Decentralized Accelerated Proximal Gradient Descent

Haishan Ye[1] *   **Ziang Zhou**[1] *   **Luo Luo**[2]   **Tong Zhang**[2]
[1]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen
[2]Department of Mathematics, The Hong Kong University of Science and Technology
hsye_cs@outlook.com   zhouza15@fudan.edu.cn   luoluo@ust.hk   tongzhang@ust.hk

Abstract

Decentralized optimization has wide applications in machine learning, signal processing, and control. In this paper, we study the decentralized composite optimization problem with a non-smooth regularization term. Many proximal gradient based decentralized algorithms have been proposed in the past. However, these algorithms do not achieve near optimal computational complexity and communication complexity. In this paper, we propose a new method which establishes the optimal computational complexity and a near optimal communication complexity. Our empirical study shows that the proposed algorithm outperforms existing state-of-the-art algorithms.

1   Introduction

In this paper, we consider the decentralized composite optimization problem where agents aim to solve the following composite convex problem defined as

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

where each function $f_i(x)$ is the loss function of agent i and known only to the agent; $r(x)$ is a non-smooth, convex function shared by all agents. The agents form a connected and undirected network. Agents can communicate with their neighbors to cooperatively solve the Problem (1).

Many machine learning problems can be formulated as Problem (1) such as the elastic net [29], the graphical Lasso [4], and sparse logistic regression [24]. Because centralized optimization suffers from the communication traffic jam on the central server and robustness of the network, the decentralized optimization has become an active research topic in machine learning. Decentralized methods for solving Problem (1) have also been widely studied in signal processing, control, and optimization communities.

Due to the wide applications, many different decentralized algorithms have been proposed in the past. We review the existing methods for the smooth non-proximal case where $r(x) = 0$. Some earlier primal algorithms were penalty based, and they can only reach sublinear convergence $\mathcal{O}(1/t)$ even for strongly convex objective functions [12, 27, 7]. The methods in [20] established the linear convergence for decentralized methods based on ADMM. More recently, gradient tracking based algorithms were proposed and they all achieved linear convergence including EXTRA [18], ESOM [11] , exact diffusion [28], NIDS [9], and many others [14, 15, 3, 17, 8]. Qu & Li [15] combined acceleration technique with gradient tracking to obtain faster convergence rate. Some other recent works [16, 22, 5, 5] studied the dual formulations of the problem, and proposed accelerated dual gradient descent to reach an optimal convergence rate for smooth problems. Recently, Ye et al. [26] proposed Mudag, which can achieve the optimal computation complexity and near optimal communication complexity for the smooth case with $r(x) = 0$.

---

*Equal Contribution

Many gradient tracking based algorithms have been extended to decentralized composite optimization problems with a non-smooth regularization term such as `PG-EXTRA` [19] and `NIDS` [9]. However, due to the non-smooth term, these algorithms can only achieve sub-linear convergence rates. Recently, the authors of [21] proposed a gradient tracking based method called `SONATA`, and established a linear convergence rate with the assumption that $f(x)$ is strongly convex. In addition, Alghunaim et al. [2] proposed a primal-dual algorithm which can achieve linear convergence rate when each $f_i(x)$ is convex. Recently, a unified framework to analyze a large group of algorithms, and showed that these algorithms can also achieve linear convergence rates with nonsmooth regularization term such as `EXTRA (PG-EXTRA)` [18], `NIDS` [9], and `Harnessing` [14] in the work of [1, 25]. Despite intensive studies in the literature, the convergence rates of these previous algorithms do not match the optimal convergence rate. Moreover, the communication complexities achieved by algorithms analyzed in the framework of Xu et al. [25] and Alghunaim et al. [1] are sub-optimal.

In this paper, we propose a novel algorithm that can establish the optimal computation complexity and a near optimal communication complexity. Our method is closely related to `Mudag` of [26], which can only handle the non-composite case with $r(x) = 0$. We summarize our contributions as follows:

1. We proposed a novel decentralized primal proximal algorithm that can achieve the optimal computational complexity $\mathcal{O}(\sqrt{\kappa_g}\log(\frac{1}{\epsilon}))$ and an almost optimal communication complexity $\mathcal{O}\left(\sqrt{\frac{\kappa_g}{1-\lambda_2(W)}}\log(\frac{p(M,L,\mu)}{q(M,L,\mu)})\log\frac{1}{\epsilon}\right)$, which matches the lower bound up to a $\log$ factor. κ_g is the global condition number of $f(x)$. M and L are the smoothness parameter of local $f_i(x)$ and $f(x)$, respectively. p and q are polynomials of M, L and μ of order no larger than three. To the best of our knowledge, our work is the first near optimal decentralized proximal algorithm.

2. Our algorithm does not require each individual function to be (strongly) convex. Thus, our algorithm has a wide application range since $f_i(x)$ may not be convex in many machine learning problems. The strong convexity condition of each $f_i(x)$ is required in the algorithms suited in the framework of [25] to achieve linear convergence rates. `SONATA` can also achieve linear convergence rate without this condition [21]. However, `SONATA` has to construct successive convex approximation function to deal with the non-convexity of $f_i(x)$ which requires extra computation burden.

2 Problem Set-up

We first introduce some properties of objective functions that will be used in this paper. Then we introduce the Gossip matrix associated with the network. Finally, we reformulate the Problem (1).

2.1 Function Properties

Global L-Smoothness   We call $f(x)$ is L-smooth, that is, for any $y, x \in \mathbb{R}^d$, it holds that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

Global μ-Strong Convexity   We call $f(x)$ is μ-strongly convex, that is, for any $y, x \in \mathbb{R}^d$, it holds that

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

Local M-Smoothness   For each $f_i(x)$ in Eqn. (1), and any $y, x \in \mathbb{R}^d$, it holds that

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

Local ν-Strong Convexity   For each $f_i(x)$ in Eqn. (1), and any $y, x \in \mathbb{R}^d$, it holds that

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

Based on the smoothness and strong convexity, we can define global and local condition number of the objective function respectively as follows

$$\kappa_g = L/\mu \quad \text{and} \quad \kappa_\ell = M/\nu.$$

It is well known that $\kappa_g \leq \kappa_\ell$ and $L \leq M$. Furthermore, if $f(x)$ is μ-strongly convex and r is convex, then one can easily check that $h(x) = f(x) + r(x)$ is also μ-strongly convex.

2.2 Gossip Matrix

We use $W \in \mathbb{R}^{m \times m}$ to denote the gossip matrix associated with the network and use $\lambda_2(W)$ to denote the second largest eigenvalue of W. The gossip matrix W has the following important properties:

1. W is symmetric with $W_{i,j} \neq 0$ if and only if agents i and j are connected or $i = j$.
2. $\mathbf{0} \preceq W \preceq I, W\mathbf{1} = \mathbf{1}$, null$(I - W) = \text{span}(\mathbf{1})$.

We use I to denote the $m \times m$ identity matrix and $\mathbf{1} = [1, \ldots, 1]^\top \in \mathbb{R}^m$ denotes the vector with all ones.

By above two properties, one can achieve averaging x_i's in different agents by several steps of local communication (by multiplying W). Instead of directly multiplying W several times, Liu & Morse [10] proposed a more efficient way to achieve averaging described in Algorithm 2 which has the following important proposition.

Proposition 1. *Let $\mathbf{x}^K$ be the output of Algorithm 2 and $\bar{x} = \frac{1}{m}\mathbf{1}^\top \mathbf{x}^0$. Then it holds that*

$$\bar{x} = \frac{1}{m}\mathbf{1}^\top \mathbf{x}^K, \quad \text{and} \quad \left\| \mathbf{x}^K - \mathbf{1}\bar{x} \right\| \leq \left(1 - \sqrt{1 - \lambda_2(W)} \right)^K \left\| \mathbf{x}^0 - \mathbf{1}\bar{x} \right\|,$$

where $\lambda_2(W)$ is the second largest eigenvalue of W.

2.3 Problem Reformulation

Denote by $x_i \in \mathbb{R}^d$ the local copy of the variable of x for agent i. We introduce the aggregated variable $\mathbf{x}$ and the aggregated objective function $H(\mathbf{x})$ as

$$H(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(x_i) + \frac{1}{m}\sum_{i=1}^{m} r(x_i) \text{ with } \mathbf{x} = [x_1, \cdots, x_m]^\top. \tag{2}$$

We can reformulate Problem (1) as

$$\min_{\mathbf{x} \in \mathbb{R}^{m \times d}} H(\mathbf{x}) \quad \text{subject to} \quad x_1 = x_2 = \cdots = x_m. \tag{3}$$

We denote the aggregated smooth component, the aggregated non-smooth component, and $\nabla F(\mathbf{x})$ as

$$F(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(x_i), \ R(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} r(x_i), \ \nabla F(\mathbf{x}) = \frac{1}{m}[\nabla f_1(x_1), \cdots, \nabla f_m(x_m)]^\top.$$

Moreover, we denote the local proximal operator and aggregate proximal operator as

$$\mathbf{prox}_{\eta,r}(x) = \operatorname*{argmin}_{z \in \mathbb{R}^d} \left(r(z) + \frac{1}{2\eta}\|z - x\|^2 \right), \quad \mathbf{prox}_{m\eta,R}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^{m \times d}} \left(R(\mathbf{z}) + \frac{1}{2m\eta}\|\mathbf{z} - \mathbf{x}\|^2 \right). \tag{4}$$

3 Decentralized Accelerated Proximal Gradient Descent

In this section, we propose a novel decentralized proximal gradient descent algorithm achieving the optimal computational complexity and near optimal communication complexity. We describe our decentralized accelerated proximal gradient descent method (DAPG) in Algorithm 1. The use of multi-consensus (FastMix) and gradient tracking (that is, each time we track the gradient by communicating $s_t + \nabla F(\mathbf{y}_{t+1}) - \nabla F(\mathbf{y}_t)$) is motivated by Mudag of [26]. However, the actual algorithm is quite different, due to the need to handle the proximal term.

Algorithm 1 DPAG

1: **Input:** $\mathbf{x}_0^{(i)} = \mathbf{x}_0^{(j)}$ for $1 \leq i, j, \leq m$, $\mathbf{y}_0 = \mathbf{x}_0$, $\mathbf{s}_0 = \nabla F(\mathbf{x}_0)$, $\eta = \frac{1}{L}$ and $\alpha = \sqrt{\frac{\mu}{L}}$, $K = \mathcal{O}\left(\frac{1}{\sqrt{1-\lambda_2(W)}} \log \frac{p(M,L,\mu)}{q(M,L,\mu)}\right)$, where p, q are polynomials with order less than 3.
2: **for** $t = 0, \ldots, T$ **do**
3:     $\mathbf{x}_{t+1} = \text{FastMix}(\mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t), K)$
4:     $\mathbf{y}_{t+1} = \text{FastMix}\left(\mathbf{x}_{t+1} + \frac{1-\alpha}{1+\alpha}(\mathbf{x}_{t+1} - \mathbf{x}_t), K\right)$
5:     $\mathbf{s}_{t+1} = \text{FastMix}(\mathbf{s}_t + \nabla F(\mathbf{y}_{t+1}) - \nabla F(\mathbf{y}_t), K)$
6: **end for**
7: **Output:** $\bar{x}_T$.

Algorithm 2 FastMix

1: **Input:** $\mathbf{x}^0 = \mathbf{x}^{-1}$, K, W, step size $\eta_w = \frac{1-\sqrt{1-\lambda_2^2(W)}}{1+\sqrt{1-\lambda_2^2(W)}}$.
2: **for** $k = 0, \ldots, K$ **do**
3:     $\mathbf{x}^{k+1} = (1 + \eta_w)W\mathbf{x}^k - \eta_w \mathbf{x}^{k-1}$;
4: **end for**
5: **Output:** $\mathbf{x}^K$.

3.1 Sketch of the Main Proof Techniques

The main idea behind DAPG is trying to approximate the centralized accelerated proximal gradient descent by Nesterov's acceleration, multi-consensus, and gradient tracking. We first define the average of the aggregated variables defined in Algorithm 1 as follows

$$\bar{x}_t = \frac{1}{m} \sum_{i=0}^{m} \mathbf{x}_t^{(i)}, \quad \bar{y}_t = \frac{1}{m} \sum_{i=0}^{m} \mathbf{y}_t^{(i)}, \quad \bar{s}_t = \frac{1}{m} \sum_{i=0}^{m} \mathbf{s}_t^{(i)}, \quad \bar{g}_t = \frac{1}{m} \sum_{i=0}^{m} \nabla f_i(\mathbf{y}_t^{(i)}), \quad (5)$$

where $\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$ indicates the i-th row of matrix $\mathbf{x}$ and $\mathbf{y}$. We can regard these averaged variables as the approximation of their centralized counterparts.

In our method, instead of solving Problem (1) over the network, we minimize the reformulated function (3). Our algorithm tries to use accelerated proximal gradient descent to minimize $H(\mathbf{x})$ where the acceleration helps to achieve a fast convergence rate. To deal with the consensus constraints $x_i = x_j$ for $1 \leq i, j \leq m$, we resort to the multi-consensus and gradient-tracking techniques. We then show that variables such as $\mathbf{x}_t^i, \mathbf{y}_t^i$ and $\mathbf{s}_t^i$ in agent i will converge to their centralized counterparts $\bar{x}_t, \bar{y}_t$ and $\bar{s}_t$ as t increases.

Once local variables and gradients can approximate their centralized counterparts, we can show that our algorithm has convergence properties similar to that of the centralized accelerated proximal gradient descent. This implies that our algorithm can also achieve the optimal computational complexity [13].

Previously multi-consensus was regarded as communication-unfriendly because, without gradient-tracking, one has to increase the communication times to achieve high precision consensuses [6, 14]. However, it was shown in [26] that the combination of multi-consensus and gradient-tracking leads to communication-efficiency. We follow this argument, and show that the proposed algorithm also achieves near optimal communication complexity in the proximal case.

3.2 Complexity Analysis

Following earlier work, we measure the computational complexity by the number of times that the gradient of $f(x)$ is computed, and we measure the communication complexity by the times of local communications, which is presented as Wx in our algorithm. Similar to the analysis of accelerated proximal gradient descent method, we define the Lyapunov function as follows

$$V_t \triangleq h(\bar{x}_t) - h(x^*) + \frac{\mu}{2} \|\bar{v}_t - x^*\|^2, \quad (6)$$

where $\bar{v}_t = \bar{x}_{t-1} + \frac{1}{\alpha}(\bar{x}_t - \bar{x}_{t-1})$, with $\alpha = \sqrt{\frac{\mu}{L}}$.

4

In the following lemma, we will show how the multi-consensus ('FastMix' operation in Algorithm 2) helps to achieve the consensus and bound the error between local variables and their centralized counterparts.

Lemma 1. *Suppose $f(x)$ is L-smooth and μ-strongly convex. Assume each $f_i(x)$ is M-smooth and r is a proper and lower-semicontinuous convex function. Let $\mathbf{z}_t = [\|\mathbf{x}_t - \mathbf{1}\bar{x}_t\|, \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\|, \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|]^\top$, where $\mathbf{1} = [1, \dots, 1]^\top \in \mathbb{R}^m$. It holds that*

$$\mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t + 8\rho M\sqrt{m}[0, 0, \sqrt{\frac{2}{\mu}V_t}]^\top, \tag{7}$$

where ρ and $\mathbf{A}$ are defined as

$$\rho = \left(1 - \sqrt{1 - \lambda_2(W)}\right)^K, \quad \mathbf{A} \triangleq \rho \cdot \begin{pmatrix} 0 & 2 & 2\eta \\ 2 & 4 & 4\eta \\ 2M & M(13 + 4M\eta) & 1 + 6M\eta \end{pmatrix}$$

Furthermore, we have

$$\mathbf{z}_{t+1} \leq \mathbf{A}^{t+1}\mathbf{z}_0 + 8M\rho\sqrt{\frac{2m}{\mu}} \cdot \sum_{i=0}^{t} \mathbf{A}^{t-i}[0, 0, \sqrt{V_i}]^\top.$$

If the spectral norm of $\mathbf{A}$ is less than 1 then V_t will converge to zero. It follows that $\|\mathbf{z}_t\|$, which can be regarded as a measure of total decentralized error compared to its centralized counterpart, will converge to zero. That is, the proposed method can well approximate accelerated proximal gradient descent. The following Lemmas show how the Lyapunov function V_t decreases at the disturbance caused by the decentralized setting.

Lemma 2. *The Lyapunov function V_t of Algorithm 1 defined in Eqn. (6) has the following property*

$$V_{t+1} \leq (1 - \alpha)V_t + D_1 \sqrt{V_t} \cdot \|\mathbf{z}_t\| + D_2 \|\mathbf{z}_t\|^2, \tag{8}$$

where $\mathbf{z}_t$ is defined in Lemma 1 and the constants D_1, D_2 are defined as follows

$$D_1 = \frac{4}{\sqrt{m}} \left(21L + 2\sqrt{L\mu}(2 + 2M\eta) + 3\mu(2 + 2M\eta) + 8 + \frac{\mu}{L} + 3\sqrt{\frac{\mu}{L}}\right)$$

$$D_2 = \frac{8}{m}(1 + \mu) \cdot \left(9 \cdot (12 + 8L + 4M)^2 + 16L(L + 1) + \frac{133 + 79L}{L}\right).$$

By Lemma 1, we can obtain that the convergence properties of $\|\mathbf{z}_t\|$ is determined by the value of ρ and V_t. At the same time, Lemma 2 shows that the value of $\|\mathbf{z}_t\|$ will affect the convergence properties of V_t in turn. In the following lemma, we give the condition to achieve $\|\mathbf{A}\|_2 \leq \frac{1}{2}$.

Lemma 3. *Let $\|\mathbf{A}\|_2$ be the spectral norm of matrix $\mathbf{A}$ defined in Lemma 1. With the constant D_3 defined as*

$$D_3 = 9 + 6\eta + 15M + 4M^2\eta + 6M\eta,$$

if $\rho < \frac{1}{2D_3}$, we will have $\|\mathbf{A}\|_2 < \frac{1}{2}$.

In the following lemma, we will show that once $\alpha \leq \frac{1}{2}$, the Lyapunov function V_t will converge with the rate $1 - \frac{\alpha}{2}$ once we choose a proper ρ. We can observe that we can get a slightly slower convergence rate than vanilla accelerated proximal gradient descent and this is caused by the error brought by the decentralized setting.

Lemma 4. *Assuming that $\alpha \leq \frac{1}{2}$, $f(x)$ is L-smooth and μ-strongly convex, each $f_i(x)$ is M-smooth, $r(x)$ is a proper and lower-semicontinuous convex function, then Algorithm 1 has the following convergence rate*

$$V_{t+1} \leq \left(1 - \frac{\alpha}{2}\right)^{t+1} (V_0 + C\|\mathbf{z}_0\|^2), \text{ with } C = \frac{D_1^2}{\alpha} + 2D_2, \tag{9}$$

if ρ satisfies the conditions in Lemma 3 and

$$\rho < \frac{\alpha}{2(D_1 D_4 + D_2 D_4^2)}, \text{ with } D_4 = 24M\sqrt{\frac{2m}{\mu}} + 2\frac{D_3}{\sqrt{C}}.$$

Moreover, we can bound $\|\mathbf{z}_t\|$ as

$$\|\mathbf{z}_t\|^2 \leq 2\rho \cdot D_4^2 \left(1 - \frac{\alpha}{2}\right)^t (V_0 + C\|\mathbf{z}_0\|^2). \tag{10}$$

With the above core lemmas, we give the detailed computation complexity and communication complexity of our algorithm in the following theorem.

Theorem 1 (Main Theorem). *Assume that $f(x)$ is L-smooth and μ-strongly convex, each $f_i(x)$ is M-smooth, $\alpha \leq \frac{1}{2}$, r is a proper and lower-semicontinuous convex function. Letting K satisfy that*

$$K = \sqrt{\frac{\kappa_g}{1 - \lambda_2(W)}} \log(\rho^{-1}), \text{ with } \rho < \min\left\{\frac{\alpha}{2(D_1 D_4 + D_2 D_4^2)}, \frac{1}{2D_3}\right\},$$

then, it holds that for the output $\bar{x}_T$ of Algorithm 1,

$$h(\bar{x}_T) - h(x^*) \leq \left(1 - \frac{\alpha}{2}\right)^T \left(h(\bar{x}_0) - h(x^*) + \frac{\mu \|\bar{x}_0 - x^*\|^2}{2} + C \sum_{i=1}^m \|\nabla f_i(\bar{x}_0) - \nabla f(\bar{x}_0)\|^2\right).$$

To achieve $h(\bar{x}_T) - h(x^*) < \epsilon$ and $\|\mathbf{x}_T - \mathbf{1}x^*\|^2 = \mathcal{O}(\epsilon/\mu)$, the computational and communication complexities of Algorithm 1 are

$$T = \mathcal{O}\left(\sqrt{\kappa_g} \log \frac{1}{\epsilon}\right), \text{ and } Q = \mathcal{O}\left(\sqrt{\frac{\kappa_g}{1 - \lambda_2(W)}} \log \frac{p(M, L, \mu)}{q(M, L, \mu)} \log(\frac{1}{\epsilon})\right),$$

where each $\mathcal{O}(\cdot)$ contains a universal constant and $p(M, L, \mu), q(M, L, \mu)$ are polynomials with order less than 3.

Proof. Because ρ satisfies the conditions in Lemma 4, combining with the definition of V_t in Eqn. (6), we obtain that

$$h(\bar{x}_T) - h(x^*) \leq V_T \leq \left(1 - \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^T \left(h(\bar{x}_0) - h(x^*) + \frac{\mu}{2}\|\bar{x}_0 - x^*\|^2 + C\|\mathbf{z}_0\|^2\right)$$

$$\leq \exp\left(-\frac{T}{2}\sqrt{\frac{\mu}{L}}\right)\left(h(\bar{x}_0) - h(x^*) + \frac{\mu}{2}\|\bar{x}_0 - x^*\|^2 + C\|\mathbf{z}_0\|^2\right),$$

where C is defined in Lemma 4. Because when $t = 0$ we set all nodes as the same status , i.e., $\mathbf{x}_0 = \mathbf{1}\bar{x}_0$ and $\mathbf{y}_0 = \mathbf{1}\bar{y}_0$, we have $\|\mathbf{z}_0\|^2 = \sum_{i=1}^m \|\nabla f_i(\bar{x}_0) - \nabla f_i(x^*)\|^2$. Thus, to achieve $h(\bar{x}_T) - h(x^*) < \epsilon$, T requires to be

$$T = 2\sqrt{\kappa_g} \log \frac{h(\bar{x}_0) - h(x^*) + \frac{\mu}{2}\|\bar{x}_0 - x^*\|^2 + C\sum_{i=1}^m \|\nabla f_i(\bar{x}_0) - \nabla f_i(x^*)\|^2}{\epsilon} = \mathcal{O}(\sqrt{\kappa_g} \log \frac{1}{\epsilon}).$$

By Eqn. (10) of Lemma 4, we have

$$\|\mathbf{z}_T\|^2 \leq 2\rho^2 D_4^2 \left(1 - \frac{\alpha}{2}\right)^T (V_0 + C\|\mathbf{z}_0\|)^2 = \mathcal{O}(\epsilon/\mu).$$

Therefore, we can obtain

$$\|\mathbf{x}_T - \mathbf{1}x^*\|^2 \leq 2(\|\mathbf{x}_T - \mathbf{1}\bar{x}_T\|^2 + \|\mathbf{1}\bar{x}_T - \mathbf{1}x^*\|^2) \leq \|\mathbf{z}_T\|^2 + \frac{4}{\mu}V_T = \mathcal{O}(\epsilon/\mu).$$

The bound of K can be obtained by Proposition 1. By the properties of D_1, D_2, D_3, C, D_4, it is easy to check that D_3 and $\frac{\alpha}{2(D_1 D_4 + D_2 D_4^2)}$ are independent of m. Thus, to achieve the conditions of ρ in Lemma 2, 3, and 4, ρ only needs to be less than $\frac{p}{q}$, where $p(M, L, \mu)$ and $q(M, L, \mu)$ are polynomials of M, L and μ with orders less than 3. Thus, by Proposition 1, we only require that $K \leq \frac{1}{\sqrt{1-\lambda_2(W)}} \log \frac{p}{q}$. Combining with the computation complexity, we can obtain the total communication complexity as

$$Q = \mathcal{O}\left(\sqrt{\frac{\kappa_g}{1 - \lambda_2(W)}} \log \frac{p(M, L, \mu)}{q(M, L, \mu)} \log \frac{1}{\epsilon}\right).$$

$\square$

Remark 1. *We can observe that `DAPG` establishes the optimal computational complexity [13] and a near optimal communication complexity which matches the lower bound up to a $\log$ term independent of ϵ [16]. This is the best computation and communication complexity of decentralized proximal algorithms can achieve. Before our work, the fastest decentralized proximal algorithm is `NIDS` which establishes computation and communication complexities both of $\mathcal{O}\left(\max\left\{\kappa_\ell, \frac{1}{1-\lambda_2(W)}\right\}\right)$ [9, 25]. We can observe that the complexities of `DAPG` are much less than that of `NIDS`.*
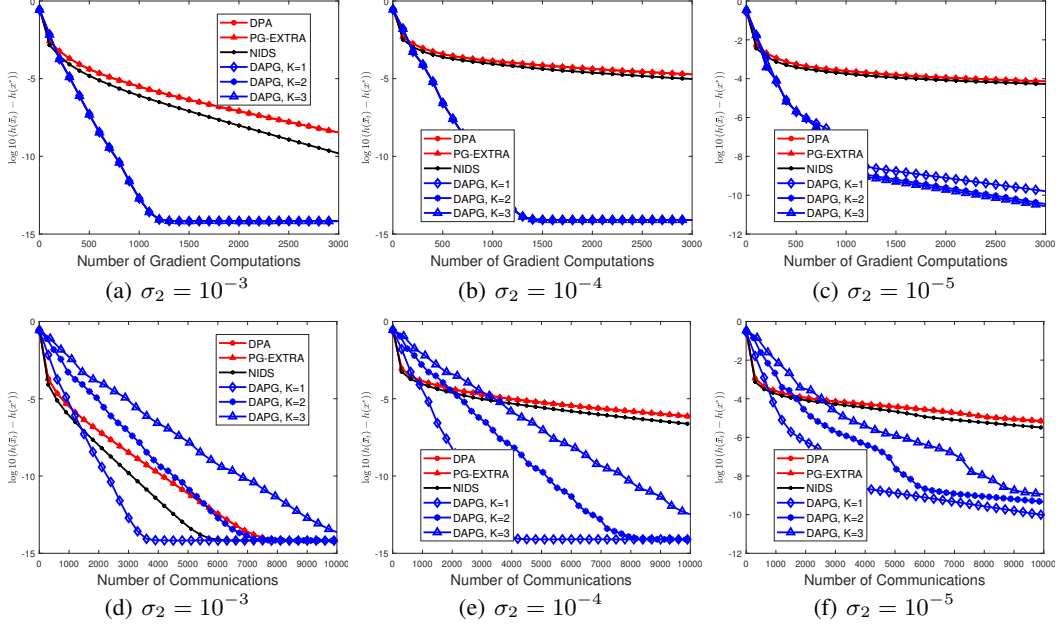
Figure 1: Experiment on 'a9a'. We compare the computation cost of algorithms in the top row and compare the communication cost of algorithms in the bottom row.

Remark 2. *The complexities of* `DAPG` *are linear to* $\sqrt{\kappa_g}$ *instead of* κ_ℓ. *In contrast, the complexities of algorithms fitting into the framework of [25] are linear to* κ_ℓ. *Note that* κ_ℓ *can be infinitely larger than* κ_g. *For example, for* $f(x) = \frac{1}{2}(f_1(x) + f_2(x)), x \in \mathbb{R}^2$, *with* $f_1(x) = x_1^2$ *and* $f_2(x) = x_1^2 + x_2^2$, *then it holds that* $\kappa_g = 2$ *and* $\kappa_\ell = \infty$. *Thus, the complexities of* `DAPG` *are better than those depending on the local condition number.*

Remark 3. `DAPG` *does not require each* $f_i(x)$ *be convex to achieve fast convergence rate. Thus,* `DAPG` *has a wide application range since* $f_i(x)$ *may be non-convex in some machine learning applications.* `SONATA` *can also be applied in these applications [21]. However,* `SONATA` *has to construct an SCA surrogate function of the non-convex function* $f_i(x)$ *to deal with the local non-convexity. Furthermore,* `DAPG` *only takes a cheap proximal mapping each iteration while* `SONATA` *has to minimize a sub-problem. Hence,* `DAPG` *is a simpler and easier to implement algorithm.*

4 Experiments

In the previous sections, we have given the theoretical analysis of our algorithm. In this section, we will validate the effectiveness and computational efficiency of our algorithm empirically. We will conduct experiments on the sparse logistic regression problem where $f(x)$ is general strongly convex. The sparse logistic regression is defined as

$$F(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) + \sigma_1\|x\|_1 + \frac{\sigma_2}{2}\|x\|^2, \text{ with } f_i(x) = \frac{1}{n}\sum_{j=1}^{n} \log[1 + \exp(-b_j\langle a_j, x\rangle)] \quad (11)$$

where $a_j \in \mathbb{R}^d$ is the j-th input vector, and $b_j \in \{-1, 1\}$ is the corresponding label.

Experiments Setting In our experiments, we consider random networks where each pair of agents have a connection with a probability of $p = 0.1$. We set $W = I - \frac{L}{\lambda_1(L)}$ where L is the Laplacian matrix associated with a weighted graph, and $\lambda_1(L)$ is the largest eigenvalue of L. We set $m = 100$, that is, there exists 100 agents in this network. In our experiments, the gossip matrix W satisfies $1 - \lambda_2(W) = 0.05$.

We conduct experiments on the datasets 'w8a' and 'w9a' which can be downloaded in libsvm datasets. For 'w8a', we set $n = 497$ and $d = 300$. For 'a9a', we set $n = 325$ and $d = 123$. We set $\sigma_1 = 10^{-4}$
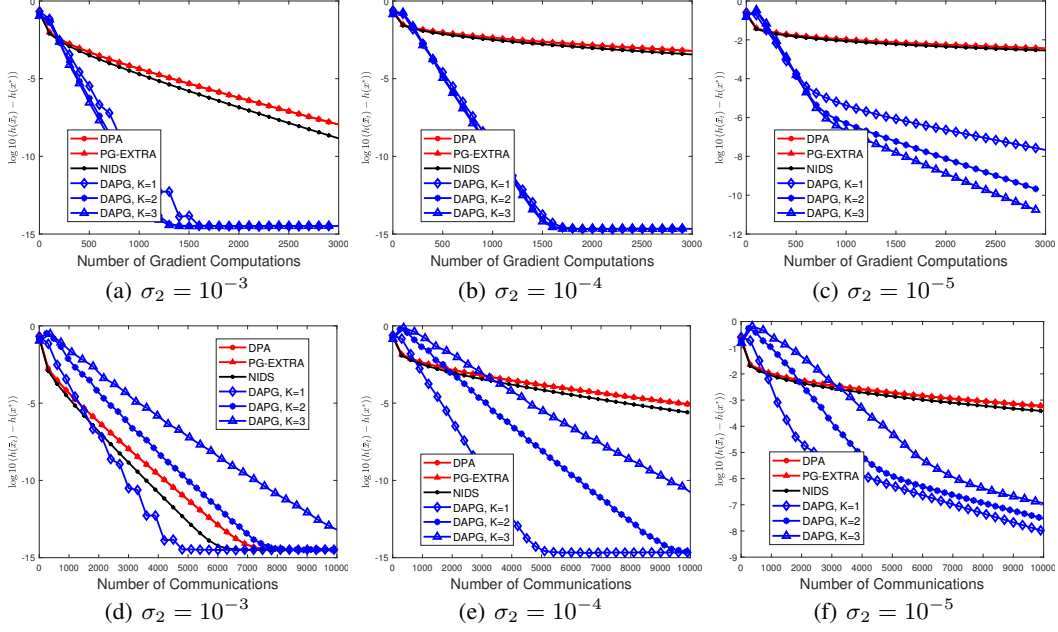
Figure 2: Experiment on 'w8a'. We compare the computation cost of algorithms in the top row and compare the communication cost of algorithms in the bottom row.

for all datasets and set σ_2 as 10^{-3}, 10^{-4} and 10^{-5} to control the condition number of the objective function.

Comparison with Existing Works We compare our work with state-of-the-art algorithms PG-EXTRA [19], NIDS [9] and Decentralized Proximal Algorithm (DPA) [2]. In the experiments, we set $K = 1, 2, 3$ respectively. The parameters of all algorithms are well-tuned. We report experiment results in Figure 1 and 2. We can observe that DAPG takes much less computational cost than other algorithms because DAPG uses Nesterov's acceleration to achieve a faster convergence rate. This matches our theoretical analysis of the computation complexity. We can further observe that the advantage of DAPG is more clear when σ_2 is small. This is because a small σ_2 commonly leads to a large condition number and the computation complexity of DAPG is linear to $\sqrt{\kappa_g}$ instead of κ_ℓ. DAPG also shows great advantages over other state-of-the-art decentralized proximal algorithms on the communication cost. Though DAPG takes three times of local communication while other algorithms communicate only once for each iteration, DAPG still requires much less communication costs because of its fast convergence rate when σ_2 is small.

5   Conclusion

In this paper, we studied the decentralized composite optimization problem with a non-smooth regularization term. We proposed a novel algorithm that achieves the optimal computation complexity and a near optimal communication complexity matching the lower bound up to a $\log$ term independent of ϵ. This is the best known communication complexity that decentralized proximal gradient algorithms can achieve. Furthermore, DAPG does not require each individual function $f_i(x)$ to be convex to achieve a fast convergence rate. Hence, our algorithm has a wider range of applications in machine learning than other decentralized proximal gradient descent algorithms which require $f_i(x)$ to be convex. Finally, the complexities of our algorithms depend on the global condition number κ_g instead of the local one. Since $\kappa_g \leq \kappa_\ell$ and κ_ℓ can be infinitely larger than κ_g, our algorithm can achieve much better performance than the algorithms whose complexities depend on κ_ℓ. The experiments also validate the advantages of our algorithm.

6 Broader Impact

Our work focuses on the theory of decentralized optimization and proposes a novel decentralized proximal algorithm. This will help us to design new decentralized proximal algorithms. Because decentralized optimization has wide applications in machine learning, sensor networks, and multi-robot system, our work may be used in these areas.

Acknowledgments and Disclosure of Funding

References

[1] Alghunaim, S. A., Ryu, E., Yuan, K., & Sayed, A. H. (2020). Decentralized proximal gradient algorithms with linear convergence rates. *IEEE Transactions on Automatic Control*.

[2] Alghunaim, S. A., Yuan, K., & Sayed, A. H. (2019). A linearly convergent proximal gradient algorithm for decentralized optimization. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada* (pp. 2844–2854).

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

[4] Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.

[5] Hendrikx, H., Bach, F., & Massoulié, L. (2019). An accelerated decentralized stochastic proximal algorithm for finite sums. In *Advances in Neural Information Processing Systems* (pp. 952–962).

[6] Jakovetić, D., Xavier, J., & Moura, J. M. (2014). Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5), 1131–1146.

[7] Lan, G. & Monteiro, R. D. (2013). Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138(1-2), 115–139.

[8] Li, B., Cen, S., Chen, Y., & Chi, Y. (2019a). Communication-efficient distributed optimization in networks with gradient tracking. *arXiv preprint arXiv:1909.05844*.

[9] Li, Z., Shi, W., & Yan, M. (2019b). A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17), 4494–4506.

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

[11] Mokhtari, A., Shi, W., Ling, Q., & Ribeiro, A. (2016). A decentralized second-order method with exact linear convergence rate for consensus optimization. *IEEE Trans. Signal Inf. Process. over Networks*, 2(4), 507–522.

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

[13] Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

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

[15] Qu, G. & Li, N. (2019). Accelerated distributed nesterov gradient descent. *IEEE Transactions on Automatic Control*.

[16] Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., & Massoulié, L. (2017). Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3027–3036).: JMLR. org.

[17] Shen, Z., Mokhtari, A., Zhou, T., Zhao, P., & Qian, H. (2018). Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication. In *International Conference on Machine Learning* (pp. 4624–4633).

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

[19] Shi, W., Ling, Q., Wu, G., & Yin, W. (2015b). A proximal gradient algorithm for decentralized composite optimization. *IEEE Trans. Signal Process.*, 63(22), 6013–6023.

[20] Shi, W., Ling, Q., Yuan, K., Wu, G., & Yin, W. (2014). On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7), 1750–1761.

[21] Sun, Y., Daneshmand, A., & Scutari, G. (2019). Convergence rate of distributed optimization algorithms based on gradient tracking. *arXiv preprint arXiv:1905.02637*.

[22] Uribe, C. A., Lee, S., Gasnikov, A., & Nedić, A. (2018). A dual approach for optimal algorithms in distributed optimization over networks. *arXiv preprint arXiv:1809.00710*.

[23] Xiao, L. & Boyd, S. (2004). Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1), 65–78.

[24] Xiao, L. & Zhang, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4), 2057–2075.

[25] Xu, J., Tian, Y., Sun, Y., & Scutari, G. (2020). Distributed algorithms for composite optimization: Unified and tight convergence analysis. *CoRR*, abs/2002.11534.

[26] Ye, H., Luo, L., Zhou, Z., & Zhang, T. (2020). Multi-consensus decentralized accelerated gradient descent. *CoRR*, abs/2005.00797.

[27] Yuan, K., Ling, Q., & Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3), 1835–1854.

[28] Yuan, K., Ying, B., Zhao, X., & Sayed, A. H. (2019). Exact diffusion for distributed optimization and learning - part I: algorithm development. *IEEE Trans. Signal Process.*, 67(3), 708–723.

[29] Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

A Proof of Proposition 1

Proof of Proposition 1. By the update rule of Algorithm 2 and the fact that $W^\infty = \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ [23], we have

$$W^\infty \mathbf{x}^K = W^\infty \mathbf{x}^{K-1} + \eta_w \left(W^\infty \mathbf{x}^{K-1} - W^\infty \mathbf{x}^{K-2} \right).$$

We can obtain that

$$W^\infty \left(\mathbf{x}^K - \mathbf{x}^{K-1} \right) = \eta_w \left(W^\infty \mathbf{x}^{K-1} - W^\infty \mathbf{x}^{K-2} \right).$$

Note that $\mathbf{x}^0 = \mathbf{x}^{-1}$ in Algorithm 2, we can obtain that for any $k = 0, \ldots, K$, we have

$$W^\infty \left(\mathbf{x}^k - \mathbf{x}^{k-1} \right) = 0.$$

Therefore, we can obtain the identity $W^\infty \mathbf{x}^K = W^\infty \mathbf{x}^0$, which implies the result. The convergence rate of Algorithm 2 can be found in [10]. $\qquad\square$

B Collection of Lemmas

We list several important lemmas that will be used in our proofs.

Lemma 5 ([13]). *Letting $\tilde{\nabla}h(x)$ the generalized gradient of $h(x)$ defined as*

$$\tilde{\nabla}h(x) \triangleq \frac{x - \boldsymbol{prox}_{\eta,r}(x - \eta \nabla f(x))}{\eta}, \ \text{ with } \eta = \frac{1}{L} \text{ being the step size,} \tag{12}$$

then it holds that $\tilde{\nabla}h(x^*) = 0$ if x^* minimizes $h(x)$.

Lemma 6. *Letting $\boldsymbol{prox}^{(i)}_{\eta m, R}(\mathbf{x})$ denote the i-th row of the matrix $\boldsymbol{prox}_{\eta m, R}(\mathbf{x})$ (defined in Eqn. (4)), we have the following equation*

$$\boldsymbol{prox}^{(i)}_{\eta m, R}(\mathbf{x}) = \boldsymbol{prox}_{\eta,r}(\mathbf{x}^{(i)}).$$

Proof. By the definition of the proximal operators, we have

$$
\begin{aligned}
\mathbf{prox}_{\eta m, R}(\mathbf{x}) &= \underset{\mathbf{z}}{\operatorname{argmin}} \left(R(\mathbf{z}) + \frac{1}{2\eta m} \|\mathbf{z} - \mathbf{x}\|_F^2 \right) \\
&= \underset{\mathbf{z}}{\operatorname{argmin}} \left(\frac{1}{m} \sum_{i=1}^{m} r(\mathbf{z}^{(i)}) + \sum_{i=1}^{m} \frac{1}{2\eta m} \left\| \mathbf{z}^{(i)} - \mathbf{x}^{(i)} \right\|^2 \right) \\
&= \underset{\mathbf{z}}{\operatorname{argmin}} \left(\sum_{i=1}^{m} r(\mathbf{z}^{(i)}) + \sum_{i=1}^{m} \frac{1}{2\eta} \left\| \mathbf{z}^{(i)} - \mathbf{x}^{(i)} \right\|^2 \right) \\
&= \begin{pmatrix} \operatorname{argmin}_z \left(r(z) + \frac{1}{2\eta} \left\| z - \mathbf{x}^{(1)} \right\| \right)^\top \\ \vdots \\ \operatorname{argmin}_z \left(r(z) + \frac{1}{2\eta} \left\| z - \mathbf{x}^{(m)} \right\| \right)^\top \end{pmatrix}.
\end{aligned}
$$

Therefore, we have the following equation

$$\mathbf{prox}^{(i)}_{\eta m, R}(\mathbf{x}) = \mathbf{prox}_{\eta,r}(\mathbf{x}^{(i)}).$$

$\qquad\square$

Lemma 7. *For $\bar{x}_t$, $\bar{y}_t$ and $\bar{v}_t$ defined in Eqn. (5) and (6), then we can obtain that*

$$\bar{y}_t - \bar{x}_t = \alpha(\bar{v}_t - \bar{y}_t)$$

$$\bar{y}_{t+1} = \frac{\bar{x}_{t+1} + \alpha\bar{v}_{t+1}}{1 + \alpha}.$$

Proof. First using the definition of $\bar{v}_t$ we have

$$\frac{\bar{x}_{t+1} + \alpha\bar{v}_{t+1}}{1 + \alpha} = \frac{\bar{x}_{t+1} + \alpha[\bar{x}_t + \frac{1}{\alpha}(\bar{x}_{t+1} - \bar{x}_t)]}{1 + \alpha}$$
$$= \bar{x}_{t+1} + \frac{1 - \alpha}{1 + \alpha}(\bar{x}_{t+1} - \bar{x}_t)$$
$$= \bar{y}_{t+1}$$

Then we can have

$$\bar{y}_t - \bar{x}_t = \alpha(\bar{v}_t - \bar{y}_t).$$

$\square$

Lemma 8. *Let $h(x)$ be μ-strongly convex. For $\bar{y}_t$, x^* and V_t defined in Eqn. (6) and (5), we have the following inequality,*

$$\|\bar{y}_t - x^*\| \le \sqrt{\frac{2}{\mu}V_t}.$$

Proof. Combined Lemma 7, the optimality of x^*, and the μ-strongly convexity of $h(x)$, we obtain that

$$\|\bar{y}_t - x^*\| = \left\|\frac{\bar{x}_t + \alpha\bar{v}_t}{1 + \alpha} - x^*\right\| \le \frac{1}{1 + \alpha}\|\bar{x}_t - x^*\| + \frac{\alpha}{1 + \alpha}\|\bar{v}_t - x^*\| \le \sqrt{\frac{2}{\mu}V_t}.$$

$\square$

Lemma 9. *If r is a proper and semi-continuous function, for any z, w and $v = \boldsymbol{prox}_{\eta,r}(w)$, it holds that*

$$r(v) \le r(z) + \frac{1}{\eta}(v - w)^\top(z - v).$$

Proof. For any given w, we have

$$v = \mathbf{prox}_{\eta,r}(w) = \underset{v}{\operatorname{argmin}} \frac{1}{2\eta}\|v - w\|^2 + r(v).$$

Thus, we can obtain that

$$0 \in \partial\left(\frac{1}{2\eta}\|v - w\|^2 + r(v)\right) = -\frac{1}{\eta}(w - v) + \partial r(v),$$

where ∂ here denotes the subgradient. Then we can get that

$$\frac{1}{\eta}(w - v) \in \partial r(v).$$

According to the definition of subgradient, we have for all z,

$$r(z) \ge r(v) - \frac{1}{\eta}(v - w)^\top(z - v).$$

Thus we obtain

$$r(v) \le r(z) + \frac{1}{\eta}(v - w)^\top(z - v),$$

for all z, w and $v = \mathbf{prox}_{\eta,r}(w)$.

$\square$

Lemma 10 (Properties of gradients). *Letting $f(x)$ be L-smooth and each $f_i(x)$ to be M-smooth, we have following properties on the aggregate gradients and generalized gradients.*

$$\|\nabla F(\mathbf{y}) - \nabla F(\mathbf{x})\| \le M\|\mathbf{y} - \mathbf{x}\|, \tag{13}$$

$$\|\bar{g}_t - \nabla f(\bar{y}_t)\| \le \frac{M}{\sqrt{m}}\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\|. \tag{14}$$

Furthermore, if we set $\eta = \frac{1}{L}$, we have the $(3L)$-smooth property for the generalized gradient (defined in Eqn. (12))

$$\left\|\tilde{\nabla}h(x) - \tilde{\nabla}h(y)\right\| \le \left(\frac{2}{\eta} + L\right)\|x - y\| = 3L\|x - y\|. \tag{15}$$

Proof. Eqn. (13) can be obtained from the M-smoothness of the local functions $f_i(x)$ as follows

$$\|\nabla F(\mathbf{y}) - \nabla F(\mathbf{x})\| = \sqrt{\sum_i^m \left\| \nabla f_i(\mathbf{y}^{(i)}) - \nabla f_i(\mathbf{x}^{(i)}) \right\|^2}$$

$$\leq \sqrt{M^2 \sum_i^m \left\| \mathbf{y}^{(i)} - \mathbf{x}^{(i)} \right\|^2}$$

$$= M \|\mathbf{y} - \mathbf{x}\|.$$

To prove the Eqn. (14), we have

$$\|\bar{g}_t - \nabla f(\bar{y}_t)\| = \left\| \frac{1}{m} \sum_{i=0}^m \left[\nabla f_i(\mathbf{y}_t^{(i)}) - \nabla f_i(\bar{y}_t) \right] \right\|$$

$$= \left\| \sum_{i=1}^m \frac{\nabla f_i(\mathbf{y}_t^{(i)}) - \nabla f_i(\bar{y}_t)}{m} \right\|$$

$$\leq M \sum_{i=0}^m \frac{\left\| \mathbf{y}_t^{(i)} - \bar{y}_t \right\|}{m}$$

$$\leq M \sqrt{\sum_{i=0}^m \frac{\left\| \mathbf{y}_t^{(i)} - \bar{y}_t \right\|^2}{m}}$$

$$= M \frac{1}{\sqrt{m}} \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\|,$$

where the first inequality is due to M-smoothness of f_i, and the second inequality is because of Jensen's inequality. Then we can prove Eqn. (15) using L-smoothness of $f(x)$ and the non-expansiveness of proximal operator

$$\left\| \tilde{\nabla} h(x) - \tilde{\nabla} h(y) \right\| = \left\| \frac{x - \mathbf{prox}_{\eta,r}(x - \eta \nabla f(x))}{\eta} - \frac{y - \mathbf{prox}_{\eta,r}(y - \eta \nabla f(y))}{\eta} \right\|$$

$$\leq \frac{1}{\eta} \|x - y\| + \frac{1}{\eta} \left\| \mathbf{prox}_{\eta,r}(x - \eta \nabla f(x)) - \mathbf{prox}_{\eta,r}(y - \eta \nabla f(y)) \right\|$$

$$\leq \frac{1}{\eta} \|x - y\| + \frac{1}{\eta} \|(x - \eta \nabla f(x)) - (y - \eta \nabla f(y))\|$$

$$\leq \left(\frac{2}{\eta} + L \right) \|x - y\|$$

$$= 3L \|x - y\|,$$

where the last inequality is due to the L-smoothness of $f(x)$.

$\square$

Lemma 11. *Let $\mathbf{prox}_{m\eta,R}(\cdot)$ denote the proximal operator defined in Eqn. (4). For any $\mathbf{x} \in \mathbb{R}^{m \times d}$, we have*

$$\left\| \mathbf{prox}_{\eta m,R}(\frac{1}{m} \mathbf{1}\mathbf{1}^\top \mathbf{x}) - \frac{1}{m} \mathbf{1}\mathbf{1}^\top \mathbf{prox}_{\eta m,R}(\mathbf{x}) \right\| \leq \|\mathbf{x} - \mathbf{1}\bar{x}_t\|.$$

Proof. Using Lemma 6, we expand the sum

$$\left\| \mathbf{prox}_{\eta m, R}(\frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{x}) - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{x}) \right\|$$

$$= \sqrt{\left\| \mathbf{prox}_{\eta m, R}(\frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{x}) - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{x}) \right\|^2}$$

$$= \sqrt{m \left\| \mathbf{prox}_{\eta, r}(\frac{1}{m}\mathbf{1}^\top \mathbf{x}) - \frac{1}{m}\sum_{i=1}^{m} \mathbf{prox}_{\eta, r}(\mathbf{x}^{(i)}) \right\|^2}.$$

Using the inequality $(\sum_{i=1}^{m} a_i)^2 \leq m \sum_{i=1}^{m} a_i^2$, and the non-expansiveness of proximal operator, we can obtain that

$$\sqrt{m \left\| \mathbf{prox}_{\eta, r}(\frac{1}{m}\mathbf{1}^\top \mathbf{x}) - \frac{1}{m}\sum_{i=1}^{m} \mathbf{prox}_{\eta, r}(\mathbf{x}^{(i)}) \right\|^2}$$

$$= \sqrt{m \left\| \frac{1}{m}\sum_{i=1}^{m} \left(\mathbf{prox}_{\eta, r}(\frac{1}{m}\mathbf{1}^\top \mathbf{x}) - \mathbf{prox}_{\eta, r}(\mathbf{x}^{(i)}) \right) \right\|^2}$$

$$\leq \sqrt{\sum_{i=1}^{m} \left\| \left(\mathbf{prox}_{\eta, r}(\frac{1}{m}\mathbf{1}^\top \mathbf{x}) - \mathbf{prox}_{\eta, r}(\mathbf{x}^{(i)}) \right) \right\|^2}$$

$$\leq \sqrt{\sum_{i=1}^{m} \left\| \left(\frac{1}{m}\mathbf{1}^\top \mathbf{x} - \mathbf{x}^{(i)} \right) \right\|^2}$$

$$= \| \mathbf{x} - \mathbf{1}\bar{x}_t \|.$$

$\square$

Lemma 12. *Define the estimated generalized gradient*

$$G_t = \eta^{-1} \left(\mathbf{y}_t - \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) \right), and, G_t^{(i)} = \eta^{-1} \left(\mathbf{y}_t^{(i)} - \mathbf{prox}_{\eta, r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) \right). \quad (16)$$

Letting the functions satisfy the properties in Lemma 1, we have following error bound for the estimated generalized gradient

$$\left\| \frac{\eta}{m}\mathbf{1}^\top G_t - \eta \tilde{\nabla} h(\bar{y}_t) \right\| \leq \frac{4 + 2M\eta}{\sqrt{m}} \| \mathbf{y}_t - \mathbf{1}\bar{y}_t \| + \frac{2\eta}{\sqrt{m}} \| \mathbf{s}_t - \mathbf{1}\bar{s}_t \|,$$

$$\left\| \frac{\eta}{m}\mathbf{1}\cdot\mathbf{1}^\top G_t - \eta\mathbf{1}\tilde{\nabla} h(\bar{y}_t) \right\| \leq (4 + 2M\eta) \| \mathbf{y}_t - \mathbf{1}\bar{y}_t \| + 2\eta \| \mathbf{s}_t - \mathbf{1}\bar{s}_t \|,$$

$$\left\| \eta G_t - \eta\mathbf{1}\tilde{\nabla} h(\bar{y}_t) \right\| \leq (7 + 2M\eta) \| \mathbf{y}_t - \mathbf{1}\bar{y}_t \| + 4\eta \| \mathbf{s}_t - \mathbf{1}\bar{s}_t \|,$$

where the generalized gradient $\tilde{\nabla} h(\cdot)$ is defined in Eqn. (12).

Proof. Using Lemma 6, we have the i-th index of G_i is

$$\eta^{-1} \left(\mathbf{y}_t^{(i)} - \mathbf{prox}_{\eta m, R}^{(i)}(\mathbf{y}_t - \eta \mathbf{s}_t) \right)$$

$$= \eta^{-1} \left(\mathbf{y}_t^{(i)} - \mathbf{prox}_{\eta, r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) \right) \quad (17)$$

$$= G_t^{(i)}.$$

Therefore $G_t^{(i)}$ is the i-th index of G_t.

Combining with the inequality $(\sum_{i=1}^{m} a_i)^2 \le m \sum_{i=1}^{m} a_i^2$, we can obtain

$$\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\| = \sqrt{\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\eta G_t^{(i)} - \eta\tilde{\nabla}h(\bar{y}_t)\right)\right\|^2}$$

$$\le \sqrt{\frac{1}{m}\cdot\sum_{i=1}^{m}\left\|\left(\eta G_t^{(i)} - \eta\tilde{\nabla}h(\bar{y}_t)\right)\right\|^2}.$$

By the definition of the estimated generalized gradient G_t and $G_t^{(i)}$, we have

$$\sqrt{\frac{1}{m}\cdot\sum_{i=1}^{m}\left\|\left(\eta G_t^{(i)} - \eta\tilde{\nabla}h(\bar{y}_t)\right)\right\|^2}$$

$$=\sqrt{\frac{1}{m}}\cdot\sqrt{\sum_{i=1}^{m}\left\|\left(\mathbf{y}_t^{(i)} - \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta\mathbf{s}_t^{(i)})\right) - \left(\bar{y}_t - \mathbf{prox}_{\eta,r}(\bar{y}_t - \eta\nabla f(\bar{y}_t))\right)\right\|^2}$$

$$\le\sqrt{\frac{1}{m}}\cdot\sqrt{\sum_{i=1}^{m}\left(2\left\|\mathbf{y}_t^{(i)} - \bar{y}_t\right\|^2 + 2\left\|\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta\mathbf{s}_t^{(i)}) - \mathbf{prox}_{\eta,r}(\bar{y}_t - \eta\nabla f(\bar{y}_t))\right\|^2\right)}$$

$$\le\sqrt{\frac{1}{m}}\cdot\sqrt{\sum_{i=1}^{m}\left(2\left\|\mathbf{y}_t^{(i)} - \bar{y}_t\right\|^2 + 2\left\|(\mathbf{y}_t^{(i)} - \eta\mathbf{s}_t^{(i)}) - (\bar{y}_t - \eta\nabla f(\bar{y}_t))\right\|^2\right)}$$

$$=\sqrt{\frac{1}{m}}\cdot\sqrt{2\left\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\right\|^2 + 2\left\|\eta\mathbf{s}_t - \eta\mathbf{1}\nabla f(\bar{y}_t) + \mathbf{y}_t - \mathbf{1}\bar{y}_t\right\|^2}$$

$$\le 4\left\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\right\| + 2\eta\left\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\right\| + 2\eta\left\|\mathbf{1}\bar{s}_t - \mathbf{1}\nabla f(\bar{y}_t)\right\|$$

$$\le\sqrt{\frac{1}{m}}\cdot\left((4 + 2M\eta)\left\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\right\| + 2\eta\left\|\bar{s}_t - \mathbf{1}\bar{s}_t\right\|\right),$$

where the second inequality is due to the non-expansiveness of proximal operator, and the last inequality is from Eqn. (14). Combining above two inequalities, we can obtain that

$$\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|$$

$$\le\sqrt{\frac{1}{m}}\sqrt{\sum_{i=1}^{m}\left\|\left(\eta G_t^{(i)} - \eta\tilde{\nabla}h(\bar{y}_t)\right)\right\|^2}$$

$$\le\frac{4 + 2M\eta}{\sqrt{m}}\left\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\right\| + \frac{2\eta}{\sqrt{m}}\left\|\bar{s}_t - \mathbf{1}\bar{s}_t\right\|.$$

Because $\left\|\frac{\eta}{m}\mathbf{1}\mathbf{1}^\top G_t - \eta\mathbf{1}\tilde{\nabla}h(\bar{y}_t)\right\|$ is $\sqrt{m}$ times of $\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|$, we can obtain the matrix version of the inequality

$$\left\|\frac{\eta}{m}\mathbf{1}\mathbf{1}^\top G_t - \eta\mathbf{1}\tilde{\nabla}h(\bar{y}_t)\right\| = \sqrt{m}\cdot\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|$$

$$\le(4 + 2M\eta)\cdot\left\|\mathbf{y}_t - \mathbf{1}\bar{y}_t\right\| + 2\eta\left\|\mathbf{s}_t - \mathbf{1}\bar{s}_t\right\|.$$

15

Finally, combining above results, we can obtain the gap of the approximated generalized gradients

$$
\begin{aligned}
&\left\| \eta G_t - \eta \mathbf{1} \tilde{\nabla} h(\bar{y}_t) \right\| \\
&\leq \left\| \eta G_t - \frac{\eta}{m} \mathbf{1} \cdot \mathbf{1}^\top G_t \right\| + \left\| \frac{\eta}{m} \mathbf{1} \cdot \mathbf{1}^\top G_t - \eta \mathbf{1} \tilde{\nabla} h(\bar{y}_t) \right\| \\
&\leq \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + \left\| \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) - \frac{1}{m} \mathbf{1}\mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) \right\| \\
&\quad + (4 + 2M\eta) \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + 2\eta \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\| \\
&\leq \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + \left\| \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) - \mathbf{prox}_{\eta m, R}(\mathbf{1}\bar{y}_t - \eta \mathbf{1}\bar{s}_t) \right\| + 2\eta \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\| \\
&\quad + \left\| \mathbf{prox}_{\eta m, R}(\mathbf{1}\bar{y}_t - \eta \mathbf{1}\bar{s}_t) - \frac{1}{m} \mathbf{1}\mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) \right\| + (4 + 2M\eta) \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| \\
&= (7 + 2M\eta) \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + 4\eta \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|,
\end{aligned}
$$

where the second inequality is from the definition of the G_t and the last inequality is from the non-expansiveness of proximal operator and Lemma 11. □

Lemma 13. *Letting* $\mathbf{s}_t^{(i)}$ *be the i-th row of* $\mathbf{s}_t$ *defined in Algorithm 1, and functions satisfy the properties described in Lemma 1, we have*

$$
\sum_{i=1}^m \left\| \mathbf{s}_t^{(i)} - \nabla f(\mathbf{x}_t^{(i)}) \right\|^2 \leq 2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + 8M^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2.
$$

Proof. Using the inequality that $(a + b)^2 \leq 2a^2 + 2b^2$, we have

$$
\begin{aligned}
\sum_{i=1}^m \left\| \mathbf{s}_t^{(i)} - \nabla f(\mathbf{x}_t^{(i)}) \right\|^2 &\leq 2 \sum_{i=1}^m \left\| \mathbf{s}_t^{(i)} - \bar{s}_t \right\|^2 + 2 \sum_{i=1}^m \left\| \bar{s}_t - \nabla f(\mathbf{x}_t^{(i)}) \right\|^2 \\
&\leq 2 \sum_{i=1}^m \left\| \mathbf{s}_t^{(i)} - \bar{s}_t \right\|^2 + 4 \sum_{i=1}^m \left\| \bar{s}_t - \nabla f(\bar{x}_t) \right\|^2 + 4 \sum_{i=1}^m \left\| \nabla f(\bar{x}_t) - \nabla f(\mathbf{x}_t^{(i)}) \right\|^2 \\
&\leq 2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + 4M^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2 + 4L^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2 \\
&\leq 2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + 8M^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2,
\end{aligned}
$$

where the third inequality is from Eqn. (14) and the L-smoothness of $f(x)$, the last inequality is due to $L \leq M$. □

C   Proof of core lemmas

C.1   Proof of Lemma 1

Proof. Reformulate the Eqn. (5) as $\bar{x}_t = \frac{1}{m} \mathbf{1}^\top \cdot \mathbf{x}_t$, $\bar{y}_t = \frac{1}{m} \mathbf{1}^\top \cdot \mathbf{y}_t$, and $\bar{s}_t = \frac{1}{m} \mathbf{1}^\top \cdot \mathbf{s}_t$, $\bar{g}_t = \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{y}_t^{(i)})$, one can check the following equations by induction and similar results can be found in [26, 14]

$$
\begin{aligned}
\bar{y}_{t+1} &= \bar{x}_{t+1} + \frac{1-\alpha}{1+\alpha}(\bar{x}_{t+1} - \bar{x}_t) \\
\bar{s}_{t+1} &= \bar{g}_{t+1}.
\end{aligned}
\tag{18}
$$

For simplicity, we denote FastMix$(\cdot, K)$ operation as $\mathbb{T}(\cdot)$. From Proposition 1 we can know that

$$
\left\| \mathbb{T}(\mathbf{x}) - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{x} \right\| \leq \rho \left\| \mathbf{x} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{x} \right\|.
\tag{19}
$$

First, we have

$$
\begin{aligned}
\|\mathbf{1}\bar{x}_{t+1} - \mathbf{x}_{t+1}\| \leq &\rho \left\| \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) \right\| \\
\leq &\rho \left\| \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) - \mathbf{prox}_{\eta m, R}\left(\mathbf{1}(\bar{y}_t - \eta \bar{s}_t)\right) \right\| \\
&+ \rho \left\| \mathbf{prox}_{\eta m, R}\left(\mathbf{1}(\bar{y}_t - \eta \bar{s}_t)\right) - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) \right\| \\
\leq &\rho \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + \rho\eta \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + \rho \|(\mathbf{y}_t - \eta \mathbf{s}_t) - \mathbf{1}(\bar{y}_t - \eta \bar{s}_t)\| \\
\leq &2\rho \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + 2\rho\eta \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|,
\end{aligned}
\tag{20}
$$

where the first inequality is due to Eqn. (19), and the third inequality is because of Lemma 11 and the non-expansiveness of proximal operator.

Using the definition of $\mathbf{y}_{t+1}$ in Algorithm 1 and the property of FastMix operation, we have

$$
\begin{aligned}
\|\mathbf{y}_{t+1} - \mathbf{1}\bar{y}_{t+1}\| \leq &\frac{2\rho}{1+\alpha} \|\mathbf{x}_{t+1} - \mathbf{1}\bar{x}_{t+1}\| + \rho\frac{1-\alpha}{1+\alpha} \|\mathbf{x}_t - \mathbf{1}\bar{x}_t\| \\
\leq &4\rho^2 \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + 4\rho^2\eta \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + 2\rho \|\mathbf{x}_t - \mathbf{1}\bar{x}_t\|,
\end{aligned}
$$

where the last inequality is from Eqn. (20).

Now we are going to bound the value of $\|\mathbf{s}_{t+1} - \mathbf{1}\bar{s}_{t+1}\|$. Combining the results we have obtain on $\|\mathbf{y}_{t+1} - \mathbf{1}\bar{y}_{t+1}\|$, we have

$$
\begin{aligned}
&\|\mathbf{s}_{t+1} - \mathbf{1}\bar{s}_{t+1}\| \\
\leq &\rho \|\mathbf{s}_t + \nabla F(\mathbf{y}_{t+1}) - \nabla F(\mathbf{y}_t) - \mathbf{1} \cdot (\bar{s}_t + \bar{g}_{t+1} - \bar{g}_t)\| \\
\leq &\rho \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + \rho M \|\mathbf{y}_{t+1} - \mathbf{y}_t\| \\
\leq &\rho \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + \rho M \|\mathbf{y}_{t+1} - \mathbf{1}\bar{y}_{t+1}\| + \rho M \|\mathbf{1}\bar{y}_{t+1} - \mathbf{1}\bar{y}_t\| + \rho M \|\mathbf{1}\bar{y}_t - \mathbf{y}_t\| \\
\leq &\rho \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + \rho M \left(4\rho^2 \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + 4\rho^2\eta \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + 2\rho \|\mathbf{x}_t - \mathbf{1}\bar{x}_t\|\right) \\
&+ \rho M \|\mathbf{1}\bar{y}_{t+1} - \mathbf{1}\bar{y}_t\| + \rho M \|\mathbf{1}\bar{y}_t - \mathbf{y}_t\| \\
= &\left(\rho + 4\rho^3 M\eta\right) \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + 2\rho^2 M \|\mathbf{x}_t - \mathbf{1}\bar{x}_t\| + \left(\rho M + 4\rho^3 M\right) \|\mathbf{1}\bar{y}_t - \mathbf{y}_t\| + \rho M \|\mathbf{1}\bar{y}_{t+1} - \mathbf{1}\bar{y}_t\|,
\end{aligned}
$$

where the second inequality is because it holds that $\left\|\mathbf{x} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{x}\right\| \leq \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^{m \times d}$ from [14].

Then we only need to consider the term $\|\mathbf{1}\bar{y}_{t+1} - \mathbf{1}\bar{y}_t\|$. Using the iteration of average variables illustrated in Eqn. (18), we have

$$
\begin{aligned}
&\|\mathbf{1}\bar{y}_{t+1} - \mathbf{1}\bar{y}_t\| \\
= &\left\| \frac{2}{1+\alpha}\mathbf{1}\bar{x}_{t+1} - \frac{1-\alpha}{1+\alpha}\mathbf{1}\bar{x}_t - \mathbf{1}\bar{y}_t \right\| \\
= &\left\| \frac{2}{1+\alpha} \cdot \frac{1}{m}\mathbf{1}\mathbf{1}^\top \left(\mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t)\right) - \frac{1-\alpha}{1+\alpha} \cdot \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{x}_t - \frac{1}{m}\mathbf{1}\mathbf{1}^\top \mathbf{y}_t \right\| \\
= &\left\| \frac{2}{1+\alpha} \cdot \frac{1}{m}\mathbf{1}\mathbf{1}^\top \left(\mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) - \mathbf{y}_t\right) - \frac{1-\alpha}{1+\alpha}\mathbf{1}\left(\bar{x}_t - \bar{y}_t\right) \right\| \\
\leq &\frac{2}{1+\alpha} \cdot \left\| \frac{1}{m}\mathbf{1}\mathbf{1}^\top \left(\mathbf{y}_t - \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t)\right) - \eta\mathbf{1} \cdot \tilde{\nabla}h(\bar{y}_t) \right\| \\
&+ \frac{2}{1+\alpha} \left\| \eta\mathbf{1} \cdot \tilde{\nabla}h(\bar{y}_t) \right\| + \frac{1-\alpha}{1+\alpha} \left(\|\mathbf{1}\bar{x}_t - \mathbf{1}x^*\| + \|\mathbf{1}\bar{y}_t - \mathbf{1}x^*\|\right).
\end{aligned}
$$

Furthermore, by Lemma 12, we can obtain that

$$
\frac{2}{1+\alpha} \cdot \left\| \frac{1}{m}\mathbf{1}\mathbf{1}^\top \left(\mathbf{y}_t - \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t)\right) - \eta\mathbf{1} \cdot \tilde{\nabla}h(\bar{y}_t) \right\| \leq (8+4M\eta) \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + 4\eta \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|.
$$

Furthermore, by Lemma 8 and the fact that $\tilde{\nabla}h(x^*) = 0$, we can obtain

$$\frac{2}{1+\alpha} \left\| \eta\mathbf{1} \cdot \tilde{\nabla}h(\bar{y}_t) \right\| + \frac{1-\alpha}{1+\alpha} \left(\|\mathbf{1}\bar{x}_t - \mathbf{1}x^*\| + \|\mathbf{1}\bar{y}_t - \mathbf{1}x^*\| \right)$$

$$\leq \frac{2\eta\sqrt{m}}{1+\alpha} \left\| \tilde{\nabla}h(\bar{y}_t) - \tilde{\nabla}h(x^*) \right\| + \frac{1-\alpha}{1+\alpha} \cdot 2 \cdot \sqrt{\frac{2}{\mu}V_t}$$

$$\leq \left(\frac{6\eta\sqrt{m}L}{1+\alpha} + \frac{1-\alpha}{1+\alpha} \cdot 2 \right) \sqrt{\frac{2}{\mu}V_t}$$

$$\leq \left(6\sqrt{m} + 2 \right) \sqrt{\frac{2}{\mu}V_t},$$

where the second inequality is because of the Eqn.(15). Thus, we can obtain that

$$\|\mathbf{1}\bar{y}_{t+1} - \mathbf{1}\bar{y}_t\| \leq (8 + 4M\eta) \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + 4\eta \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + \left(6\sqrt{m} + 2 \right) \sqrt{\frac{2}{\mu}V_t}.$$

Combining above results, we can bound the value of $\|\mathbf{s}_{t+1} - \mathbf{1}\bar{s}_{t+1}\|$ as follows

$$\|\mathbf{s}_{t+1} - \mathbf{1}\bar{s}_{t+1}\| \leq \left(\rho + 4\rho^3 M\eta \right) \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + 2\rho^2 M \|\mathbf{x}_t - \mathbf{1}\bar{x}_t\| + \left(\rho M + 4\rho^3 M \right) \|\mathbf{1}\bar{y}_t - \mathbf{y}_t\|$$

$$+ \rho M \|\mathbf{1}\bar{y}_{t+1} - \mathbf{1}\bar{y}_t\|$$

$$\leq \left(\rho + 4\rho^3 M\eta \right) \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + 2\rho^2 M \|\mathbf{x}_t - \mathbf{1}\bar{x}_t\| + \left(\rho M + 4\rho^3 M \right) \|\mathbf{1}\bar{y}_t - \mathbf{y}_t\|$$

$$+ \rho M \left((8 + 4M\eta) \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\| + 4\eta \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + \left(6\sqrt{m} + 2 \right) \sqrt{\frac{2}{\mu}V_t} \right)$$

$$\leq \left(\rho + 4\rho^3 M\eta + 4\rho M\eta \right) \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\| + 2\rho^2 M \|\mathbf{x}_t - \mathbf{1}\bar{x}_t\|$$

$$+ \left(9\rho M + 4\rho^3 M + 4\rho M^2\eta \right) \|\mathbf{1}\bar{y}_t - \mathbf{y}_t\| + 8\rho M \sqrt{m} \sqrt{\frac{2}{\mu}V_t}.$$

If we denote $\mathbf{z}_t = [\|\mathbf{x}_t - \mathbf{1}\bar{x}_t\|, \|\mathbf{y}_t - \mathbf{1}\bar{y}_t\|, \|\mathbf{s}_t - \mathbf{1}\bar{s}_t\|]^\top$, then due to $\rho < 1$, $L \leq M$, and $1 \leq m$, we can have

$$\mathbf{z}_{t+1} \leq \mathbf{A}'\mathbf{z}_t + 8\rho M \sqrt{m} [0, 0, \sqrt{\frac{2}{\mu}V_t}]^\top$$

where

$$\mathbf{A}' = \begin{pmatrix} 0 & 2\rho & 2\rho\eta \\ 2\rho & 4\rho^2 & 4\rho^2\eta \\ 2\rho^2 M & 9\rho M + 4\rho^3 M + 4\rho M^2\eta & \rho + 4\rho^3 M\eta + 4\rho M\eta \end{pmatrix}$$

$$\leq \begin{pmatrix} 0 & 2\rho & 2\rho\eta \\ 2\rho & 4\rho & 4\rho\eta \\ 2\rho M & \rho M(13 + 4M\eta) & \rho(1 + 6M\eta) \end{pmatrix}$$

$$= \rho \begin{pmatrix} 0 & 2 & 2\eta \\ 2 & 4 & 4\eta \\ 2M & M(13 + 4M\eta) & 1 + 6M\eta \end{pmatrix} = \mathbf{A}$$

Finally, we simplify the result and obtain that

$$\mathbf{z}_{t+1} \leq \mathbf{A}\mathbf{z}_t + 8\rho M \sqrt{m} [0, 0, \sqrt{\frac{2}{\mu}V_t}]^\top.$$

$\square$

C.2 Proof of Lemma 2

Proof. By the μ-strongly convexity , L-smoothness of $f(x)$ and the property of proximal operator showed in Lemma 9, we can have for any given $z \in \mathbb{R}^d$

$$
\begin{aligned}
&h(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)})) = f(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)})) + r(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}))\\
&= f(\mathbf{y}_t^{(i)} + \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)}) + r(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}))\\
&\leq f(\mathbf{y}_t^{(i)}) + \nabla f(\mathbf{y}_t^{(i)})^\top \left(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right) + \frac{L}{2} \left\| \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right\|^2\\
&\quad + r(z) + \frac{1}{\eta}(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} + \eta \mathbf{s}_t^{(i)})^\top (z - \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}))\\
&\leq h(z) - \nabla f(\mathbf{y}_t^{(i)})^\top (z - \mathbf{y}_t^{(i)}) - \frac{\mu}{2} \left\| z - \mathbf{y}_t^{(i)} \right\|^2 + \nabla f(\mathbf{y}_t^{(i)})^\top \left(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right)\\
&\quad + \frac{L}{2} \left\| \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right\|^2 + \frac{1}{\eta}(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} + \eta \mathbf{s}_t^{(i)})^\top (z - \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)})).
\end{aligned}
\tag{21}
$$

Multiplying $1 - \alpha$ on both sides of Eqn. (21) and setting $z = \bar{x}_t$, we get

$$
\begin{aligned}
&(1 - \alpha)h(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}))\\
&\leq (1 - \alpha)h(\bar{x}_t) - (1 - \alpha)\nabla f(\mathbf{y}_t^{(i)})^\top (\bar{x}_t - \mathbf{y}_t^{(i)}) - \frac{\mu(1 - \alpha)}{2} \left\| \bar{x}_t - \mathbf{y}_t^{(i)} \right\|^2\\
&\quad + (1 - \alpha)\nabla f(\mathbf{y}_t^{(i)})^\top \left(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right) + \frac{L(1 - \alpha)}{2} \left\| \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right\|^2\\
&\quad + \frac{(1 - \alpha)}{\eta}(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} + \eta \mathbf{s}_t^{(i)})^\top (\bar{x}_t - \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)})).
\end{aligned}
\tag{22}
$$

Similarly multiplying α on both sides of Eqn. (21) and setting $z = x^*$, we obtain that

$$
\begin{aligned}
&\alpha(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}))\\
&\leq \alpha h(x^*) - \alpha \nabla f(\mathbf{y}_t^{(i)})^\top (x^* - \mathbf{y}_t^{(i)}) - \frac{\mu \alpha}{2} \left\| x^* - \mathbf{y}_t^{(i)} \right\|^2\\
&\quad + \alpha \nabla f(\mathbf{y}_t^{(i)})^\top \left(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right) + \frac{\alpha L}{2} \left\| \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} \right\|^2\\
&\quad + \frac{\alpha}{\eta}(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)}) - \mathbf{y}_t^{(i)} + \eta \mathbf{s}_t^{(i)})^\top (x^* - \mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)})).
\end{aligned}
\tag{23}
$$

Adding above two inequalities and by the definition of G_t and $G_t^{(i)}$ in Lemma 12, we have

$$
\begin{aligned}
&h(\mathbf{prox}_{\eta,r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)})) - h(x^*)\\
&\leq (1 - \alpha)\left(h(\bar{x}_t) - h(x^*)\right) - \alpha \cdot \frac{\mu}{2} \left\| x^* - \mathbf{y}_t^{(i)} \right\|^2 + \frac{1}{2L} \left\| G_t^{(i)} \right\|^2\\
&\quad + \left(\mathbf{s}_t^{(i)} - G_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)}) \right)^\top \left((1 - \alpha)\bar{x}_t + \alpha x^* + \eta G_t^{(i)} - \mathbf{y}_t^{(i)} \right).
\end{aligned}
\tag{24}
$$

Note that by Jensen's inequality, we can get that

$$
\|x^* - \bar{y}_t\| = \left\| x^* - \frac{1}{m}\sum_{i=1}^m \mathbf{y}_t^{(i)} \right\| \leq \sqrt{\frac{1}{m}\sum_{i=1}^m \left\| x^* - \mathbf{y}_t^{(i)} \right\|^2}.
$$

Then averaging Eqn. (24) from $i = 1$ to $i = m$, we have

$$
h(\bar{x}_{t+1}) - h(x^*)
$$

$$
\leq \frac{1}{m} \sum_{i=1}^{m} h(\mathbf{prox}_{\eta, r}(\mathbf{y}_t^{(i)} - \eta \mathbf{s}_t^{(i)})) - h(x^*)
$$

$$
\leq (1 - \alpha)(h(\bar{x}_t) - h(x^*)) + \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2L} \left\| G_t^{(i)} \right\|^2
$$

$$
+ \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{s}_t^{(i)} - G_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)}) \right)^\top \left((1-\alpha)\bar{x}_t + \alpha x^* + \eta G_t^{(i)} - \mathbf{y}_t^{(i)} \right) - \alpha \cdot \frac{\mu}{2} \| x^* - \bar{y}_t \|^2
$$

$$
= (1 - \alpha)(h(\bar{x}_t) - h(x^*)) - \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2L} \left\| G_t^{(i)} \right\|^2 + \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)}) \right)^\top ((1-\alpha)\bar{x}_t + \alpha x^* - \bar{y}_t)
$$

$$
- \frac{1}{m} \sum_{i=1}^{m} (G_t^{(i)} - \tilde{\nabla} h(\bar{y}_t))^\top ((1-\alpha)\bar{x}_t + \alpha x^* - \bar{y}_t) + \frac{\eta}{m} \sum_{i=1}^{m} \left(\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)}) \right)^\top \left(G_t^{(i)} - \tilde{\nabla} h(\bar{y}_t) \right)
$$

$$
+ \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)}) \right)^\top \left(\bar{y}_t - \mathbf{y}_t^{(i)} \right) - \frac{1}{m} \sum_{i=1}^{m} (G_t^{(i)} - \tilde{\nabla} h(\bar{y}_t))^\top \left(\bar{y}_t - \mathbf{y}_t^{(i)} \right)
$$

$$
- \frac{1}{m} \sum_{i=1}^{m} \tilde{\nabla} h(\bar{y}_t)^\top \left(((1-\alpha)\bar{x}_t + \alpha x^* - \bar{y}_t) - \eta \left(\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)}) \right) + \left(\bar{y}_t - \mathbf{y}_t^{(i)} \right) \right) - \alpha \cdot \frac{\mu}{2} \| x^* - \bar{y}_t \|^2,
$$

(25)

where the first inequality is because of the convexity of $h(x)$, the second inequality is from Eqn. (24), and the last equality is a reorganization of above expression.

For $\left\| G_t^{(i)} \right\|^2$, we have following equality

$$
\left\| G_t^{(i)} \right\|^2 = \left\| G_t^{(i)} - \tilde{\nabla} h(\bar{y}_t) + \tilde{\nabla} h(\bar{y}_t) \right\|^2
$$

$$
= \left\| G_t^{(i)} - \tilde{\nabla} h(\bar{y}_t) \right\|^2 + 2 \left\langle G_t^{(i)} - \tilde{\nabla} h(\bar{y}_t), \tilde{\nabla} h(\bar{y}_t) \right\rangle + \left\| \tilde{\nabla} h(\bar{y}_t) \right\|^2.
$$

(26)

Now we bound another term of V_{t+1} as follows

$$
\frac{\mu}{2} \| \bar{v}_{t+1} - x^* \|^2
$$

$$
= \frac{\mu}{2} \left\| \bar{x}_t + \frac{1}{\alpha}(\bar{x}_{t+1} - \bar{x}_t) - x^* \right\|^2
$$

$$
= \frac{\mu}{2} \left\| \bar{x}_t + \frac{1}{\alpha}(\frac{1}{m} \mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) - \bar{x}_t) - x^* \right\|^2
$$

$$
\leq \frac{\mu}{2} \left\| \bar{x}_t + \frac{1}{\alpha}(\bar{y}_t - \eta \tilde{\nabla} h(\bar{y}_t) - \bar{x}_t) - x^* \right\|^2 + \frac{\mu}{2} \left\| \bar{y}_t - \eta \tilde{\nabla} h(\bar{y}_t) - \frac{1}{m} \mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) \right\|^2
$$

$$
+ \mu \left\| \bar{x}_t + \frac{1}{\alpha}(\bar{y}_t - \eta \tilde{\nabla} h(\bar{y}_t) - \bar{x}_t) - x^* \right\| \cdot \left\| \bar{y}_t - \eta \tilde{\nabla} h(\bar{y}_t) - \frac{1}{m} \mathbf{1}^\top \mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta \mathbf{s}_t) \right\|.
$$

Because of $\bar{y}_t - \bar{x}_t = \alpha(\bar{y}_t - \bar{v}_t)$, we have

$$
\bar{x}_t + \frac{1}{\alpha}(\bar{y}_t - \eta \tilde{\nabla} h(\bar{y}_t) - \bar{x}_t) - x^* = \bar{v}_t - \alpha(\bar{v}_t - \bar{y}_t) - \frac{\eta}{\alpha} \tilde{\nabla} h(\bar{y}_t) - x^*
$$

$$
= (1 - \alpha)(\bar{v}_t - x^*) + \alpha(\bar{y}_t - x^*) - \frac{\eta}{\alpha} \tilde{\nabla} h(\bar{y}_t).
$$

20

Thus, we can obtain

$$\frac{\mu}{2}\left\|\bar{x}_t + \frac{1}{\alpha}(\bar{y}_t - \eta\tilde{\nabla}h(\bar{y}_t) - \bar{x}_t) - x^*\right\|^2$$

$$=\frac{\mu - \mu\alpha}{2}\|\bar{v}_t - x^*\|^2 + \frac{\mu\alpha}{2}\|\bar{y}_t - x^*\|^2 - \alpha\left\langle\tilde{\nabla}h(\bar{y}_t), (1-\alpha)\bar{v}_t + \alpha\bar{y}_t - x^*\right\rangle$$

$$+ \frac{1}{2L}\left\|\tilde{\nabla}h(\bar{y}_t)\right\|^2 - \frac{\alpha(\mu - \mu\alpha)}{2}\|\bar{y}_t - \bar{v}_t\|^2$$

$$\leq\frac{\mu - \mu\alpha}{2}\|\bar{v}_t - x^*\|^2 + \frac{\mu\alpha}{2}\|\bar{y}_t - x^*\|^2 - \alpha\left\langle\tilde{\nabla}h(\bar{y}_t), (1-\alpha)\bar{v}_t + \alpha\bar{y}_t - x^*\right\rangle + \frac{1}{2L}\left\|\tilde{\nabla}h(\bar{y}_t)\right\|^2$$

$$=\frac{\mu - \mu\alpha}{2}\|\bar{v}_t - x^*\|^2 + \frac{\mu\alpha}{2}\|\bar{y}_t - x^*\|^2 + \left\langle\tilde{\nabla}h(\bar{y}_t), \alpha x^* + (1-\alpha)\bar{x}_t - \bar{y}_t\right\rangle + \frac{1}{2L}\left\|\tilde{\nabla}h(\bar{y}_t)\right\|^2,$$

where the last equality comes from Lemma 7. We can further obtain that

$$\frac{\mu}{2}\|\bar{v}_{t+1} - x^*\|^2$$

$$=\frac{\mu}{2}\left\|\bar{x}_t + \frac{1}{\alpha}(\bar{y}_t - \eta\tilde{\nabla}h(\bar{y}_t) - \bar{x}_t) - x^*\right\|^2 + \frac{\mu}{2}\left\|\bar{y}_t - \eta\tilde{\nabla}h(\bar{y}_t) - \frac{1}{m}\mathbf{1}^\top\mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta\mathbf{s}_t)\right\|^2$$

$$+ \frac{\mu}{2}\left\langle\bar{x}_t + \frac{1}{\alpha}(\bar{y}_t - \eta\tilde{\nabla}h(\bar{y}_t) - \bar{x}_t) - x^*, \bar{y}_t - \eta\tilde{\nabla}h(\bar{y}_t) - \frac{1}{m}\mathbf{1}^\top\mathbf{prox}_{\eta m, R}(\mathbf{y}_t - \eta\mathbf{s}_t)\right\rangle$$

$$\leq\frac{\mu - \mu\alpha}{2}\|\bar{v}_t - x^*\|^2 + \frac{\mu\alpha}{2}\|\bar{y}_t - x^*\|^2 + \left\langle\tilde{\nabla}h(\bar{y}_t), \alpha x^* + (1-\alpha)\bar{x}_t - \bar{y}_t\right\rangle + \frac{1}{2L}\left\|\tilde{\nabla}h(\bar{y}_t)\right\|^2$$

$$+ \frac{\mu}{2}\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|^2 + \frac{\mu}{2}\left\|\bar{x}_t + \frac{1}{\alpha}(\bar{y}_t - \eta\tilde{\nabla}h(\bar{y}_t) - \bar{x}_t) - x^*\right\|\cdot\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|.$$

$$(27)$$

Recalling for any $a > 0, b > 0, \{a_i > 0\}$, it holds that $ab \leq \frac{a^2 + b^2}{2}$, $\frac{1}{m}\sum_{i=1}^m a_i \leq \sqrt{\frac{1}{m}\sum_{i=1}^m a_i^2}$, and $(a+b)^2 \leq 2a^2 + 2b^2$. Combining with Cauchy's inequality, Eqn. (25), Eqn. (26) and Eqn. (27), we can obtain that

$$V_{t+1}$$

$$=h(\bar{x}_t) - h(x^*) + \frac{\mu}{2}\|\bar{x}_t - x^*\|^2$$

$$\leq(1-\alpha)V_t + \left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|\cdot\left\|\tilde{\nabla}h(\bar{y}_t)\right\|$$

$$+ \sqrt{\frac{1}{m}\sum_{i=1}^m\left\|\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)})\right\|^2}\cdot\|(1-\alpha)\bar{x}_t + \alpha x^* - \bar{y}_t\|$$

$$+ \left\|\frac{1}{m}\mathbf{1}^\top G_t - \tilde{\nabla}h(\bar{y}_t)\right\|\cdot\|(1-\alpha)\bar{x}_t + \alpha x^* - \bar{y}_t\| + \frac{2\eta}{m}\sum_{i=1}^m\left\|\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)})\right\|^2$$

$$+ \frac{2}{m}\sum_{i=1}^m\left\|\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)})\right\|^2 + \frac{4}{m}\|\mathbf{1}\bar{y}_t - \mathbf{y}_t\|^2 + \frac{2}{m}\left\|G_t - \mathbf{1}\tilde{\nabla}h(\bar{y}_t)\right\|^2 + \frac{2\eta}{m}\left\|G_t - \mathbf{1}\tilde{\nabla}h(\bar{y}_t)\right\|^2$$

$$+ \left\|\tilde{\nabla}h(\bar{y}_t)\right\|\left(\eta\sqrt{\frac{1}{m}\sum_{i=1}^m\left\|\mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)})\right\|^2}\right) + \frac{\mu}{2}\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|^2$$

$$+ \frac{\mu}{2}\left((1-\alpha)\|\bar{v}_t - x^*\| + \alpha\|\bar{y}_t - x^*\| + \frac{\eta}{\alpha}\left\|\tilde{\nabla}h(\bar{y}_t)\right\|\right)\cdot\left\|\frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t)\right\|.$$

$$(28)$$

By the lemma 8, 10, 11, 12, and 13, we can collect the following inequalities

$$\left\| G_t - \mathbf{1}\tilde{\nabla}h(\bar{y}_t) \right\|^2 \leq 2\,(2M+7L)^2 \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\|^2 + 32\eta^2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2$$

$$\left\| \frac{\eta}{m}\mathbf{1}^\top G_t - \eta\tilde{\nabla}h(\bar{y}_t) \right\| \leq \frac{4+2M\eta}{\sqrt{m}} \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + \frac{2\eta}{\sqrt{m}} \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|$$

$$\sum_{i=1}^{m} \left\| \mathbf{s}_t^{(i)} - \nabla f(\mathbf{y}_t^{(i)}) \right\|^2 \leq 2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + 8M^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2$$

$$\left\| \tilde{\nabla}h(\bar{y}_t) \right\| \leq 3L\sqrt{\frac{2}{\mu}V_t}$$

$$\left\| (1-\alpha)\bar{x}_t + \alpha x^* - \bar{y}_t \right\| \leq \sqrt{\frac{2}{\mu}V_t}.$$

Replacing above inequalities to Eqn. (28), we can obtain

V_{t+1}

$$\leq (1-\alpha)V_t + \left(\frac{4+2M\eta}{\sqrt{m}} \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + \frac{2\eta}{\sqrt{m}} \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\| \right) \cdot \left(3L\sqrt{\frac{2}{\mu}V_t} \right)$$

$$+ \frac{1}{\sqrt{m}} \cdot \sqrt{2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + 8M^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2} \cdot \sqrt{\frac{2}{\mu}V_t}$$

$$+ L\left(\frac{4+2M\eta}{\sqrt{m}} \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + \frac{2\eta}{\sqrt{m}} \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\| \right) \cdot \sqrt{\frac{2}{\mu}V_t} + \frac{2(1+\eta)}{m} \left(2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + 8M^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2 \right)$$

$$+ \frac{4}{m} \left\| \mathbf{1}\bar{y}_t - \mathbf{y}_t \right\|^2 + \frac{2(1+\eta)}{m} \left(2\,(2M+7L)^2 \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\|^2 + 32\eta^2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 \right)$$

$$+ 3\sqrt{\frac{2}{\mu}V_t} \cdot \sqrt{\frac{1}{m}\left(2 \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + 8M^2 \left\| \mathbf{1}\bar{x}_t - \mathbf{x}_t \right\|^2 \right)} + \frac{\mu}{2} \left(\frac{4+2M\eta}{\sqrt{m}} \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + \frac{2\eta}{\sqrt{m}} \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\| \right)^2$$

$$+ \frac{\mu}{2}\left(\sqrt{\frac{2}{\mu}V_t} + \frac{\eta}{\alpha} \cdot 3L\sqrt{\frac{2}{\mu}V_t} \right) \cdot \left(\frac{4+2M\eta}{\sqrt{m}} \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\| + \frac{2\eta}{\sqrt{m}} \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\| \right).$$

$$\leq (1-\alpha)V_t + \left(\frac{4(4+2M\eta)}{\sqrt{m}\eta} + \frac{\mu}{2}(\frac{3}{\alpha}+1)\cdot\frac{4+2M\eta}{\sqrt{m}} \right) \sqrt{\frac{2}{\mu}V_t} \cdot \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\|$$

$$+ \frac{8\sqrt{2}L}{\sqrt{m}} \cdot \sqrt{\frac{2}{\mu}V_t} \cdot \left\| \mathbf{x}_t - \mathbf{1}\bar{x}_t \right\| + \left(\frac{8}{\sqrt{m}} + \frac{4\sqrt{2}}{\sqrt{m}} + \frac{\mu}{2}(\frac{3}{\alpha}+1)\cdot\frac{2\eta}{\sqrt{m}} \right) \sqrt{\frac{2}{\mu}V_t} \cdot \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|$$

$$+ \frac{(2\mu+68+68\eta)\,\eta^2}{m} \left\| \mathbf{s}_t - \mathbf{1}\bar{s}_t \right\|^2 + \frac{16\,(\eta+1)\,L^2}{m} \left\| \mathbf{x}_t - \mathbf{1}\bar{x}_t \right\|^2$$

$$+ \left(\frac{4}{m} + \frac{\mu(4+2M\eta)^2}{m} + \frac{4(1+\eta)(2M+7L)^2}{m} \right) \left\| \mathbf{y}_t - \mathbf{1}\bar{y}_t \right\|^2$$

$$\leq (1-\alpha)V_t + \frac{4}{\sqrt{m}}\left(21L + 2\sqrt{L\mu}(2+2M\eta) + 3\mu(2+2M\eta) + 8 + \frac{\mu}{L} + 3\sqrt{\frac{\mu}{L}} \right) \cdot \left\| \mathbf{z}_t \right\| \sqrt{\frac{2}{\mu}V_t}$$

$$+ \frac{8}{m}(1+\mu) \cdot \left(9\cdot(12+8L+4M)^2 + 16L(L+1) + \frac{133+79L}{L} \right) \left\| \mathbf{z}_t \right\|^2,$$

where the last inequality is because of Cauchy's inequality and $L \leq M$. $\qquad\square$

C.3   Proof of Lemma 3

Proof. It is easy to check that spectral norm of a matrix with all the indices positive will be less that the sum of all its indices, then we have $\left\| \mathbf{A} \right\|_2 \leq \left\| \mathbf{A} \right\|_F \leq \rho D_3$, where $\left\| \mathbf{A} \right\|_F$ is the Frobenius norm of $\mathbf{A}$ and

$$D_3 = \sum_{i,j} \mathbf{A}_{i,j}/\rho = (9 + 6\eta + 15M + 4M^2\eta + 6M\eta).$$

Therefore, if $\rho < \frac{1}{2D_3}$, we will have $\|\mathbf{A}\|_2 < \frac{1}{2}$. □

C.4 Proof of Lemma 4

Our proof is based on the fact that error term $\|\mathbf{z}_t\|$ in Eqn. (8) can be controlled by ρ. In total, what we will do in this lemma are all about bounding the term $D_1 \sqrt{V_t} \cdot \|\mathbf{z}_t\|$ and $D_2 \|\mathbf{z}_t\|^2$ to the level of $\frac{\alpha}{2} V_t$ in an induction way.

Proof. We will prove our result by induction. First, when $t = 0$, by Lemma 2 and the definition of C, we have

$$
\begin{aligned}
V_1 &\leq (1-\alpha)V_0 + D_1 \|\mathbf{z}_0\| \sqrt{V_0} + D_2 \|\mathbf{z}_0\|^2 \\
&\leq (1-\alpha)V_0 + \frac{1}{2}\alpha V_0 + \frac{D_1^2}{2\alpha} \cdot \|\mathbf{z}_0\|^2 + D_2 \|\mathbf{z}_0\|^2 \\
&\leq \left(1 - \frac{\alpha}{2}\right) \left(V_0 + 2\left(\frac{D_1^2}{2\alpha} + D_2\right)\|\mathbf{z}_0\|^2\right) \\
&= \left(1 - \frac{\alpha}{2}\right) \left(V_0 + C\|\mathbf{z}_0\|^2\right).
\end{aligned}
$$

Therefore, we have that Eqn. (9) holds for $t = 1$. Next, we assume that for $i = 1, \ldots, t$, it holds that

$$
V_t \leq \left(1 - \frac{\alpha}{2}\right)^t \left(V_0 + C\|\mathbf{z}_0\|^2\right).
$$

By Lemma 3, we known that the spectral norm of $\mathbf{A}$ is upper bounded by $\frac{1}{2}$ and ρD_3. Then we can obtain that

$$
\begin{aligned}
\|\mathbf{z}_t\| &\leq 8M\rho \sqrt{\frac{2m}{\mu}} \sum_{j=0}^{t-1} 2^{-(t-1-j)} \sqrt{V_j} + 2^{-(t-1)} \cdot \rho D_3 \|\mathbf{z}_0\| \\
&\leq 8M\rho \sqrt{\frac{2m}{\mu}} \sum_{j=0}^{t-1} 2^{-(t-1-j)} \left(\sqrt{1-\frac{\alpha}{2}}\right)^j \left(\sqrt{V_0} + \sqrt{C}\|\mathbf{z}_0\|\right) + 2^{-t} \cdot (2\rho D_3 \|\mathbf{z}_0\|) \\
&= 8M\rho \sqrt{\frac{2m}{\mu}} \frac{2\left(\sqrt{1-\frac{\alpha}{2}}\right)^t - 2^{-(t-1)}}{2\sqrt{1-\frac{\alpha}{2}} - 1} \left(\sqrt{V_0} + \sqrt{C}\|\mathbf{z}_0\|\right) + 2^{-t} \cdot (2\rho D_3 \|\mathbf{z}_0\|) \\
&\leq 24M\rho \sqrt{\frac{2m}{\mu}} \left(\sqrt{1-\frac{\alpha}{2}}\right)^t \left(\sqrt{V_0} + \sqrt{C}\|\mathbf{z}_0\|\right) + \left(\sqrt{1-\frac{\alpha}{2}}\right)^t \cdot (2\rho D_3 \|\mathbf{z}_0\|) \\
&\leq \left(\sqrt{1-\frac{\alpha}{2}}\right)^t \left(24M\sqrt{\frac{2m}{\mu}} \left(\sqrt{V_0} + \sqrt{C}\|\mathbf{z}_0\|\right) + 2\rho D_3 \|\mathbf{z}_0\|\right) \\
&\leq \rho \cdot D_4 \left(\sqrt{1-\frac{\alpha}{2}}\right)^t \left(\sqrt{V_0} + \sqrt{C}\|\mathbf{z}_0\|\right),
\end{aligned}
$$

where $D_4 = 24M\sqrt{\frac{2m}{\mu}} + 2D_3$, and the first inequality is from splitting $\|\mathbf{A}\|_2^t = \|\mathbf{A}\|_2^{t-1} \cdot \|\mathbf{A}\|_2 \leq \left(\frac{1}{2}\right)^{t-1} \cdot \rho D_3$.

Therefore, we can obtain that

$$
\|\mathbf{z}_t\|^2 \leq 2\rho \cdot D_4^2 \left(1 - \frac{\alpha}{2}\right)^t (V_0 + C\|\mathbf{z}_0\|^2).
$$

Combining with the induction hypothesis that $V_t \leq \left(1 - \frac{\alpha}{2}\right)^t \left(V_0 + C\|\mathbf{z}_0\|^2\right)$, we have

$$
\begin{aligned}
\sqrt{V_t} \cdot \|\mathbf{z}_t\| &\leq \rho D_4 \left(1 - \frac{\alpha}{2}\right)^t \sqrt{V_0 + C\|\mathbf{z}_0\|^2} \cdot \left(\sqrt{V_0} + \sqrt{C}\|\mathbf{z}_0\|\right) \\
&\leq 2\rho D_4 \left(1 - \frac{\alpha}{2}\right)^t \left(V_0 + C\|\mathbf{z}_0\|^2\right).
\end{aligned}
$$

Using Eqn. (8), we can get if $\rho < \frac{\alpha}{2\left(D_1 D_4 + D_2 D_4^2\right)}$,

$$
\begin{aligned}
V_{t+1} \leq &(1-\alpha)V_t + D_1\sqrt{V_t} \cdot \|\mathbf{z}_t\| + D_2\|\mathbf{z}_t\|^2 \\
\leq &(1-\alpha)\left(1-\frac{\alpha}{2}\right)^t\left(V_0 + C\|\mathbf{z}_0\|^2\right) \\
&+ \rho \cdot \left(D_1 D_4 + D_2 D_4^2\right)\left(1-\frac{\alpha}{2}\right)^t\left(V_0 + C\|\mathbf{z}_0\|^2\right) \\
\leq &\left(1-\frac{\alpha}{2}\right)^{t+1}\left(V_0 + C\|\mathbf{z}_0\|^2\right).
\end{aligned}
$$

$\square$