|y_i^{\rm out} - \bar{y}^{\rm out}\right\|^2\right] \le \epsilon \tag{36}$$

for each $i \in [m]$. Applying the smoothness of f (Assumption 2.1), we have

$$f(y_i^{\rm out}) - f(\bar{y}^{\rm out}) \le \frac{L}{2}\left\|y_i^{\rm out} - \bar{y}^{\rm out}\right\|^2 \le \frac{\epsilon}{2}$$

for each $i \in [m]$. Together we have

$$f(y_i^{\rm out}) - f(x^*) = f(y_i^{\rm out}) - f(\bar{y}^{\rm out}) + f(\bar{y}^{\rm out}) - f(x^*) \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

for each $i \in [m]$. This implies that output $\mathbf{y}^{\rm out}$ is an ϵ-suboptimal solution in expectation.

Then we analyze the complexity of Algorithm 2 by following the parameter settings of Theorem 3.7. As we set $p = \max\{\theta_1, \theta_2\}$, when it holds that $\theta_1 > \theta_2$, we set the auxiliary constant $\lambda = 2/3$, then we have

$$\frac{2}{\theta_1 + \theta_2 - \frac{\theta_2}{\lambda}} + \frac{2}{p(1-\lambda)} \le \frac{4}{\theta_1} + \frac{6}{\theta_1} = \frac{10}{\theta_1}.$$

Else if $\theta_1 \le \theta_2$, we set the auxiliary constant $\lambda = 2\theta_2/(2\theta_2 + \theta_1) \in [2/3, 1)$ and can obtain that

$$\frac{2}{\theta_1 + \theta_2 - \frac{\theta_2}{\lambda}} + \frac{2}{p(1-\lambda)} \le \frac{4}{\theta_1} + \frac{2}{\theta_2(1-\lambda)} \le \frac{4}{\theta_1} + \frac{6}{\theta_1} = \frac{10}{\theta_1}.$$

Therefore, by substituting the parameters setting into (35), we can obtain

$$\left(\frac{\kappa}{\eta} + \frac{2}{\theta_1 + \theta_2 - \frac{\theta_2}{\lambda}} + \frac{2}{p(1-\lambda)}\right)\log\frac{1}{\epsilon} = \mathcal{O}\left(\sqrt{\kappa}\log\frac{1}{\epsilon}\right)$$

and thus

$$T = \mathcal{O}\left(\sqrt{\kappa}\log\frac{1}{\epsilon}\right).$$

Then we directly achieve the communication complexity by $TK = \tilde{\mathcal{O}}\big(\sqrt{\kappa/\gamma}\log(1/\epsilon)\big)$ and the expected LIFO complexity by $T(mnp + \sum_{i=1}^m\sum_{j=1}^n q_{i,j}) = \mathcal{O}\left(\left(\sqrt{mn\bar{\kappa}_{\max}} + mn\,\right)\log(1/\epsilon)\right)$.

Then we consider the computation time. As discussed in Section 3.2.1, it contains the full-batch gradient computation at snapshot points and mini-batch gradient computation at all points. For the snapshot points, it takes

$$np\cdot T = \mathcal{O}\left(\max\left\{\sqrt{\frac{n\bar{\kappa}_{\max}}{m}}, n\right\}\cdot\log\frac{1}{\epsilon}\right).$$

For the mini-batch costs, it takes

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\max_{i\in[m]} Y_i^t\right] = \mathcal{O}\left(\sqrt{\frac{n\bar{\kappa}_{\max}}{m}} + \sqrt{\kappa}\cdot(\ln mn)^2\right),$$

where we bound $\mathbb{E}\left[\max_{i\in[m]} Y_i^t\right]$ by using Theorem 3.3.

Summing over above results, we can obtain the overall computation time complexity

$$\tilde{\mathcal{O}}\left(\left(\sqrt{\frac{n\bar{\kappa}_{\max}}{m}} + n + \sqrt{\kappa}\right)\log\frac{1}{\epsilon}\right),$$

which finishes the proof. □

D. Proof of Section 4

The proofs for lower complexity bounds in Section 4 follow the construction of Hendrikx et al. (2021, Theorem 4.1 and Corollary 4.3). We modify the instances of Hendrikx et al. (2021) by considering different condition numbers κ, $\bar{\kappa}_{\max}$ and $\bar{\kappa}$.

D.1. Black-box Procedure

We use the notion of black-box procedure by following the definition of Hendrikx et al. (2021) and we present its details for completeness. The following definition does not include the oracles of dual gradient and proximal operation, since we focus on dual-free methods.

Specifically, we consider the black-box procedure for distributed algorithms on a system of m nodes, that respect:

- **Local Memory:** Each node $i \in [m]$ has a local memory $\mathcal{M}_{i,t}$ at time t. The values in this local memory can only come from either local computation $\mathcal{M}_{i,t}^C$ or communication $\mathcal{M}_{i,t}^G$, so that for all $i \in [m]$, $\mathcal{M}_{i,t} \subseteq \mathcal{M}_{i,t}^C \cup \mathcal{M}_{i,t}^G$.
- **Local Computation:** Each node i can, at time t, compute $\nabla f_{i,\xi_{t,i}}(x)$, where $\xi_{t,i} \in [n]$ can be arbitrarily chosen by the algorithm and $x \in \mathcal{M}_{i,t-1}$. This is equivalent to that
$$\mathcal{M}_{i,t}^G = \text{Span}\left(\{x, \nabla f_{i,\xi_{t,i}}(x) : x \in \mathcal{M}_{i,t-1}\}\right).$$
- **Local Communication:** Each node i can, at time t, share a value to its neighbours so that for all $i \in [m]$,
$$\mathcal{M}_{i,t}^C = \text{Span}\left(\bigcup_{j\in\mathcal{N}(i),\tau\in[t-1]} \mathcal{M}_{j,t-\tau}\right).$$
- **Output Value:** Each node i must, at time t, specify one vector x_i^t in its memory as the local output of the algorithm, that is, for all $i \in [m]$, $x_i^t \in \mathcal{M}_{i,t}$.

D.2. Proof of Theorem 4.1

Proof of Theorem 4.1. We consider the network of m nodes associated to graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edge. We let Q be a subset of $\mathcal{V}$ and $Q_\Delta^c = \{v \in \mathcal{V} : \text{dis}(v, Q) \geq \Delta\}$, where $\text{dis}(v, Q)$ is the distance from v to Q, i.e. the smallest distance between v and node $v' \in Q$, where we assume the distance between neighbour nodes is 1. Denote the number of nodes in Q and Q_Δ^c as $|Q|$ and $|Q_\Delta^c|$, and assume that $|Q| \geq |Q_\Delta^c|$ without loss of generality.

For any $L, \mu > 0$ such that $L/\mu \in [1, \kappa]$. we define the functions $\psi_i^Q : \ell_2 \to \mathbb{R}$ as

$$\psi_i^Q(y) = \frac{1}{2|Q|}\left[\frac{\mu}{3}\|y\|^2 + \frac{L-\mu}{4}(y^T M_1 y - e_1^T y)\right], \quad \text{if} \quad i \in Q$$

$$\psi_i^Q(y) = \frac{1}{2|Q_\Delta^c|}\left[\frac{\mu}{3}\|y\|^2 + \frac{L-\mu}{4}y^T M_2 y\right], \quad \text{if} \quad i \in Q_\Delta^c$$
$$\psi_i^Q(y) = \frac{\mu}{6(m-|Q_\Delta^c|-|Q|)}\|y\|^2, \quad \text{otherwise,}$$

where M_1 is the infinite block diagonal matrices with $\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}$, $M_2 = \begin{bmatrix}1 & 0\\ 0 & M_1\end{bmatrix}$ and $e_1 = [1, 0, 0, \dots] \in \ell_2$.

Then we can construct the individual functions $f_{i,j} : \ell_2^n \to \mathbb{R}$, and objective function $f : \ell_2^n \to \mathbb{R}$ as

$$f_{i,j}(x) = \psi_i^Q(x_j) \quad \text{and} \quad f(x) = \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n f_{i,j}(x),$$

where $x = [x_1; \dots; x_n] \in \ell_2^n$ and each $x_j \in \ell_2$.

We can verify that $f(\cdot)$ is $L/(mn)$-smooth and $\mu/(mn)$-strongly convex since it holds $0 \preceq M_1, M_2 \preceq 2I$. Thus, the condition number in our construction satisfies $\kappa \geq L/\mu$.

Next, we consider the solution of $\min_{x\in\ell_2^n} f(x)$. We start from the analysis for the solution of problem $\min_{y\in\ell_2}\psi(y)$ where $\psi(y) = (1/m)\sum_{i=1}^m \psi_i^Q(y)$. The definition of $\psi_i^Q(\cdot)$ implies

$$\psi(y) = \frac{1}{m}\sum_{i=1}^m \psi_i^Q(y) = \frac{L-\mu}{8}\left(y^\top(M_1+M_2)y - e_1^\top y\right) + \frac{\mu}{2}\|y\|_2^2.$$

Since $\psi(\cdot)$ is strongly convex, it has the unique minimizer $y^* = [y^*(0), y^*(1), \dots]^\top$. The optimality condition $\nabla\psi(y^*) = 0$ implies we have

$$\mu y_k^* + \frac{L-\mu}{4}[2y^*(k) - y^*(k-1) - y^*(k+1)] = 0$$

for all $k \geq 1$ and $y^*(0) = 1$. By induction, we can show that $y^*(k) = q^k$ with $q = (\sqrt{L/\mu}-1)/(\sqrt{L/\mu}+1)$.

We can further obtain that the minimizer $x^* = [x_1^*; \dots; x_n^*] \in \ell_2^n$ of $f(\cdot)$ satisfies that $x_j^* = y^*$ for all $j \in [n]$. For a sequence $\{x_i^t\}_{i=1}^m$ with $x_i^t = [x_{i,1}^t; \dots; x_{i,n}^t] \in \ell_2^n$ and $x_{i,j}^t \in \ell_2$ generated by a black-box optimization procedure as defined in Section D.1 with $x_i^0 = 0$ for $i \in [m]$, we then have

$$\sum_{i=1}^m\sum_{j=1}^n \left\|x_{i,j}^t - x_j^*\right\|^2 \geq \sum_{i=1}^m\sum_{j=1}^n\sum_{l\geq k_j(t)}\left\|x_j^*(l)\right\|^2 = \sum_{i=1}^m\sum_{j=1}^n \frac{q^{2k_j(t)}}{1-q^2} = \sum_{j=1}^n \frac{mq^{2k_j(t)}}{1-q^2},$$

where $k_j(t)$ is the first index such that $x_{i,j}^t(l) = 0$ for all nodes $i \in [n]$ and $l \geq k_j(t)$. Then it holds that

$$\mathbb{E}\left[\frac{\|x^t - x^*\|^2}{\|x^0 - x^*\|^2}\right] = \mathbb{E}\left[\sum_{i=1}^m\sum_{j=1}^n \frac{\left\|x_{i,j}^t - x_j^*\right\|^2}{\left\|x_{i,j}^0 - x_j^*\right\|^2}\right] \leq \frac{m(1-q)}{1-q^2}\sum_{j=1}^n \mathbb{E}\left[q^{2k_j(t)}\right] \tag{37}$$

since we have $x_{i,j}^0 = 0$ and $x_{i,j}^* = y^*$ for all $i \in [m], j \in [n]$. Note that the upper bound on $k_j(t)$ leads to the lower bound on the expected error. Based on (37), we can provide a lower bound for communication round complexity as follows.

Consider time t and corresponding $k_j(t)$. We here discuss the time $t' > t$ such that $k_j(t') = k_j(t) + 2$. Let us first provide some straightforward intuition. For initial point $x = 0 \in \ell_2$, to make $x(l+1)$ non-zero, the structure of M_1 and M_2 means we require $x(l)$ non-zero; then ensure it is in the memory of node $i \in Q$ and call a LIFO on the node if l is odd (or node $i' \in Q_\Delta^c$ if l is even). It is clear that any LIFO (or equivalently, local computation) on $f_{i,j}(\cdot)$ with $i \notin Q \cup Q_\Delta^c$ is not helpful to increase $k_j(t)$.

Suppose. $k_j(t)$ is odd and there exists vector $x \in \ell_2$ with $x(k_j(t)) \neq 0$ in the memory of node $i \in Q$. To achieve $k_j(t') = k_j(t) + 2$, we have to first call an oracle of $f_{i,j}(\cdot)$, i.e. do local computation, on the node such that at time t'' we have $k_j(t'') = k_j(t) + 1$. Then the message with $(k_j(t)+1)$-th coordinate nonzero generated at node i must be sent to node

$i' \in Q_\Delta^c$. Then we can call the oracle on $f_{i',j}(\cdot)$ for $i' \in Q_\Delta^c$ and at time t' we have $k_j(t') = k_j(t) + 2$. This procedure takes at least Δ communication rounds. By the message with $(k_j(t)+2)$-th coordinate nonzero generated at node i' sent back, it costs at least another Δ communication rounds to let nodes in Q receive the message. When $k_j(t)$ is even, this procedure starts from a node $i' \in Q_\Delta^c$ is almost the same as the case of odd $k_j(t)$. Since we have $k_j(0) = 1$ and $k_j(t) \le 2$ for all $t < \Delta$, we have

$$k_j(t) \le 2 + \frac{t}{\Delta}. \tag{38}$$

Combining (37) and (38), we have

$$\frac{(1-q^2)}{(1-q)mn}\mathbb{E}\left[\sum_{i=1}^m\sum_{j=1}^n \frac{\left\|x_{i,j}^t - x_j^*\right\|^2}{\left\|x_{i,j}^0 - x_j^*\right\|^2}\right] \ge \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{4+\frac{2t}{\Delta}},$$

where we use

$$q = 1 - \frac{2}{\sqrt{L/\mu}+1} = 1 - \frac{2}{\sqrt{\kappa}+1}.$$

Therefore, finding an ϵ-suboptimal solution requires at least communication round complexity of

$$\Omega\left(\sqrt{\kappa}\Delta\log\left(\frac{1}{\epsilon}\right)\right).$$

Based on the constructing of Scaman et al. (2017), when $m \ge \sqrt{3/\gamma}$, there exists a linear graph such that $\Delta = \Theta(\sqrt{1/\gamma})$. Thus, we can obtain the lower complexity bound on communication as

$$\Omega\left(\sqrt{\frac{\kappa}{\gamma}}\log\left(\frac{1}{\epsilon}\right)\right).$$

□

D.3. Proof of Theorem 4.2

Proof of Theorem 4.2. We consider the cases of $\kappa \le n\bar{\kappa}_{\max}/m$ and $\kappa \ge n\bar{\kappa}_{\max}/m$ separately.

(a) First case: $\kappa \le n\bar{\kappa}_{\max}/m$. We employ the similar analysis in the proof of Theorem 4.1. Here, we give the details for completeness.

We consider the network of m nodes associated to graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edge. We let Q be a subset of $\mathcal{V}$ and $Q_\Delta^c = \{v \in \mathcal{V} : \mathrm{dis}(v, Q) \ge \Delta\}$, where $\mathrm{dis}(v, Q)$ is the distance from v to Q. Denote the number of nodes in Q and Q_Δ^c as $|Q|$ and $|Q_\Delta^c|$, and assume that $|Q| \ge |Q_\Delta^c|$.

For any $L, \mu > 0$ such that $L/\mu \in [1, \kappa]$. We define the functions $\psi_i^Q : \ell_2 \to \mathbb{R}$ as

$$\begin{aligned}
\psi_i^Q(y) &= \frac{1}{2|Q|}\left[\frac{\mu}{3}\|y\|^2 + \frac{L-\mu}{4}(y^T M_1 y - e_1^T y)\right], \quad \text{if} \quad i \in Q \\
\psi_i^Q(y) &= \frac{1}{2|Q_\Delta^c|}\left[\frac{\mu}{3}\|y\|^2 + \frac{L-\mu}{4} y^T M_2 y\right], \quad \text{if} \quad i \in Q_\Delta^c \\
\psi_i^Q(y) &= \frac{\mu}{6(m - |Q_\Delta^c| - |Q|)}\|y\|^2, \quad \text{otherwise,}
\end{aligned}$$

where M_1 is the infinite block diagonal matrices with $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, $M_2 = \begin{bmatrix} 1 & 0 \\ 0 & M_1 \end{bmatrix}$ and $e_1 = [1, 0, 0, \dots] \in \ell_2$.

Then we can construct the individual functions $f_{i,j} : \ell_2^n \to \mathbb{R}$, and objective function $f : \ell_2^n \to \mathbb{R}$ as

$$f_{i,j}(x) = \psi_i^Q(x_j) \quad \text{and} \quad f(x) = \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n f_{i,j}(x),$$

where $x = [x_1; \dots; x_n] \in \ell_2^n$ and each $x_j \in \ell_2$.

We can verify that $f(\cdot)$ is $L/(mn)$-smooth and $\mu/(mn)$-strongly convex since it holds $0 \preceq M_1, M_2 \preceq 2I$. Thus, the condition number in our construction satisfies that

$$\kappa \geq L/\mu \qquad \text{and} \qquad \bar{\kappa}_{\max} \geq \frac{1}{n}\max_{i\in[m]}\sum_{j=1}^{n}\frac{L_{i,j}}{\frac{\mu}{mn}} = \frac{mnL}{|Q|\mu}.$$

Next, we consider the solution of $\min_{x\in\ell_2^n} f(x)$. We start from the analysis for the solution of problem $\min_{y\in\ell_2}\psi(y)$ where $\psi(y) = (1/m)\sum_{i=1}^{m}\psi_i^Q(y)$. The definition of $\psi_i^Q(\cdot)$ implies

$$\psi(y) = \frac{1}{m}\sum_{i=1}^{m}\psi_i^Q(y) = \frac{L-\mu}{8}\left(y^\top(M_1+M_2)y - e_1^\top y\right) + \frac{\mu}{2}\left\|y\right\|_2^2.$$

Since $\psi(\cdot)$ is strongly convex, it has the unique minimizer $y^* = [y^*(0), y^*(1), \dots]^\top$. The optimality condition $\nabla\psi(y^*) = 0$ implies we have

$$\mu y_k^* + \frac{L-\mu}{4}[2y^*(k) - y^*(k-1) - y^*(k+1)] = 0$$

for all $k \geq 1$ and $y^*(0) = 1$. By induction, we can show that $y^*(k) = q^k$ with $q = (\sqrt{L/\mu}-1)/(\sqrt{L/\mu}+1)$.

For a sequence $\{x_i^t\}_{i=1}^m$ with $x_i^t = [x_{i,1}^t; \dots; x_{i,n}^t] \in \ell_2^n$ and $x_{i,j}^t \in \ell_2$ generated by a black-box optimization procedure as defined in Section D.1 with $x_i^0 = 0$ for $i \in [m]$, the minimizer $x^* = [x_1^*; \dots; x_n^*] \in \ell_2^n$ of $f(\cdot)$ satisfies that $x_j^* = y^*$ for all $j \in [n]$. We then have

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|x_{i,j}^t - x_j^*\right\|^2 \geq \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l\geq k_j(t)}\left\|x_j^*(l)\right\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{q^{2k_j(t)}}{1-q^2} = \sum_{j=1}^{n}\frac{mq^{2k_j(t)}}{1-q^2},$$

where $k_j(t)$ is the first index such that $x_{i,j}^t(l) = 0$ for all nodes $i \in [n]$ and $l \geq k_j(t)$. Then it holds that

$$\mathbb{E}\left[\frac{\left\|x^t - x^*\right\|^2}{\left\|x^0 - x^*\right\|^2}\right] = \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left\|x_{i,j}^t - x_j^*\right\|^2}{\left\|x_{i,j}^0 - x_j^*\right\|^2}\right] \leq \frac{m(1-q)}{1-q^2}\sum_{j=1}^{n}\mathbb{E}\left[q^{2k_j(t)}\right] \tag{39}$$

since we have $x_{i,j}^0 = 0$ and $x_{i,j}^* = y^*$ for all $i \in [m], j \in [n]$. Note that the upper bound on $k_j(t)$ leads to the lower bound on the expected error. Then based on (39), we can provide a lower bound for computation step complexity as follows.

Without loss of generality, we consider the initial point $x^0 = 0$. The definition of M_1 and M_2 results that only the LIFO call (or local computation) at nodes $i \in Q$ can increase $k_j(t)$ when $k_j(t)$ is odd (and only the LIFO call on $i' \in Q_\Delta^c$ can increase $k_j(t)$ when $k_j(t)$ is even). Furthermore, $k_j(t)$ can only be changed by calling the LIFO of the component j on node i (for odd $k_j(t)$) or i' (for even $k_j(t)$). Thus, we can upper bound $k_j(t)$ by the number of LIFO calls of component j for all $i \in [m]$ are called. Since at one computation step, a node i can only call LIFO of one component $f_{i,\xi_i(t)}$, we have

$$k_j(t) \leq 1 + \sum_{l=1}^{t}\sum_{i\in Q}\mathbb{1}(\xi_i(t) = j) + \sum_{l=1}^{t}\sum_{i'\in Q_\Delta^c}\mathbb{1}(\xi_{i'}(t) = j)$$

where $\mathbb{1}(\cdot)$ is the indicator function. This leads to

$$\begin{aligned}\sum_{j=1}^{n}k_j(t) \leq& n + \sum_{j=1}^{n}\sum_{l=1}^{t}\sum_{i\in Q}\mathbb{1}(\xi_i(t) = j) + \sum_{j=1}^{n}\sum_{l=1}^{t}\sum_{i'\in Q_\Delta^c}\mathbb{1}(\xi_{i'}(t) = j)\\ \leq& n + |Q|t + |Q_\Delta^c|t \leq n + 2|Q|t.\end{aligned}$$

By Jensen's inequality, we obtain

$$\frac{1}{n}\sum_{j=1}^{n}q^{2k_j(t)} \geq q^{\frac{2}{n}\sum_{j=1}^{n}k_j(t)} \geq q^{2+4|Q|t}. \tag{40}$$

Combining (39) and (40) and taking $|Q| = |Q_\Delta^c| = 1$, we have

$$\frac{(1-q^2)}{(1-q)mn}\mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left\|x_{i,j}^t - x_{i,j}^*\right\|^2}{\left\|x_{i,j}^0 - x_{i,j}^*\right\|^2}\right] \geq \left(1 - \frac{2n}{\sqrt{n\bar{\kappa}_{\max}/m + n}}\right)^{2+\frac{4t}{n}},$$

where we use

$$q = 1 - \frac{2n}{n\sqrt{L/\mu} + n} = 1 - \frac{2n}{\sqrt{n|Q|\bar{\kappa}_{\max}/m} + n}.$$

Thus, we require at least

$$\Omega\left(\left(\sqrt{\frac{n\bar{\kappa}_{\max}}{m}} + n\right)\log\left(\frac{1}{\epsilon}\right)\right)$$

computation steps to achieve an ϵ-suboptimal solution when $n\bar{\kappa}_{\max}/m \geq \kappa$.

(b) Second case: $\kappa \geq n\bar{\kappa}_{\max}/m$. In this case, we aim to prove a lower bound of $\Omega\left(\sqrt{\kappa}\log\left(1/\epsilon\right)\right)$ on computation steps to achieve an ϵ-suboptimal solution. Note that when $m \leq n$, from Proposition A.1, it is clear that $n\bar{\kappa}_{\max}/m \geq \kappa$, thus it should be categorized to the case (a). Hence, we focus on the case of $m > n$.

We consider the instance that all $f_{i,j}$ are identical. This means we have $f_{i,j}(\cdot) = f(\cdot)$ for all $i \in [m], j \in [n]$, where $f : \ell_2 \to \mathbb{R}$. For any L-smooth and μ-strongly convex function $f(\cdot)$, also $f_{i,j}(\cdot)$, we have

$$\kappa = \bar{\kappa}_{\max} = \frac{L}{\mu}, \quad \text{and thus} \quad \frac{n\bar{\kappa}_{\max}}{m} = \frac{n}{m}\kappa \leq \kappa.$$

Since we set $f_{i,j}(\cdot) = f(\cdot)$ for all $i \in [m]$ and $j \in [n]$, the problem can be regarded as minimizing $f(\cdot)$ on a single machine by full-batch methods. In this case, the complexity of computation steps corresponds to the complexity of deterministic first-order algorithms. From the well-known lower bound analysis of Nesterov (2018), we have the lower bound $\Omega\left(\sqrt{\kappa}\log(1/\epsilon)\right)$.

Combining above two cases, we obtain the lower bound on computation steps

$$\Omega\left(\left(\sqrt{\frac{n\bar{\kappa}_{\max}}{m}} + \sqrt{\kappa} + n\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

□

D.4. Proof of Theorem 4.3

We first provide a lower bound for finite-sum problem on single machine by following Agarwal & Bottou (2015); Nesterov (2018); Woodworth & Srebro (2016).

Lemma D.1. *Let $\bar{\kappa} \geq 1$ and $n \in \mathbb{N}$. There exist n functions $f_j : \ell_2^n \to \mathbb{R}$ such that each f_j is smooth and convex and objective function $f \triangleq 1/n\sum_{j=1}^n f_j$ is μ-strongly convex such that $\bar{\kappa} \geq \sum_{j=1}^n L_j/(n\mu)$, where L_j is the smoothness parameter of $f_j(\cdot)$ for all $j \in [n]$. Then for any black box procedure, finding a point $x \in \ell_2^n$ such that $\mathbb{E}[\|x^t - x^*\|^2] \leq \varepsilon$ requires at least*

$$\Omega\left(\left(\sqrt{n\bar{\kappa}} + n\right)\log\left(\frac{1}{\varepsilon}\right)\right)$$

incremental first-order oracle calls, where x^* is the minimizer of $f(\cdot)$.

Proof. We first construct individual functions $f_j : \ell_2^n \to \mathbb{R}$ and objective function f as

$$f_j(x) = \psi(x_j) \quad \text{and} \quad f(x) = \frac{1}{n}\sum_{j=1}^{n} f_j(x),$$

where $x = [x_1; \dots ; x_n] \in \ell_2^n$ with $x_j \in \ell_2$ for all $j \in [n]$. We define $\psi : \ell_2 \to \mathbb{R}$ as

$$\psi(y) = \frac{L-\sigma}{8}\left(y^\top M y - e_1^\top y\right) + \frac{\sigma}{2}\left\|y\right\|_2^2,$$

where $M = \begin{bmatrix} 1 & \\ & M_0 \end{bmatrix}$ and $M_0 = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}$ is an infinite tridiagonal matrix and $\sigma > 0$

We can verify $0 \preceq M \preceq 4I$, which implies $f(\cdot)$ is σ/n-strongly convex and each $f_j(\cdot)$ is L-smooth. Thus, it holds that $\bar{\kappa} \geq nL/\mu$.

Similar to the proof of Theorem 4.1, we can obtain that the minimizer of $\psi(\cdot)$ is $y^* = [y^*(1); y^*(2); \dots] \in \ell_2$ with the k-th coordinate being $y^*(k) = q^k$, where $q = (\sqrt{L/\sigma} - 1)/(\sqrt{L/\sigma} + 1)$. It implies the minimizer of $f(\cdot)$ is $x^* = [x_1^*; \dots ; x_n^*] \in \ell_2^n$ with $x_j^* = y^*$ for each $j \in [n]$.

Without loss of generality, we assume the initial point is $x^0 = 0$. Then it holds that

$$\sum_{j=1}^n \left\|x_j^t - x_j^*\right\|^2 \geq \sum_{j=1}^n \sum_{l \geq k_j(t)} \left\|x_j^*(l)\right\|^2 = \sum_{j=1}^n \frac{q^{2k_j(t)}}{1-q^2}, \tag{41}$$

where $k_j(t)$ is the first index such that $x_j^t(l) = 0$ for all $l \geq k_j(t)$. From $x^0 = 0$ we have that $\left\|x_j^0 - x_j^*\right\|^2 = 1/(1-q)$ for all $j \in [n]$, thus it holds that

$$\mathbb{E}\left[\frac{\left\|x^t - x^*\right\|^2}{\left\|x^0 - x^*\right\|^2}\right] = \mathbb{E}\left[\sum_{j=1}^n \frac{\left\|x_j^t - x_j^*\right\|^2}{\left\|x_j^0 - x_j^*\right\|^2}\right] \geq \frac{1-q}{1-q^2}\sum_{j=1}^n \mathbb{E}\left[q^{2k_j(t)}\right].$$

We then give lower bound on $\mathbb{E}\left[q^{2k_j(t)}\right]$. The structure of M means to make the $(k+1)$-th coordinate $x_j(k+1)$ be non-zero requires that $x_j(k)$ is non-zero and call incremental first-order oracle $\nabla f_j(\cdot)$. Thus we can obtain that

$$k_j(t) \leq \sum_{l=1}^t \mathbb{1}\left\{\zeta(t) = j\right\},$$

where $\zeta(t)$ be the index of incremental first-order oracle which is accessed at the t-th step. Then it holds that

$$\sum_{j=1}^n k_j(t) \leq \sum_{l=1}^t \sum_{j=1}^n \mathbb{1}\left\{\zeta(t) = j\right\} = t.$$

Using Jensen's inequality, we have

$$\frac{1}{n}\sum_{j=1}^n q^{k_j(t)} \geq q^{\frac{1}{n}\sum_{j=1}^n k_j(t)} \geq q^{\frac{t}{n}}. \tag{42}$$

By substituting (42) into (41), we can obtain that

$$\mathbb{E}\left[\frac{\left\|x^t - x^*\right\|^2}{\left\|x^0 - x^*\right\|^2}\right] \geq \frac{1-q}{1-q^2}\sum_{j=1}^n \mathbb{E}\left[q^{2k_j(t)}\right] \geq \frac{n(1-q)}{1-q^2}\sum_{j=1}^n q^{\frac{2t}{n}}.$$

Since $q = (\sqrt{L/\sigma} - 1)/(\sqrt{L/\sigma} + 1)$ and $\bar{\kappa} \geq nL/\sigma$, we need at least

$$\Omega\left(\left(\sqrt{n\bar{\kappa}} + n\right)\log\left(\frac{1}{\varepsilon}\right)\right)$$

first-order oracle calls to achieve $\mathbb{E}[\left\|x^t - x^*\right\|^2/\left\|x^0 - x^*\right\|^2] \leq \varepsilon$. □

Applying Lemma D.1, we can prove Theorem 4.3 straightforwardly.

Proof of Theorem 4.3. Consider the instance in the proof of Lemma D.1, but with $m \times n$ individual functions. Then it suggests the lower complexity bound

$$\Omega\left(\left(\sqrt{mn\bar{\kappa}} + mn\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

on incremental first-order oracle calls.

In decentralized optimization, we allocate the $m \times n$ individual functions $f_{i,j}$ for $i \in [m]$ and $j \in [n]$ in the proof of Lemma D.1 on a fully connected network with m nodes. Then it leads to the lower complexity bound

$$\Omega\left(\left(\sqrt{mn\bar{\kappa}} + mn\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

on local first-order oracle complexity.

□