Decentralized Strongly Convex Optimization with Less Local First-Order Oracle Complexity

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we study the decentralized optimization problem of minimizing the strongly convex objective that is the sum of smooth convex functions stored across a network with m local agents. We propose an efficient algorithm for finding an ϵ-suboptimal solution within at most $\mathcal{O}((m+\sqrt{m\kappa})\log(1/\epsilon))$ local first-order oracle calls and $\tilde{\mathcal{O}}(\sqrt{\kappa/\alpha}\log(1/\epsilon))$ communication rounds, where κ is the condition number of the objective and α is the spectral gap of the gossip matrix. Both of our local first-order oracle complexity and communication complexity nearly match the corresponding lower bounds. The proposed algorithm allows only few of the agents compute their local gradients during one iteration, which significantly reduces the total computational cost. In contrast, the existing decentralized convex optimization algorithms require all of the agents computes their local gradients during every iteration, which leads to at least $\Omega(m\sqrt{\kappa}\log(1/\epsilon))$ local first-order oracle complexity totally.

1 Introduction

In this paper, we focus on solving the decentralized optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x) \tag{1.1}$$

on an undirected connected network with m agents, where the objective function $f:\mathbb{R}^d\to\mathbb{R}$ is μ-strongly-convex, and the local function $f_i:\mathbb{R}^d\to\mathbb{R}$ on the i-th agent is L-smooth and convex. Decentralized algorithms desire all of the m agents to solve the optimization problem cooperatively and each of the agents is only allowed to communicate with its neighbors.

First-order algorithms for decentralized convex optimization have been extensively studied in recent years [11, 13, 15, 16, 19–21, 25, 27, 29, 31, 32, 37–39]. Scaman et al. [29] showed that achieving an ϵ-suboptimal solution of problem (1.1) requires at least $\Omega(\sqrt{\kappa}\log(1/\epsilon))$ gradient steps and $\Omega(\sqrt{\kappa/\alpha}\log(1/\epsilon))$ communication steps, where κ is the condition number of the objective and α is the spectral gap of the gossip matrix. They attempted to match these lower bounds by proposing multi-step dual accelerated (MSDA) method. However, the iteration of MSDA relies on accessing the dual gradients of local functions, which may be intractable. We are more interested in dual-free methods [7, 11, 12, 14, 32, 37] that only require the local gradient calls during the iterations. In particular, Kovalev et al. [11] applied the idea of primal dual framework [3, 4, 6, 18] and Chebyshev acceleration [17, 28] to design optimal proximal alternating predictor-corrector (OPAPC)

Table 1: We summarize local first-order oracle complexity and communication complexity of proposed KNOT and previous work. We use $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide logarithmic factors of κ and m.

Methods	# Local First-Order Oracle	# Communication
APM-C [15]	$\mathcal{O}\big(m\sqrt{\kappa}\log(1/\varepsilon)\big)$	$\mathcal{O}\big(\sqrt{\kappa/\alpha}\log^2(1/\varepsilon)\big)$
OPAPC [11]	$\mathcal{O}\big(m\sqrt{\kappa}\log(1/\varepsilon)\big)$	$\mathcal{O}\big(\sqrt{\kappa/\alpha}\log(1/\varepsilon)\big)$
Acc-GT+CA [13]	$\mathcal{O}\big(m\sqrt{\kappa}\log(1/\varepsilon)\big)$	$\mathcal{O}\big(\sqrt{\kappa/\alpha}\log(1/\varepsilon)\big)$
KNOT (Theorem 1)	$\mathcal{O}\big((m+\sqrt{m\kappa}\,)\log(1/\varepsilon)\big)$	$\tilde{\mathcal{O}}\big(\sqrt{\kappa/\alpha}\log(1/\varepsilon)\big)$
Lower Bounds [29, 34]	$\tilde{\Omega}\big(m+\sqrt{m\kappa}\log(1/\varepsilon)\big)$	$\Omega\big(\sqrt{\kappa/\alpha}\log(1/\varepsilon)\big)$

methods, which avoid dual gradient computation and the later one matches both of the lower bounds for gradient steps and communication steps provided by Scaman et al. [29]. Li and Lin [13] established the algorithm by incorporating gradient tracking [16, 21, 25, 31, 35, 36] and Chebyshev acceleration [17, 28] into Nesterov's acceleration [22] (Acc-GT+CA), which achieves the same computation and communication complexities.

We notice that the existing statements on optimality of gradient steps for above first-order decentralized algorithms can be refined [7, 11, 14, 37]. Concretely, the existing "optimal" first-order algorithms for decentralized convex optimization require all of the agents computing their local gradient during every iteration. Hence, each "gradient step" of these algorithm contains m local gradient calls, resulting at least $\Omega(m\sqrt{\kappa}\log(1/\epsilon))$ local gradient complexity in total.[1] In fact, the atomic operation of the first-order decentralized algorithm is computing the gradient of one local function, which implies allowing only few of the agents computing their local gradients during one iteration potentially makes the algorithm be more computation efficient. In practice, decentralized optimization are usually applied to networks with limited computational resources (e.g. mobile devices [33], wireless sensors [26] and smart home appliances [9]), which also encourages us to design decentralized algorithms with less local computational cost to reduce the energy consumption.

We consider the problem of minimizing objective function in problem (1.1) on a single machine. It is well known that accelerated stochastic gradient methods [1, 10, 24] can achieve an ϵ-suboptimal solution of such finite-sum problem within $\mathcal{O}((m+\sqrt{m\kappa}\,)\log(1/\epsilon))$ individual component gradient calls. Since the objective of decentralized optimization problem also has the finite-sum structure, it implies the known local gradient oracle complexity $\mathcal{O}(m\sqrt{\kappa}\log(1/\epsilon))$ of existing first-order algorithms [7, 11, 14, 32, 37] maybe not optimal. This naturally leads to the following question

Can we design a decentralized first-order algorithm with less local gradient calls?

In this paper, we give a positive answer to above question by proposing Katyusha-type Near-Optimal decenTralized algorithm (KNOT). Our method allows the agents to skip the step of computing local gradient during most of the iterations, which significantly improves the total computational efficiency. We prove that KNOT can achieve an ϵ-suboptimal solution of problem (1.1) within at most $\mathcal{O}((m+\sqrt{m\kappa}\,)\log(1/\epsilon))$ local gradient oracle complexity and $\tilde{\mathcal{O}}(\sqrt{\kappa/\alpha}\,\log(1/\epsilon))$ communication complexity, which nearly matches the corresponding lower bounds [29, 34]. We compare the theoretical results of proposed KNOT and previous methods in Table 1.

Paper Organization In section 2, we introduce the notations and settings throughout this paper. In section 3, we propose a new decentralized optimization algorithm and provide its convergence analysis. In section 4, we give a discussion for the optimality of proposed algorithm. In section 5, we provide numerical experiments to validate our theory. We conclude our work in section 6. All proofs are deferred to appendix.

[1] Recall that finding an ϵ-suboptimal solution of smooth and strongly convex function on single machine requires at least $\Omega(\sqrt{\kappa}\log(1/\epsilon))$ gradient calls in terms of the objective function [22].

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

1: $\mathbf{v}^{-1} = \mathbf{v}^0$
2: $\beta = \dfrac{1 - \sqrt{1 - \lambda_2^2(W)}}{1 + \sqrt{1 - \lambda_2^2(W)}}$
3: **for** $k = 0, \dots, K$
4: $\quad \mathbf{v}^{k+1} = (1+\beta)W\mathbf{v}^k - \beta\mathbf{v}^{k-1}$
5: **end for**
6: **Output:** $\mathbf{v}^K$

2 Preliminaries

We use $\|\cdot\|$ to present the Frobenius norm of the matrix and the Euclidean norm of the vector. We introduce the aggregated notations

$$\mathbf{x} = [x_1, \cdots, x_m]^\top, \quad \nabla F(\mathbf{x}) = [\nabla f_1(x_1), \cdots, \nabla f_m(x_m)]^\top \quad \text{and} \quad \bar{x} = \frac{1}{m}\mathbf{1}^\top \mathbf{x},$$

where x_i is the local variable on the i-th agent and $\nabla f_i(x_i)$ is the corresponding local gradient.

We impose the following assumptions on the decentralized optimization problem (1.1).

Assumption 1. *We assume each local function $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, i.e., there exists constant $L > 0$ such that*

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

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

Assumption 2. *We assume each local function $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex, i.e., we have*

$$f_i(y) - f_i(x) \ge \langle \nabla f_i(x), y - x\rangle$$

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

Assumption 3. *We assume the global function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, i.e., there exists constant $\mu > 0$ such that*

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

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

Assumption 1 implies $f(\cdot)$ is L-smooth and we define $\kappa \triangleq L/\mu$ as its condition number.

The strong convexity means the objective $f(\cdot)$ has unique minimizer x^*. We say $\hat{x} = [\hat{x}_1, \dots, \hat{x}_m]^\top$ is an ϵ-suboptimal solution of the decentralized optimization problem (1.1) if $f(\hat{x}_i) - f(x^*) \le \epsilon$ holds for any $i = 1, \dots, m$.

We let $W \in \mathbb{R}^{m \times m}$ be the gossip matrix associated with the network of m agents and it satisfies the following assumption.

Assumption 4. *We assume the gossip matrix W is symmetric and $W_{i,j} \ne 0$ if and only if the i-th and the j-th agents are connected in the network. We also assume W satisfies $\mathbf{0} \preceq W \preceq I, W\mathbf{1} = \mathbf{1}$ and $\mathrm{null}(I - W) = \mathrm{span}(\mathbf{1})$.*

We define the spectral gap of the gossip matrix as $\alpha \triangleq 1 - \lambda_2(W)$ which describes the connectivity of the network, where $\lambda_2(W)$ is the second largest eigenvalue of W. Each communication step can be performed by a multiplication of W by an aggregated variable. It is popular to reduce the consensus error by Chebyshev acceleration [17, 28]. We present the details in Algorithm 1 and its has the following property.

Proposition 1. *Let* $\mathbf{v}^0$ *and* $\mathbf{v}^t$ *be the input and output of Algorithm 1 respectively, and* $\bar{v} = \frac{1}{m}\mathbf{1}^\top \mathbf{v}^0$*. Then we have* $\bar{v} = \frac{1}{m}1^\top \mathbf{x}^K$ *and* $\|\mathbf{v}^t - \mathbf{1}\bar{v}\| \leq \left(1 - \sqrt{1-\lambda_2(W)}\right)^K \left\|\mathbf{v}^0 - \mathbf{1}\bar{v}\right\|$.

3 The Algorithm and Main Results

In this section, we introduce the insight and the design of Katyusha-type Near-Optimal decenTralized algorithm (KNOT). We also provide complexity analysis to show the advantage of KNOT formally.

3.1 Motivation

Before studying the decentralized optimization, we first give a brief review of the algorithms for solving the finite-sum optimization problem

$$\min_{\hat{x}\in\mathbb{R}^d} \hat{f}(\hat{x}) \triangleq \frac{1}{m}\sum_{i=1}^m \hat{f}_i(\hat{x}) \tag{3.1}$$

on a single machine, where the objective function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly-convex, and each individual component function $\hat{f}_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex. It is well known that accelerated gradient descent (AGD) [22] achieves the optimal full gradient complexity $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ for solving problem (3.1), while each of its iteration requires m incremental first-order oracle (IFO) calls. A popular way to reduce the iteration is stochastic gradient descent (SGD), while it only converges sublinearly. Variance reduction [1, 5, 8, 10, 23, 24, 30] is a widely used technique to improve the convergence rate of SGD, e.g. stochastic variance reduced gradient (SVRG) method iterate with gradient estimator

$$\hat{v}^t = \nabla\hat{f}(\hat{w}) + \frac{1}{b}\sum_{j\in\mathcal{S}_t}\left(\nabla\hat{f}_j(\hat{x}^t) - \nabla\hat{f}_j(\hat{w})\right), \tag{3.2}$$

where $\hat{w}$ is a snapshot point that is updated infrequently and $\mathcal{S}_t$ is a random subset of $\{1,\dots,m\}$ with candidate b. Katyusha method [1] iterates with variance reduced estimator $\hat{v}^t$ by involving the negative momentum and achieves the near optimal IFO complexity of $\mathcal{O}((m+\sqrt{m\kappa})\log(1/\varepsilon))$, which is better than $\mathcal{O}(m\sqrt{\kappa}\log(1/\varepsilon))$ of AGD.

Note that the objective function in decentralized optimization problem (1.1) also has the finite-sum structure, which motivates us to improve the computational efficiency by introducing some gradient estimator like variance reduced estimator shown in (3.2).

3.2 The Algorithm

We propose Katyusha-type Near-Optimal decenTralized algorithm (KNOT) in Algorithm 2. The gradient tracking steps (line 10 and 21) indicates the algorithm performs the following update

$$\begin{cases}
\bar{x}^t = \theta_1\bar{z}^t + \theta_2\bar{w}^t + (1-\theta_1-\theta_2)\bar{y}^t \\
\bar{s}^t = \bar{v}^t = \bar{u}^t + \dfrac{1}{m}\displaystyle\sum_{i=1}^m \frac{\xi_i^t}{q}\left(\nabla f_i(x_i^t) - \nabla f_i(w_i^t)\right) \\
\bar{z}^{t+1} = \dfrac{1}{1+\eta\sigma}\left(\eta\sigma\bar{x}^t + \bar{z}^t - \dfrac{\eta}{L}\bar{s}^t\right) \\
\bar{y}^{t+1} = \bar{x}^t + \theta_1\left(\bar{z}^{t+1} - \bar{z}^t\right) \\
\bar{w}^{t+1} = \begin{cases} \bar{y}^t, & \text{if } \zeta^{t+1} = 1 \\ \bar{w}^t, & \text{if } \zeta^{t+1} = 0 \end{cases} \\
\bar{u}^{t+1} = \bar{g}^{t+1} = \dfrac{1}{m}\displaystyle\sum_{i=1}^m f_i(w_i^{t+1})
\end{cases}$$

Algorithm 2 Katyusha-type Near-Optimal decenTralized algorithm (KNOT)

Input: initial point $\bar{w}^0$, probabilities p and q, number of consensus steps K and K_{out}, total iteration numbers T, parameters L, μ, θ_1 and θ_2

$\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{w}^0 = \mathbf{1}\bar{w}_0,\ \mathbf{v}^{-1} = \mathbf{s}^{-1} = \mathbf{0},\ \mathbf{g} = \mathbf{u}^0 = \nabla F(\mathbf{w}^0)$

$\eta = 1/(13\theta_1),\ \sigma = \mu/L$

for $t = 0, \ldots, T$

    $\mathbf{x}^t = \texttt{AccGossip}(\theta_1 \mathbf{z}^t + \theta_2 \mathbf{w}^t + (1 - \theta_1 - \theta_2)\mathbf{y}^t, K)$

    parallel for $i = 1, \ldots, m$ **do**

        draw $\xi_i^t \sim$ Bernoulli(q)

        $v_i^t = u_i^t + \dfrac{\xi_i^t}{q}\left(\nabla f_i(x_i^t) - \nabla f_i(w_i^t)\right)$

    end parallel for

    $\mathbf{s}^t = \texttt{AccGossip}\left(\mathbf{s}^{t-1} + \mathbf{v}^t - \mathbf{v}^{t-1}, K\right)$

    $\mathbf{z}^{t+1} = \texttt{AccGossip}\left(\dfrac{1}{1+\eta\sigma}\left(\eta\sigma\mathbf{x}^t + \mathbf{z}^t - \dfrac{\eta}{L}\mathbf{s}^t\right), K\right)$

    $\mathbf{y}^{t+1} = \texttt{AccGossip}\left(\mathbf{x}^t + \theta_1(\mathbf{z}^{t+1} - \mathbf{z}^t), K\right)$

    draw $\zeta^{t+1} \sim$ Bernoulli(p)

    parallel for $i = 1, \ldots, m$ **do**

        $\tilde{w}_i^{t+1} = \begin{cases} y_i^t, & \text{if } \zeta^{t+1} = 1 \\ w_i^t, & \text{if } \zeta^{t+1} = 0 \end{cases}$

    end parallel for

    $\mathbf{w}^{t+1} = \texttt{AccGossip}\left(\tilde{\mathbf{w}}^{t+1}, K\right)$

    parallel for $i = 1, \ldots, m$ **do**

        $g_i^{t+1} = \begin{cases} \nabla f_i(w_i^{t+1}), & \text{if } \zeta^{t+1} = 1 \\ g_i^t, & \text{if } \zeta^{t+1} = 0 \end{cases}$

    end parallel for

    $\mathbf{u}^{t+1} = \texttt{AccGossip}\left(\mathbf{u}^t + \mathbf{g}^{t+1} - \mathbf{g}^t, K\right)$

end for

Output: $\mathbf{x}_{\text{out}} = \texttt{AccGossip}(\mathbf{x}_T, K_{\text{out}})$.

on the mean vectors, which is similar to mini-batch version of Loopless Katyusha (`L-Katyusha`) [1, 10, 24]. The consensus error of the variables can be bounded by the property (Proposition 1) of the subroutine `AccGossip` (Algorithm 1), which encourages KNOT achieves the similar convergence result to Katyusha.

The computational efficiency of KNOT mainly comes from the local gradient estimator

$$v_i^t = u_i^t + \frac{\xi_i^t}{q}\left(\nabla f_i(x_i^t) - \nabla f_i(w_i^t)\right),$$

where u_i^t is an estimator of the local gradient at the snapshot point w_i^t and ξ_i^t is a random variable drawn from Bernoulli distribution with parameter q. If we set $q = b/m$ for some $b \in \{1, ..., m\}$, the mean of local gradient estimators $v_1^t, \ldots, v_m^t$ can be rewritten as

$$\bar{v}^t = \bar{u}^t + \frac{1}{b}\sum_{i \in \mathcal{I}^t}\left(\nabla f_i(x_i^t) - \nabla f_i(w_i^t)\right),$$

where $\mathcal{I}^t = \{i : \xi_i^t = 1\}$ and $\bar{u}^t$ can be regarded as an estimator of the global gradient at point $\bar{w}^t$. This implies the computations of local gradient $\nabla f_i(x_i^t)$ and $\nabla f_i(w_i^t)$ are performed on $|\mathcal{I}^t|$ agents with $\mathbb{E}\big[|\mathcal{I}^t|\big] = b$. Hence, the mean vector $\bar{v}^t$ plays a similar role to the variance reduced gradient estimator with batch-size $b = mq$, which leads to the algorithm requires less local gradient computation. Additionally, KNOT only computes the local gradient for all of the m agents when $\zeta^{t+1} = 1$ (line 19), which follows the idea of loopless framework for variance reduction [10, 24]. Since, we draw ζ^{t+1} from Bernoulli distribution with parameter p, the small p leads to $\zeta^{t+1} = 1$ occurs infrequently and the computational cost for this case is not expensive in expectation. KNOT also enjoys the parallel speed up property like Katyusha [1], which means to the appropriate settings for p and q can reduce the number of total iterations that corresponds to less communication rounds in total.

3.3 Convergence Analysis

The convergence analysis of KNOT (Algorithm 2) is based on the Lyapunov function [24] as follows

$$V^t \triangleq \mathcal{Z}^t + \mathcal{Y}^t + \mathcal{W}^t, \tag{3.3}$$

where

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

The parameters η, σ, θ_1, θ_2 in (3.3) follow the notations of Algorithm 2 and we let $\gamma \in (1/2, 1)$.

The decentralized setting leads to we cannot directly follow the analysis of existing Katyusha-type algorithms [1, 10, 24]. Different from previous work, the recursion on Lyapunov function V^t for KNOT contains the additional terms of consensus error, which is shown in the following lemma.

Lemma 1. *Under Assumption 1, 2, 3 and 4, if we choose $\eta = 1/(13\theta_1)$, Algorithm 2 holds that*

$$\begin{aligned}\mathbb{E}\left[V^{t+1}\right] \leq &\max\left(\frac{1}{1+\eta\sigma}, 1 - \left(\theta_1 + \theta_2 - \frac{\theta_2}{\gamma}\right), 1 - p(1-\gamma)\right)V^t \\ &+ \sqrt{\frac{2\eta L V^t}{(1+\eta\sigma)m}}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\| + \frac{L}{3m^2q\theta_1}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{L}{4m^2q\theta_1}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2.\end{aligned}$$

We consider the consensus error by introducing the vector

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

We describe the convergence of r_t by a linear system as follows.

Lemma 2. *Under the settings of Lemma 1, we run Algorithm 2 by taking*

$$K = \left\lceil \frac{\log(1/\rho)}{\sqrt{1-\lambda_2(W)}} \right\rceil$$

with

$$\rho \leq \frac{q\theta_1^3}{9}\min\left\{\frac{1}{2}\eta^2\sigma^2,\ \frac{1}{2}\left(\theta_1 + \theta_2 - \frac{\theta_2}{\gamma}\right),\ \frac{1}{2}p(1-\gamma)\right\}.$$

Then it holds that

$$\mathbb{E}\left[r^{t+1}\right] \leq \rho^2\left(\mathbf{B}\cdot\mathbf{A}\cdot r^t + e^t\right)$$

for some matrix $\mathbf{A} \in \mathbb{R}^{6\times 6}$, elementary matrix $\mathbf{B} \in \mathbb{R}^{6\times 6}$ and vector $e^t \in \mathbb{R}^6$ satisfy [2]

$$\|\mathbf{A}\| \leq \frac{4}{q\theta_1^2}, \qquad \|\mathbf{B}\| \leq 2 \qquad \textit{and} \qquad \left\|e^t\right\| < \frac{2}{3q\theta_1^2}\left(V^{t+1} + V^t\right).$$

[2] The expressions of $\mathbf{A}$, $\mathbf{B}$ and e_t are very complicated and we present them in appendix.

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

Theorem 1 (main result)**.** *Under Assumption 1, 2, 3 and 4, we run Algorithm 2 with*

$$p = \max\left\{\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{\kappa}}\right\}, \quad q = \min\left\{\frac{1}{\sqrt{m}}, \frac{\sqrt{\kappa}}{m}\right\}, \quad \gamma \in \left(\frac{2}{3}, 1\right)$$
$$\theta_2 = \frac{1}{2mq}, \qquad \theta_1 = \min\left\{\sqrt{\frac{mq}{\kappa p}}\theta_2, \theta_2\right\}, \qquad \eta = \frac{1}{13\theta_1}$$

and take K by following the setting of Lemma 2. Then it holds that

$$\mathbb{E}\left[V^t\right] \leq \left(1 - \min\left\{\eta\sigma, \frac{\theta_1 + \theta_2 - \theta_2/\gamma}{2}, \frac{p(1-\gamma)}{2}\right\}\right)^t \left(V^0 + \|\mathbf{r}^0\|\right)$$

and

$$\mathbb{E}\left[\frac{L}{m}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2\right] \leq \left(\frac{8}{243} + 2^{-t}\right)\left(1 - \min\left\{\eta\sigma, \frac{\theta_1 + \theta_2 - \theta_2/\gamma}{2}, \frac{p(1-\gamma)}{2}\right\}\right)^t \cdot \left(V^0 + \|\mathbf{r}^0\|\right).$$

Theorem 1 establish the linear convergence for the function value at the point of mean vectors. The Bernoulli variables in the algorithm indicate each iteration has $\mathcal{O}(m/\sqrt{\kappa} + \sqrt{m})$ local gradient calls in expectation. Hence, we obtain the upper bounds of local gradient oracle complexity and communication complexity for finding an ϵ-suboptimal solution.

Corollary 2. *Under the settings of Theorem 1, Algorithm 2 can achieve an ϵ-suboptimal solution by taking $T = \mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$ and $K_{\text{out}} = \tilde{\mathcal{O}}(\sqrt{1/\alpha})$, which requires the local first-order oracle complexity of $\mathcal{O}\left((m + \sqrt{m\kappa})\log(1/\epsilon)\right)$ and the communication complexity of $\tilde{\mathcal{O}}(\sqrt{\kappa/\alpha}\log(1/\epsilon))$ in expectation.*

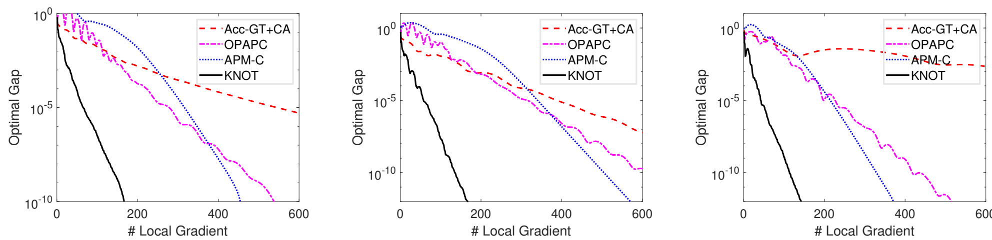
4 Discussion for the Optimality

In this section, we verify the optimality of the proposed algorithms. We first provide follow the statement for the lower bounds of decentralized strongly convex optimization provided by Kovalev et al. [11], which is a direct application of Corollary 2 from Scaman et al. [29] but does not include the dual gradient oracle.

Proposition 2. *For any $m \geq 2$ and $\alpha > 1$, there exist a gossip matrix $W \in \mathbb{R}^{m\times m}$ satisfying $1 - \lambda_2(W) = \alpha$ and a family of smooth strongly convex functions $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^m$ with condition number κ such that the following holds: for any $\epsilon > 0$, any first-order decentralized algorithm requires at least $\Omega(\sqrt{\kappa/\alpha}\log(1/\epsilon))$ communication rounds and at least $\Omega(\sqrt{\kappa}\log(1/\epsilon))$ gradient steps to output $\mathbf{x} = [x_1, \dots, x_m]^\top$ such that $f(x_i) - f(x^*) \leq \epsilon$ for all $i = 1, \dots, m$, where x^* is the minimizer of $f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x)$.*

It is worth pointing out that the concept "gradient step" described in Proposition 2 only requires the gradient computation should depend on the history local points of the corresponding agent, while it does not contain any requirement on the number of agents that participate into their local gradients computation. This implies the "gradient steps" lower bound of $\Omega(\sqrt{\kappa}\log(1/\epsilon))$ described in this proposition corresponds to the iterations number of proposed KNOT (Algorithm 2), rather than the number of local gradient calls. Hence, the result of Corollary 2 means KNOT matches the "gradient steps" lower bound and nearly matches the communication lower bound provided by Proposition 2.

Compared with the number of "gradient steps", we are more interested in the number of local gradient calls, which essentially reflects the totally computational cost of a decentralized optimization algorithm. The lower bound of local gradient calls can be established by considering the IFO calls for the finite-sum optimization problem on single machine. Woodworth and Srebro [34] provide the following lower bound for solving the finite-sum optimization problem by randomized first-order (non-distributed) methods.



(a) a9a (b) german.numer (c) australian

Figure 1: Comparison for the number of local gradient calls vs. optimal gap.

Proposition 3. *For any $m \geq 2$ and $\kappa > 161m$, there exist a family of smooth strongly convex functions $\{\hat{f}_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^m$ with condition number κ such that the following holds: for any $\epsilon > 0$, any randomized algorithm require at least $\tilde{\Omega}\left(m + \sqrt{m\kappa}\log(1/\epsilon)\right)$ IFO calls to output x such that $\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] < \epsilon$, where x^* is the minimizer of $\hat{f}(x) = \frac{1}{m}\sum_{i=1}^m \hat{f}_i(x)$.*

We can view the individual functions $\{\hat{f}_i\}_{i=1}^m$ in Proposition 3 as the local functions in decentralized optimization on a fully connected network, then the IFO lower bound $\tilde{\Omega}\left(m + \sqrt{m\kappa}\log(1/\epsilon)\right)$ just corresponds to the local gradient lower bound in our decentralized optimization problem. Hence, Proposition 3 implies the local gradient oracle complexity of proposed KNOT is near optimal.

5 Experiments

In this section, we provide the numerical experiments to evaluate the performance of proposed KNOT. We consider the ℓ_2-regularized logistic regression for binary classification. We formulate this model by the optimization problem

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

with

$$f_i(x) = \frac{1}{n}\sum_{j=1}^n \log\left(1 + \exp\left(-b_{ij} a_{ij}^\top x\right)\right) + \frac{\mu}{2}\|x\|^2,$$

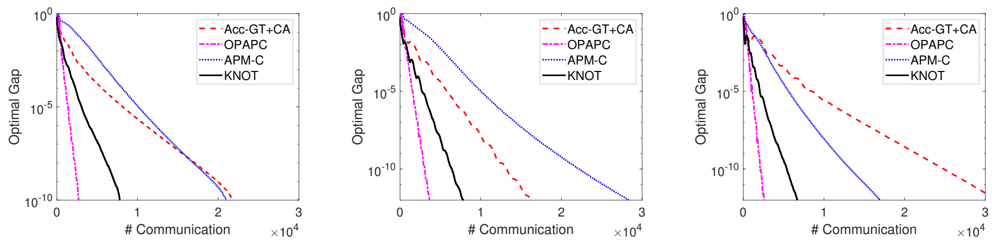
where $a_{ij} \in \mathbb{R}^d$ is the feature vector of the j-th sample on agent i, $b_{ij} \in \{-1, 1\}$ is the corresponding label and $\mu > 0$ is the hyperparamter.

We conduct our experiments on three real-world datasets "a9a", "german.numer" and "australian" which can be found in LIBSVM repository [2]. We let $m = 300$ and $\mu = 0.01$. We set the gossip matrix W by corresponding a random graph that each pair in the network is connected with probability $1/30$, which leads to $1 - \lambda_2(W) \approx 0.0382$.

We compare the proposed method KNOT with baseline algorithms ACC-GT+CA [13], OPAPC [11] and APM-C [15]. For KNOT, we set the parameters p, q, θ_1, θ_2 and η by following the settings of Theorem 1 and tune K from $\{1, 5, 10\}$. For the baseline algorithms, we also select their parameters by following the corresponding theoretical analysis.

We present the experimental results for the computational cost and communication cost in Figures 1 and 2 respectively, where the y-axis represents the optimal gap which is defined as

$$\frac{1}{m}\sum_{i=1}^m f(x_i) - f(x^*).$$



(a) a9a (b) german.numer (c) australian

Figure 2: Comparison for the number of communication rounds vs. optimal gap.

We observe that the proposed KNOT always has significantly better computational efficiency than than all of baseline methods. For the communication complexity, the result of KNOT is comparable to OPAPC and better than other baseline methods.

6 Conclusion

In this paper, we study decentralized strongly convex optimization and propose a novel method called Katyusha-type Near-Optimal decenTralized algorithm (KNOT), which avoids computing all of the local gradients in one iteration. The theoretical analysis shows that our method is near optimal to both the local first-order oracle complexity and the communication complexity. The empirical studies on regularized logistic regression problem also supports our theoretical results. We believe the idea of KNOT is not limited to first-order optimization for convex problems. It is possible to extend the framework of KNOT to solve variational inequalities. We can also try to design the second-order decentralized algorithms with less local Hessian calls.

References

[1] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *STOC*, 2017.

[2] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011. URL `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

[3] Peijun Chen, Jianguo Huang, and Xiaoqun Zhang. A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration. *Inverse Problems*, 29 (2):025011, 2013.

[4] Laurent Condat, Daichi Kitahara, Andrés Contreras, and Akira Hirabayashi. Proximal splitting algorithms: Relax them all. *arXiv preprint arXiv:1912.00137*, 2019.

[5] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.

[6] Yoel Drori, Shoham Sabach, and Marc Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.

[7] Darina Dvinskikh and Alexander Gasnikov. Decentralized and parallel primal and dual accelerated methods for stochastic convex programming problems. *Journal of Inverse and Ill-posed Problems*, 29(3):385–405, 2021.

[8] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

[9] Il-Young Joo and Dae-Hyun Choi. Distributed optimization framework for energy management of multiple smart homes with distributed energy resources. *IEEE Access*, 5:15551–15560, 2017.

[10] Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *ALT*, 2020.

[11] Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *NeurIPS*, 2020.

[12] Huan Li and Zhouchen Lin. Revisiting EXTRA for smooth distributed optimization. *SIAM Journal on Optimization*, 30(3):1795–1821, 2020.

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

[14] Huan Li, Cong Fang, Wotao Yin, and Zhouchen Lin. A sharp convergence rate analysis for distributed accelerated gradient methods. *arXiv preprint arXiv:1810.01053*, 2018.

[15] Huan Li, Cong Fang, Wotao Yin, and Zhouchen Lin. Decentralized accelerated gradient methods with increasing penalty parameters. *IEEE transactions on Signal Processing*, 68:4855–4870, 2020.

[16] Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17):4494–4506, 2019.

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

[18] Ignace Loris and Caroline Verhoeven. On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. *Inverse Problems*, 27(12):125007, 2011.

[19] Angelia Nedic. Asynchronous broadcast-based convex optimization over a network. *IEEE Transactions on Automatic Control*, 56(6):1337–1351, 2010.

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

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

[22] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[23] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

[24] Xun Qian, Zheng Qu, and Peter Richtárik. L-SVRG and L-Katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22(1):4991–5039, 2021.

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

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

[27] S. Sundhar Ram, A. Nedich, and Venugopal V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147:516–545, 2010.

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

[29] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *ICML*, 2017.

[30] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

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

[32] Zhuoqing Song, Lei Shi, Shi Pu, and Ming Yan. Optimal gradient tracking for decentralized optimization. *arXiv preprint arXiv:2110.05282*, 2021.

[33] Mu Wang, Changqiao Xu, Xingyan Chen, Lujie Zhong, Zhonghui Wu, and Dapeng Oliver Wu. Bc-mobile device cloud: A blockchain-based decentralized truthful framework for mobile device cloud. *IEEE Transactions on Industrial Informatics*, 17(2):1208–1219, 2020.

[34] Blake E. Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *NIPS*, 2016.

[35] Ran Xin and Usman A. Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.

[36] Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *CDC*, 2015.

[37] Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *arXiv preprint arXiv:2005.00797*, 2020.

[38] Haishan Ye, Ziang Zhou, Luo Luo, and Tong Zhang. Decentralized accelerated proximal gradient descent. In *NeurIPS*, 2020.

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

The theoretical analysis of KNOT is organized as follows.

- We first consider the mean vector and provide the convergence of $f(\bar{x}^t) - f(x^*)$. The result is shown in Lemma 1, and its detailed proof is shown in Appendix A.
- We then consider the consensus error, which is characterized by vector
$r^t = \frac{L}{m}\left[\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\|^2, \frac{\eta^2}{L^2}\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\|^2, \frac{\eta^2}{L^2}\left\|\mathbf{s}^t - \mathbf{1}\bar{s}^t\right\|^2, \|\mathbf{z}^t - \mathbf{1}\bar{z}^t\|^2, \|\mathbf{w}^t - \mathbf{1}\bar{w}^t\|^2, \|\mathbf{y}^t - \mathbf{1}\bar{y}^t\|^2\right]^\top.$
The recursion of r_t is established in Lemma 2, and its detailed proof is shown in Appendix B. Especially, we present the expression of $\mathbf{A}$, $\mathbf{B}$ and e_t in the statement of Lemma 15.
- We finally apply Lemma 1 and 2 to obtain our final results Theorem 1 and Corollary 2, whose detailed proofs are shown in Appendix C and D respectively.

A Proof of Lemma 1

In this section, we focus on analyzing Lyapunov function

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

where

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

The analysis is more complicated than the counterpart of L-Katyusha [10, 24] because of the consensus error aroused from the decentralized setting.

We substitute $q = b/m$ in the following proof, where b can be regarded as expected mini-batch size.

Let us begin with a useful lemma of L-smooth and convex functions for our further analysis.

Lemma 3 ([22])**.** *Under Assumption 1 and 3, it holds that*

$$\frac{1}{2L}\left\|\nabla f_i(x) - \nabla f_i(y)\right\|^2 \le f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle, \tag{A.1}$$

for all i and x, y.

We show that the average of local gradient trackers can approximate $\nabla f(\bar{x}^t)$ well first.

Lemma 4. *Under the settings of Lemma 1, Algorithm 2 holds that*

$$\bar{s}^t = \frac{1}{m}\sum_{i}^{m} v_i^t \qquad \textit{and} \qquad \mathbb{E}[\bar{s}^t] = \frac{1}{m}\sum_{i}^{m}\nabla f_i(\mathbf{x}_i^t).$$

Furthermore, we have

$$\left\|\nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t]\right\| \le \frac{L}{\sqrt{m}}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|.$$

Proof. We have

$$\begin{aligned}
\left\|\nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t]\right\|^2 &= \left\|\frac{1}{m}\sum_{i=1}^{m}\left(\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\bar{x}^t)\right)\right\|^2 \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla f_i(x_i^t) - \nabla f_i(\bar{x}^t)\right\|^2 \\
&\le \frac{L^2}{m}\sum_{i=1}^{m}\left\|x_i^t - \bar{x}^t\right\|^2 \\
&= \frac{L^2}{m}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2,
\end{aligned}$$

where the first equality is due to $\bar{s}^t = \bar{v}^t$ and

$$\mathbb{E}[\bar{v}^t] = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_i^t)$$

hold for Algorithm 2. $\square$

Then we provide some lemmas for the mean vectors.

Lemma 5. *Under the settings of Lemma 1, it holds that*

$$\begin{aligned}\mathbb{E}\left[\|\bar{s}^t - \nabla f(\bar{x}^t)\|^2\right] \le & \frac{12L}{b}\left(f(\bar{w}^t) - f(\bar{x}^t) - \left\langle \nabla f(\bar{x}^t), \bar{w}^t - \bar{x}^t \right\rangle\right) + \frac{8L^2}{mb}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 \\ & + \frac{6L^2}{mb}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2.\end{aligned} \tag{A.2}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{s}^t - \nabla f(\bar{x}^t)^2\|^2\right] \\
\overset{\text{Alg. 2}}{=} & \mathbb{E}\left[\|\bar{u}^t + \frac{1}{qm}\sum_{j=1}^{m}\xi_j\left(\nabla f_j(x_j^t) - \nabla f_j(w_j^t)\right) - \nabla f(\bar{x}^t)\|^2\right] \\
= & \mathbb{E}\left[\|\frac{1}{m}\sum_{j=1}^{m}\frac{\xi_j}{q}(\nabla f_j(x_j^t) - \nabla f_j(w_j^t)) - \mathbb{E}\left[\nabla f_j(x_j^t) - \nabla f_j(w_j^t)\right] + (\mathbb{E}[\bar{s}^t] - \nabla f(\bar{x}^t))\|^2\right] \\
\le & \frac{2}{m}\mathbb{E}\left[\|\frac{\xi_j}{q}(\nabla f_j(x_j^t) - \nabla f_j(w_j^t))\|^2\right] + 2\mathbb{E}\left[\|\mathbb{E}[\bar{s}^t] - \nabla f(\bar{x}^t))\|^2\right] \\
= & \frac{2}{mq}\mathbb{E}\left[\|\nabla f_j(x_j^t) - \nabla f_j(w_j^t)\|^2\right] + 2\mathbb{E}\left[\|\mathbb{E}[\bar{s}^t] - \nabla f(\bar{x}^t))\|^2\right] \\
\le & \frac{6}{mq}\mathbb{E}\left[\left\|\nabla f_j(x_j^t) - \nabla f_j(\bar{x}^t)\right\|^2 + \left\|\nabla f_j(\bar{x}^t) - \nabla f_j(\bar{w}^t)\right\|^2 + \left\|\nabla f_j(\bar{w}^t) - \nabla f_j(w_j^t)\right\|^2\right] \\
& + 2\mathbb{E}\left[\|\mathbb{E}[\bar{s}^t] - \nabla f(\bar{x}^t))\|^2\right] \\
\overset{3}{\le} & \frac{12L}{mq}\left(f(\bar{w}^t) - f(\bar{x}^t) - \left\langle \nabla f(\bar{x}^t), \bar{w}^t - \bar{x}^t \right\rangle\right) + \frac{8L^2}{m^2 q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{6L^2}{m^2 q}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
= & \frac{12L}{b}\left(f(\bar{w}^t) - f(\bar{x}^t) - \left\langle \nabla f(\bar{x}^t), \bar{w}^t - \bar{x}^t \right\rangle\right) + \frac{8L^2}{mb}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{6L^2}{mb}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2,
\end{aligned}$$

where the first equality is because of the fact that $\bar{u}^t = \mathbb{E}\left[\nabla f_j(x_j^t)\right]$ and the first inequality is because of the fact that $\mathbb{E}[\|z - \mathbb{E}[z]\|^2] \le \mathbb{E}[\|z\|^2]$ and the property of variance; the last inequality is because of Lemma 4.

$\square$

Lemma 6. *Under the settings of Lemma 1, we have*

$$\left\langle \bar{s}^t, x^* - \bar{z}^{t+1} \right\rangle + \frac{\mu}{2}\left\|\bar{x}^t - x^*\right\|^2 \ge \tfrac{L}{2\eta}\left\|\bar{z}^t - \bar{z}^{t+1}\right\|^2 + \mathcal{Z}^{t+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^t. \tag{A.3}$$

Proof. We start with the definition of $\mathbf{z}^{t+1}$

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

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

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

□

Lemma 7. *Under the settings of Lemma 1, we have*

$$\frac{1}{\theta_1} \left(f(\bar{y}^{t+1}) - f(\bar{x}^t) \right) - \frac{1}{24L\theta_1} \left\| \bar{s}^t - \nabla f(\bar{x}^t) \right\|^2 \leq \frac{L}{2\eta} \left\| \bar{z}^{t+1} - \bar{z}^t \right\|^2 + \left\langle \bar{s}^t, z^{t+1} - \bar{z}^t \right\rangle . \quad \text{(A.4)}$$

Proof. We have

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

where the last inequality uses the 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 8. *Under the settings of Lemma 1, we have*

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

Proof. From Algorithm 2, we know that

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

Then from the definition of $\mathcal{W}^t$ and $\mathcal{Y}^t$, the lemma naturally holds. □

Using the above lemmas, we prove Lemma 1 as follows.

Proof for Lemma 1. Combining Lemma 4, 5, 6, 7 and 8, we obtain

$$
\begin{aligned}
f(x^*) \overset{\text{Asm. 3}}{\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 \\
\overset{\text{Alg. 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 \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 \\
\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[\frac{\mu}{2} \left\| \bar{x}^t - x^* \right\|^2 + \left\langle \bar{s}^t, x^* - z^{t+1} \right\rangle + \left\langle \bar{s}^t, z^{t+1} - z^t \right\rangle \right] + \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle \\
\overset{(A.3)}{\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, z^{t+1} - z^t \right\rangle + \frac{L}{2\eta} \left\| z^t - 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{(A.4)}{\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{(A.2)}{\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{L}{3mb\theta_1} \left\| \mathbf{x}^t - \mathbf{1}\bar{x}^t \right\|^2 - \frac{L}{4mb\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{L}{3mb\theta_1} \left\| \mathbf{x}^t - \mathbf{1}\bar{x}^t \right\|^2 - \frac{L}{4mb\theta_1} \left\| \mathbf{w}^t - \mathbf{1}\bar{w}^t \right\|^2,
\end{aligned}
$$

where in the second inequality we use the convexity of $f(\cdot)$. The procedure of the algorithm means

$$
\begin{aligned}
\mathbf{x}^t & \overset{\text{Alg. 2}}{=} \theta_1 \mathbf{z}^t + \theta_2 \mathbf{w}^t + (1-\theta_1-\theta_2)\mathbf{y}^t, \\
\mathbf{z}^t - \mathbf{x}^t & \overset{\text{Alg. 2}}{=} \frac{\theta_2}{\theta_1}(\mathbf{x}^t - \mathbf{w}^t) + \frac{1-\theta_1-\theta_2}{\theta_1}(\mathbf{x}^t - \mathbf{y}^t).
\end{aligned}
$$

Combining above results, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{Z}^{t+1} + \mathcal{Y}^{t+1} \right] \leq & \frac{1}{1+\eta\sigma} \mathcal{Z}^t + (1-\theta_1-\theta_2)\mathcal{Y}^t + \frac{\theta_2}{\theta_1} (f(\bar{w}^t) - f^*) \\
& - \left\langle \nabla f(\bar{x}^t) - \mathbb{E}[\bar{s}^t], x^* - \bar{z}^t \right\rangle + \frac{L}{3mb\theta_1} \left\| \mathbf{x}^t - \mathbf{1}\bar{x}^t \right\|^2 + \frac{L}{4mb\theta_1} \left\| \mathbf{w}^t - \mathbf{1}\bar{w}^t \right\|^2.
\end{aligned}
$$

Using definition of $\mathcal{W}^t$, we get

$$\begin{aligned}\mathbb{E}\left[\mathcal{Z}^{t+1}+\mathcal{Y}^{t+1}\right] \leq& \frac{1}{1+\eta\sigma}\mathcal{Z}^t+(1-\theta_1-\theta_2)\mathcal{Y}^t+p\gamma\mathcal{W}^t\\ &+\left\|\nabla f(\bar{x}^t)-\mathbb{E}[\bar{s}^t]\right\|\left\|x^*-\bar{z}^t\right\|+\frac{L}{3mb\theta_1}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\frac{L}{4mb\theta_1}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\\ \overset{(4)}{\leq}& \frac{1}{1+\eta\sigma}\mathcal{Z}^t+(1-\theta_1-\theta_2)\mathcal{Y}^t+p\gamma\mathcal{W}^t\\ &+\frac{L}{\sqrt{m}}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|\left\|x^*-\bar{z}^t\right\|+\frac{L}{3mb\theta_1}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\frac{L}{4mb\theta_1}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2.\end{aligned}$$

Finally, we use Lemma 8 to achieve

$$\begin{aligned}&\mathbb{E}\left[\mathcal{Z}^{t+1}+\mathcal{Y}^{t+1}+\mathcal{W}^{t+1}\right]\\ \leq&\frac{1}{1+\eta\sigma}\mathcal{Z}^t+(1-\theta_1-\theta_2)\mathcal{Y}^t+p\gamma\mathcal{W}^t+(1-p)\mathcal{W}^t+\frac{\theta_2}{\gamma}\mathcal{Y}^t\\ &+\frac{L}{\sqrt{m}}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|\left\|x^*-\bar{z}^t\right\|+\frac{L}{3mb\theta_1}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\frac{L}{4mb\theta_1}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\\ =&\frac{1}{1+\eta\sigma}\mathcal{Z}^t+\left(1-\left(\theta_1+\theta_2-\frac{\theta_2}{\gamma}\right)\right)\mathcal{Y}^t+\left(1-p(1-\gamma)\right)\mathcal{W}^t\\ &+\frac{L}{\sqrt{m}}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|\left\|x^*-\bar{z}^t\right\|+\frac{L}{3mb\theta_1}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\frac{L}{4mb\theta_1}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\\ \leq&\frac{1}{1+\eta\sigma}\mathcal{Z}^t+\left(1-\left(\theta_1+\theta_2-\frac{\theta_2}{\gamma}\right)\right)\mathcal{Y}^t+\left(1-p(1-\gamma)\right)\mathcal{W}^t\\ &+\sqrt{\frac{2\eta LV^t}{(1+\eta\sigma)m}}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|+\frac{L}{3mb\theta_1}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\frac{L}{4mb\theta_1}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2,\end{aligned}$$

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

B Proof of Lemma 2

We first provide some lemmas to bound the consensus error.

Lemma 9. *Letting*

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

then under the settings of Lemma 2, we have

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

where $\rho>0$ is the parameter such that $K=\dfrac{\log(1/\rho)}{\sqrt{1-\lambda_2(W)}}$.

Lemma 10. *Under the settings of Lemma 2, it holds that*

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

Proof. From Lemma 1, we have that

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2\right] &\leq \mathbb{E}\left[\rho^2\left\|\tilde{\mathbf{w}}^{t+1}-\mathbf{1}\bar{\tilde{w}}^{t+1}\right\|^2\right] \\ &\overset{\text{Alg. 2}}{=} \rho^2 p\left\|\mathbf{y}^t-\mathbf{1}\bar{y}^t\right\|^2+\rho^2(1-p)\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2.\end{aligned}$$

□

Then we provide upper bound for $\|\mathbf{g}^{t+1}-\mathbf{g}^t\|^2$ and $\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2$.

Lemma 11. *Under the settings of Lemma 2, it holds that*

$$\begin{aligned}\left\|\mathbf{g}^{t+1}-\mathbf{g}^t\right\|^2 \leq& 4L^2\left\|\mathbf{w}^{t+1}-\mathbf{1}\bar{w}^{t+1}\right\|^2+\frac{8Lmp\gamma\theta_1}{\theta_2}\mathcal{W}^{t+1} \\ &+4L^2\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2+\frac{8Lmp\gamma\theta_1}{\theta_2}\mathcal{W}^t.\end{aligned}$$

Proof. We have

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

We can also obtain that

$$\begin{aligned}\left\|\nabla f_i(w_i^t)-\nabla f_i(x^*)\right\|^2 =& \left\|\nabla f_i(w_i^t)-\nabla f_i(\bar{w}^t)+\nabla f_i(\bar{w}^t)-\nabla f_i(x^*)\right\|^2 \\ \leq& 2\left\|\nabla f_i(w_i^t)-\nabla f_i(\bar{w}^t)\right\|^2+2\left\|\nabla f_i(\bar{w}^t)-\nabla f_i(x^*)\right\|^2 \\ \overset{(1)}{\leq}& 2L^2\left\|w_i^t-\bar{w}^t\right\|^2+2\left\|\nabla f_i(\bar{w}^t)-\nabla f_i(x^*)\right\|^2 \\ \overset{(2)}{\leq}& 2L^2\left\|w_i^t-\bar{w}^t\right\|^2+4L(f_i(\bar{w}^t)-f_i(x^*)).\end{aligned}$$

Combining above results, we achieve

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

□

Next, we target to bound $\|\mathbf{v}^t - \mathbf{1}\bar{v}^t\|^2$. We first give two auxiliary lemmas.

Lemma 12. *Under the settings of Lemma 2, it holds that*

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

Proof. We have

$$\begin{aligned}
\sum_{i=1}^m \left\|u_i^t\right\|^2 &= \sum_{i=1}^m \left\|(u_i^t - \bar{u}^t) + (\bar{u}^t - \nabla f(\bar{w}^t)) + (\nabla f(\bar{w}^t) - \nabla f(x^*))\right\|^2 \\
&\le \sum_{i=1}^m \left[3\left\|u_i^t - \bar{u}^t\right\|^2 + 3\left\|\bar{u}^t - \nabla f(\bar{w}^t)\right\|^2 + 3\left\|\nabla f(\bar{w}^t) - \nabla f(x^*)\right\|^2\right] \\
&\overset{(A.1)}{\le} 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + 3\sum_{i=1}^m \left\|\frac{1}{m}\sum_{j=1}^m (\nabla f_j(w_j^t) - \nabla f_j(\bar{w}^t))\right\|^2 + 3\sum_{i=1}^m (2L(f_i(\bar{w}^t) - f_i(x^*)) \\
&\le 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{3}{m}\sum_{i=1}^m\sum_{j=1}^m \left\|\nabla f_j(w_j^t) - \nabla f_j(\bar{w}^t)\right\|^2 + 3\sum_{i=1}^m (2L(f_i(\bar{w}^t) - f_i(x^*)) \\
&\overset{\text{Asm. 1}}{\le} 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{3L^2}{m}\sum_{i=1}^m\sum_{j=1}^m \left\|w_j^t - \bar{w}^t\right\|^2 + 3\sum_{i=1}^m (2L(f_i(\bar{w}^t) - f_i(x^*)) \\
&= 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + 3L^2\left\|w^t - \mathbf{1}\bar{w}^t\right\|^2 + \frac{6Lmp\gamma\theta_1}{\theta_2}\mathcal{W}^t.
\end{aligned}$$

□

Lemma 13. *Under the settings of Lemma 2, it holds that*

$$\begin{aligned}
&\sum_{i=1}^m \mathbb{E}\left[\left\|\frac{\xi_i}{q}(\nabla f_i(x_i^t) - \nabla f_i(w_i^t))\right\|^2\right] \\
\le&\frac{8\eta mL\theta_1}{(1+\eta\sigma)q}\mathcal{Z}^t + \frac{8mLp\gamma\theta_1}{q\theta_2}(1+\theta_2)\mathcal{W}^t + \frac{8mL\theta_1}{q}(1-\theta_1-\theta_2)\mathcal{Y}^t \\
&+ \frac{4L^2}{q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{4L^2}{q}\left\|w^t - \mathbf{1}\bar{w}^t\right\|^2.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\sum_{i=1}^m \mathbb{E}\left[\left\|\frac{\xi_i}{q}(\nabla f_i(x_i^t) - \nabla f_i(w_i^t))\right\|^2\right] \\
=&\frac{1}{q}\sum_{i=1}^m \left\|\nabla f_i(x_i^t) - \nabla f_i(w_i^t)\right\|^2 \\
\le&\frac{4}{q}\sum_{i=1}^m \left[\left\|\nabla f_i(x_i^t) - \nabla f_i(\bar{x}^t)\right\|^2 + \left\|\nabla f_i(\bar{x}^t) - \nabla f_i(x^*)\right\|^2 + \left\|\nabla f_i(x^*) - \nabla f_i(\bar{w}^t)\right\|^2 + \left\|\nabla f_i(\bar{w}^t) - \nabla f_i(w_i^t)\right\|^2\right] \\
\overset{\text{Asm. 1}}{\le}&\frac{4L^2}{q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{8mL}{q}(f(\bar{x}^t) - f(x^*)) + \frac{8mL}{q}(f(\bar{w}^t) - f(x^*)) + \frac{4L^2}{q}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
\overset{\text{Asm. 2}}{\le}&\frac{8mL}{q}(\theta_1(f(\bar{z}^t) - f(x^*)) + \theta_2(f(\bar{w}^t) - f(x^*)) + (1-\theta_1-\theta_2)(f(\bar{z}^t) - f(x^*))) \\
&+ \frac{4L^2}{q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{8mL}{q}(f(\bar{w}^t) - f(x^*)) + \frac{4L^2}{q}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
\overset{\text{Asm. 1}}{\le}&\frac{4mL^2\theta_1}{q}\left\|\bar{z}^t - x^*\right\|^2 + \frac{8mL}{q}(1+\theta_2)(f(\bar{w}^t) - f(x^*)) + \frac{8mL}{q}(1-\theta_1-\theta_2)(f(\bar{y}^t) - f(x^*)) \\
&+ \frac{4L^2}{q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{4L^2}{q}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2
\end{aligned}$$

$$
\begin{aligned}
=&\frac{8\eta mL\theta_1}{(1+\eta\sigma)q}\mathcal{Z}^t + \frac{8mLp\gamma\theta_1}{q\theta_2}(1+\theta_2)\mathcal{W}^t + \frac{8mL\theta_1}{q}(1-\theta_1-\theta_2)\mathcal{Y}^t \\
&+ \frac{4L^2}{q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + \frac{4L^2}{q}\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2.
\end{aligned}
$$

□

Now we are ready to bound $\|\mathbf{v}^t - \mathbf{1}\bar{v}^t\|^2$.

Lemma 14. *Under the settings of Lemma 2, it holds that*

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{v}^{t+1} - \mathbf{v}^t\right\|^2\right] \\
\le&\frac{4Lmp\gamma\theta_1}{\theta_2}\left(\frac{4(1+\theta_2)}{q} + 3\right)(\mathcal{W}^{t+1} + \mathcal{W}^t) + \frac{16\eta mL\theta_1}{(1+\eta\sigma)q}(\mathcal{Z}^{t+1} + \mathcal{Z}^t) \\
&+ \frac{16mL\theta_1}{q}(1-\theta_1-\theta_2)(\mathcal{Y}^{t+1} + \mathcal{Y}^t) \\
&+ 6\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{8L^2}{q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + 2L^2\left(\frac{4}{q}+3\right)\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2 \\
&+ 6\left\|\mathbf{u}^{t+1} - \mathbf{1}\bar{u}^{t+1}\right\|^2 + \frac{8L^2}{q}\left\|\mathbf{x}^{t+1} - \mathbf{1}\bar{x}^{t+1}\right\|^2 + 2L^2\left(\frac{4}{q}+3\right)\left\|\mathbf{w}^{t+1} - \mathbf{1}\bar{w}^{t+1}\right\|^2.
\end{aligned}
$$

Proof. It holds that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2\right] =& \sum_{i=1}^m \mathbb{E}\left[\left\|u_i^t + \frac{\xi_i}{q}(\nabla f_i(x_i^t) - \nabla f_i(w_i^t))\right\|^2\right] \\
=& \sum_{i=1}^m \mathbb{E}\left[\left\|u_i^t\right\|^2 + \left\|\frac{\xi_i}{q}(\nabla f_i(x_i^t) - \nabla f_i(w_i^t))\right\|^2\right] \\
\overset{(12),(13)}{\le}& \frac{2Lmp\gamma\theta_1}{\theta_2}\left(\frac{4(1+\theta_2)}{q} + 3\right)\mathcal{W}^t + \frac{8\eta mL\theta_1}{(1+\eta\sigma)q}\mathcal{Z}^t + \frac{8mL\theta_1}{q}(1-\theta_1-\theta_2)\mathcal{Y}^t \\
&+ 3\left\|\mathbf{u}^t - \mathbf{1}\bar{u}^t\right\|^2 + \frac{4L^2}{q}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|^2 + L^2\left(\frac{4}{q}+3\right)\left\|\mathbf{w}^t - \mathbf{1}\bar{w}^t\right\|^2.
\end{aligned}
$$

Then we use the fact that $\left\|\mathbf{v}^{t+1} - \mathbf{v}^t\right\|^2 \le 2\left\|\mathbf{v}^{t+1}\right\|^2 + 2\left\|\mathbf{v}^t\right\|^2$, we can obtain the result. □

Substituting the result of Lemma 11 and Lemma 14 into Lemma 9, we obtain Lemma 2. Here, we rewrite the result of Lemma 2 by taking $q = b/m$, which contains the detailed expressions of $\mathbf{A}$, $\mathbf{B}$ and e^t

Lemma 15 (The complete version of Lemma 2)**.** *Let*

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

Under the settings of Lemma 1, we run Algorithm 2 by taking

$$
K = \left\lceil \frac{\log(1/\rho)}{\sqrt{1-\lambda_2(W)}} \right\rceil
$$

with

$$
\rho \le \frac{q\theta_1^3}{9}\min\left\{\frac{1}{2}\eta^2\sigma^2,\ \frac{1}{2}\left(\theta_1+\theta_2-\frac{\theta_2}{\gamma}\right),\ \frac{1}{2}p(1-\gamma)\right\},
$$

then we have

$$
\mathbb{E}\left[r^{t+1}\right] \le \rho^2\cdot\left(\mathbf{B}\cdot\mathbf{A}\cdot r^t + e^t\right),
$$

where we define matrix $\mathbf{A}$*, elementary matrix* $\mathbf{B}$ *and vector* e^t *as*

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 8\eta^2(1+\rho^2(1-p)) & 8\eta^2\rho^2 p \\ a_{31} & 12(1+2\rho^2) & 2 & a_{34} & a_{35} & a_{36} \\ \frac{3\eta^2\sigma^2}{(1+\eta\sigma)^2} & 0 & \frac{3}{(1+\eta\sigma)^2} & \frac{3}{(1+\eta\sigma)^2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1-p & p \\ \frac{9\rho^2\theta_1^2\eta^2\sigma^2}{(1+\eta\sigma)^2} & 0 & \frac{9\rho^2\theta_1^2}{(1+\eta\sigma)^2} & 3\theta_1^2 + \frac{9\rho^2\theta_1^2}{(1+\eta\sigma)^2} & 0 & 0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 3\rho^2\theta_1^2 & 3\rho^2\theta_2^2 & 3\rho^2(1-\theta_1-\theta_2)^2 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$e^t = \begin{bmatrix} 0 \\ \frac{16\eta^2 p\gamma\theta_1}{\theta_2}(\mathcal{W}^{t+1} + \mathcal{W}^t) \\ \frac{32\eta^3\theta_1 m}{(1+\eta\sigma)b}(\mathcal{Z}^t + \mathcal{Z}^{t+1}) + \frac{8\eta^2 p\gamma\theta_1}{\theta_2}(\frac{4m(1+\theta_2)}{b} + 3 + 48\rho^2)(\mathcal{W}^t + \mathcal{W}^{t+1}) + \frac{32\eta^3\theta_1 m}{b}(1-\theta_1-\theta_2)(\mathcal{Y}^t + \mathcal{Y}^{t+1}) \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

where

$$\begin{aligned} a_{31} &= \frac{16\eta^2 m}{b}\left(1 + 9\rho^4\left[\frac{\eta^2\sigma^2\theta_1^2}{(1+\eta\sigma)^2} + (1-\theta_1-\theta_2)^2\right]\right), \\ a_{34} &= \frac{144\eta^2\rho^4 m}{b}\left(\frac{\theta_1^2}{(1+\eta\sigma)^2} + \theta_1^2(1-\theta_1-\theta_2)^2\right), \\ a_{35} &= \frac{16\eta^2 m}{b}\left[1 + \rho^2(1-p) + 9\rho^4\theta_2^2(1-p)\right] + 12\eta^2(1+8\rho^2)(1+\rho^2(1-p)), \\ a_{36} &= 4\eta^2\rho^2 p\left[\frac{4m}{b}(1+3\rho^2\theta_2^2) + 3(1+8\rho^2)\right]. \end{aligned}$$

to simplify our equation. Then we can do a simple estimation with substituting $\eta = 1/(13\theta_1)$ *to obtain that*

$$\|\mathbf{A}\| < \frac{4m}{b\theta_1^2}, \quad \|\mathbf{B}\| < 2 \quad \textit{and} \quad \left\|e^t\right\| < \frac{2m}{3b\theta_1^2}(V^{t+1} + V^t).$$

C Proof of Theorem 1

Combining Lemma 1 and 2, we have got what we need to prove our main theorem.

Proof for Theorem1. We prove the theorem by induction. By Lemma 1, the theorem holds for $t = 1$. Now, we assume that

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

holds when $t \le k$ and are going to prove it holds for $t = k+1$. We use the notation $\mathbf{A}$, $\mathbf{B}$ and e^t by following Lemma 2 (Section B). By the definition of V^t and Lemma 15, we can obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|r^k\right\|\right] \\
\le&\frac{2\rho^2 m}{3b\theta_1^2}\cdot\sum_{i=1}^{k}\left(2\rho^2\left\|\mathbf{A}\right\|\right)^{k-i}\left(V^i+V^{i-1}\right)+\left(2\rho^2\left\|\mathbf{A}\right\|\right)^k\left\|r^0\right\| \\
\le&\frac{2\rho^2 m}{3b\theta_1^2}\cdot\sum_{i=1}^{k}\left(2\rho^2\left\|\mathbf{A}\right\|\right)^{k-i}\alpha^{i-1}(\alpha+1)(V^0+\left\|r^0\right\|)+\left(2\rho^2\left\|\mathbf{A}\right\|\right)^k\left\|r^0\right\| \\
\le&\frac{2\rho^2 m(\alpha+1)}{3b\theta_1^2\alpha}\cdot\left(\frac{\alpha}{2}\right)^k\sum_{i=1}^{k}\left(\frac{\alpha}{2}\right)^{-i}\alpha^i(V^0+\left\|r^0\right\|)+\left(2\rho^2\left\|\mathbf{A}\right\|\right)^k\left\|r^0\right\| \\
=&\frac{2\rho^2 m(\alpha+1)}{3b\theta_1^2\alpha}\cdot\left(\frac{\alpha}{2}\right)^k\cdot(2^{k+1}-2)(V^0+\left\|r^0\right\|)+\left(2\rho^2\left\|\mathbf{A}\right\|\right)^k\left\|r^0\right\| \\
\le&\left(\frac{8\rho^2 m}{3b\theta_1^2}\alpha^{k-1}+\left(2\rho^2\left\|\mathbf{A}\right\|\right)^k\right)\cdot(V^0+\left\|r^0\right\|),
\end{aligned}
\tag{C.2}
$$

where the second inequality is because the condition for ρ in Theorem 1 implies $\rho^2 \le \alpha/(4\left\|\mathbf{A}\right\|)$ when $\eta = 1/(13\theta_1)$ and we assume that for all $i \le k$, $\mathbb{E}\left[V^i\right] \le \alpha^i(V^0 + \left\|r^0\right\|)$. Furthermore, we can obtain

$$
\begin{aligned}
&\mathbb{E}\left[\frac{L}{3mb\theta_1}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\frac{L}{4mb\theta_1}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2\right] \\
\le&\frac{1}{3b\theta_1}\mathbb{E}\left[\left\|r^k\right\|\right] \\
\le&\frac{1}{3b\theta_1}\left(\frac{8\rho^2 m}{3b\theta_1^2}\alpha^{k-1}+\left(2\rho^2\left\|\mathbf{A}\right\|\right)^k\right)\cdot(V^0+\left\|r^0\right\|)
\end{aligned}
\tag{C.3}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\sqrt{\frac{2\eta L V^t}{(1+\eta\sigma)m}}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|\right] \\
\le&\sqrt{2\eta V^t}\sqrt{\mathbb{E}\|\mathbf{r}^k\|} \\
\le&\sqrt{2\eta V^t}\sqrt{\left(\frac{8\rho^2 m}{3b\theta_1^2}\alpha^{k-1}+(2\rho^2\|\mathbf{A}\|)^k\right)\cdot(V^0+\|\mathbf{r}^0\|)}.
\end{aligned}
\tag{C.4}
$$

Furthermore, if we denote that

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

we have

$$
\begin{aligned}
&\mathbb{E}\left[V^{k+1}\right] \\
\le&\beta V^k+\sqrt{\frac{2\eta L V^t}{(1+\eta\sigma)m}}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|+\frac{L}{3mb\theta_1}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2+\frac{L}{4mb\theta_1}\left\|\mathbf{w}^t-\mathbf{1}\bar{w}^t\right\|^2 \\
\le&\alpha^k(V^0+\|r^0\|)\left(\beta+\frac{1}{3b\theta_1}\left(\frac{8\rho^2 m}{3b\theta_1^2}\alpha^{-1}+\left(2\rho^2\left\|\mathbf{A}\right\|\alpha^{-1}\right)^k\right)\right. \\
&\left.+\sqrt{\frac{2\eta\alpha^{-k}}{3b\theta_1}\left(\frac{8\rho^2 m}{3b\theta_1^2}\alpha^{k-1}+(2\rho^2\left\|\mathbf{A}\right\|)^k\right)}\right) \\
\le&\alpha^k(V^0+\|r^0\|)\left(\beta+\frac{1}{3b\theta_1}\left(\frac{16\rho^2 m}{3b\theta_1^2}+\left(\frac{16\rho^2 m}{b\theta_1^2}\right)^k\right)\right.
\end{aligned}
\tag{C.5}
$$

$$
\begin{aligned}
&+\sqrt{\frac{2\eta}{3b\theta_1}}\left(\sqrt{\frac{8\rho^2 m}{3b\theta_1^2}}\alpha^{-1/2}+\left(\frac{16\rho^2 m}{b\theta_1^2}\right)^{\frac{k}{2}}\right)\Bigg) \qquad\qquad \text{(C.6)}\\
\le&\alpha^k(V^0+\|r^0\|)\left(\beta+\frac{1}{3b\theta_1}\left(\frac{16m}{3b\theta_1^2}+\frac{16m}{b\theta_1^2}\right)\rho^2+\sqrt{\frac{2\eta}{3b\theta_1}}\left(2\sqrt{\frac{8m}{3b\theta_1^2}}+\sqrt{\frac{16m}{b\theta_1^2}}\right)\rho\right)\\
=&\alpha^k(V^0+\|r^0\|)\left(\beta+\frac{64m}{9b^2\theta_1^3}\rho^2+4\left(1+\sqrt{\frac{2}{3}}\right)\sqrt{\frac{2}{39}}\frac{\sqrt{m}}{b\theta_1^2}\rho\right)\\
\le&\alpha^k\left(V^0+\left\|r^0\right\|\right)\left(\max\left(\frac{1}{1+\eta\sigma},1-\left(\theta_1+\theta_2-\frac{\theta_2}{\gamma}\right),1-p(1-\gamma)\right)+\frac{9m}{b\theta_1^3}\rho\right)\\
\le&\alpha^{k+1}\left(V^0+\left\|r^0\right\|\right),
\end{aligned}
$$

where the last inequality is because of the condition of ρ in Theorem 1. The first inequality is because of Lemma 1. Inequality (C.5) is because of the equation (C.1),(C.3) and (C.4) and inequality (C.6) is because of $\alpha \ge 1/2$ and $\|\mathbf{A}\| < (4m)/(b\theta_1^2)$. Thus, the theorem also holds for $t = k+1$ and we complete the proof by induction. Furthermore, Equation C.2 and condition of ρ in Theorem 1 imply that $\rho^2 \le \alpha/(4\|\mathbf{A}\|)$. Then we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\frac{L}{m}\left\|\mathbf{x}^t-\mathbf{1}\bar{x}^t\right\|^2\right]\\
\le&\left(\frac{8}{243}+2^{-t}\right)\max\left(1-\eta\sigma,1-\frac{1}{2}\left(\theta_1+\theta_2-\frac{\theta_2}{\gamma}\right),1-\frac{1}{2}p(1-\gamma)\right)^t\cdot\left(V^0+\left\|\mathbf{r}^0\right\|\right).
\end{aligned}
$$

□

D Proof of Corollary 2

Proof. We first prove that KNOT (Algorithm 2) can find an ϵ-suboptimal solution in expectation. We run KNOT with the setting of Theorem 1 and let

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

Then Theorem 1 means

$$
\frac{L}{2}\mathbb{E}\left[\left\|\bar{z}^T-x^*\right\|^2\right]\le\frac{\epsilon}{3},\qquad \mathbb{E}[f(\bar{y}^T)-f(x^*)]\le\frac{\epsilon}{3},\quad \mathbb{E}[f(\bar{w}^T)-f(x^*)]\le\frac{\epsilon}{3}
$$

and $\mathbb{E}\left[\left\|x_i^T-\bar{x}^T\right\|^2\right]\le m\epsilon/L$ for each $i=1,\dots,m$. From Assumption 1 and 3 we obtain that

$$
\begin{aligned}
&\mathbb{E}\left[f(\bar{x}^T)-f(x^*)\right]\\
\le&\theta_1\mathbb{E}\left[f(\bar{z}^T)-f(x^*)\right]+\theta_2\mathbb{E}\left[f(\bar{w}^T)-f(x^*)\right]+(1-\theta_1-\theta_2)\mathbb{E}\left[f(\bar{y}^T)-f(x^*)\right]\\
\le&\frac{\theta_1 L}{2}\mathbb{E}\left[\left\|\bar{z}^T-x^*\right\|^2\right]+\theta_2\mathbb{E}\left[f(\bar{w}^T)-f(x^*)\right]+(1-\theta_1-\theta_2)\mathbb{E}\left[f(\bar{y}^T)-f(x^*)\right]\\
\le&\left(\theta_1+\theta_2+(1-\theta_1-\theta_2)\right)\epsilon=\epsilon.
\end{aligned}
\tag{D.1}
$$

Moreover, Proposition 1 means step $\mathbf{x}_{\text{out}}=\texttt{AccGossip}(\mathbf{x}_T,K_{\text{out}})$ with $K_{\text{out}}=\mathcal{O}\big(\sqrt{1/\alpha}\log m\big)$ (line 23 of Algorithm 2) leads to $\bar{x}^{\text{out}}=\bar{x}^T$ and

$$
L\mathbb{E}\left[\left\|x_i^{\text{out}}-\bar{x}^{\text{out}}\right\|^2\right]\le\frac{\epsilon}{3}
\tag{D.2}
$$

for each $i = 1, \ldots, m$. Applying the smoothness of f (Assumption 1) and Young's inequality, we have

$$\begin{aligned}
f(x_i^{\text{out}}) - f(\bar{x}^{\text{out}}) \leq & \langle \nabla f(\bar{x}^{\text{out}}), x_i^{\text{out}} - \bar{x}^{\text{out}} \rangle + \frac{L}{2} \left\| x_i^{\text{out}} - \bar{x}^{\text{out}} \right\|^2 \\
\leq & \frac{c}{2} \left\| \nabla f(\bar{x}^{\text{out}}) \right\|^2 + \frac{1}{2c} \left\| x_i^{\text{out}} - \bar{x}^{\text{out}} \right\|^2 + \frac{L}{2} \left\| x_i^{\text{out}} - \bar{x}^{\text{out}} \right\|^2 .
\end{aligned}$$

for any $i = 1, \ldots, m$ and any $c > 0$. Additionally, Lemma 3 implies

$$f(\bar{x}^{\text{out}}) - f(x^*) \geq \frac{1}{2L} \left\| \nabla f(\bar{x}^{\text{out}}) \right\|^2 .$$

Combing above two results with $c = 1/(2L)$, we have

$$\begin{aligned}
\mathbb{E}\left[f(x_i^{\text{out}}) - f(x^*) \right] \leq & cL\mathbb{E}\left[f(\bar{x}^{\text{out}}) - f(x^*) \right] + \left(\frac{1}{2c} + \frac{L}{2} \right) \mathbb{E}\left[\left\| x_i^{\text{out}} - \bar{x}^{\text{out}} \right\|^2 \right] \\
\leq & \frac{\epsilon}{2} + \frac{3}{2} \cdot \frac{\epsilon}{3} = \epsilon
\end{aligned}$$

for any $i = 1, \ldots, m$. This implies the output x^{out} is an ϵ-suboptimal solution in expectation.

Then we analyze the complexity of KNOT by following the parameter settings of Theorem 1.

Case 1: In the case of $m < \kappa$, we choose $b = \sqrt{m}$ and $p = 1/\sqrt{m}$. Thus,

$$\theta_1 = \min \left\{ \sqrt{\frac{b}{\kappa p}} \theta_2, \theta_2 \right\} = \sqrt{\frac{b}{\kappa p}} \theta_2 = \frac{1}{2\sqrt{\kappa}}.$$

By choosing $\gamma = 1 - \frac{1}{3}\sqrt{\frac{m}{\kappa}} \in (2/3, 1)$, we have

$$\frac{2}{\theta_1 + \theta_2 - \frac{\theta_2}{\gamma}} = \frac{2}{\theta_1 (1 - \frac{1}{3\gamma})} \leq \frac{4}{\theta_1} = 2\sqrt{\kappa},$$

which means

$$\frac{1}{\eta \sigma} + \frac{2}{\theta_1 + \theta_2 - \frac{\theta_2}{\gamma}} + \frac{2}{p(1-\gamma)} \leq \left(\frac{13}{2} + 4 + \frac{4}{1-\gamma} \right) \frac{1}{\theta_1} = \mathcal{O}(\sqrt{\kappa}\,).$$

Case 2: In the case of $m \geq \kappa$, we choose $b = \sqrt{\kappa}$ and $p = 1/\sqrt{\kappa}$. Thus

$$\theta_1 = \min \left\{ \sqrt{\frac{b}{\kappa p}} \theta_2, \theta_2 \right\} = \theta_2 = \frac{1}{2\sqrt{\kappa}}.$$

By choosing $\gamma \in (2/3, 1)$, we have

$$\frac{1}{\eta \sigma} + \frac{2}{\theta_1 + \theta_2 - \frac{\theta_2}{\gamma}} + \frac{2}{p(1-\gamma)} \leq \left(\frac{13}{2} + 4 + \frac{4}{1-\gamma} \right) \frac{1}{\theta_1} = \mathcal{O}(\sqrt{\kappa}\,).$$

Therefore, the number of iterations for KNOT to achieve an expected ϵ-suboptimal solution is

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

Thus, the expected first-order oracle complexity is

$$T \cdot (b + mp) = \mathcal{O}\left((m + \sqrt{m\kappa}\,) \log \frac{1}{\epsilon} \right),$$

and the expected communication complexity is

$$T \cdot K = \mathcal{O}\left(\frac{\sqrt{\kappa}\log(m\kappa)}{\sqrt{1-\lambda_2(W)}}\log\frac{1}{\epsilon}\right).$$

$\square$