Fast Stochastic ADMMs for Large-Scale Nonconvex Composition Data Mining

ABSTRACT

Stochastic composition optimization has been successful in many applications of data mining and machine learning such as portfolio management, policy evaluation, and recommendation systems. Although some works have devoted to study stochastic composition optimization, few focus on nonconvex nonsmooth composition problems, which are popular in many composition learning applications. In this paper, thus, we focus on the large-scale nonconvex composition problems with multiple nonsmooth regularization penalties, and propose a class of fast compositional stochastic Alternating Direction Method of Multipliers (ADMM) methods (*i.e.,* com-SVRG-ADMM and com-SAGA-ADMM) with using the variance reduced techniques. Moreover, we prove that both our com-SVRG-ADMM and com-SAGA-ADMM have convergence rate of $O(\frac{1}{T})$, where T denotes the iterate number. Our theoretical analysis also shows that the com-SVRG-ADMM and com-SAGA-ADMM have the optimal query complexities of $O(n_1^{2/3}\epsilon^{-1})$ and $O(n_2^{2/3}\epsilon^{-1})$ for finding an ϵ-approximate solution, respectively, where n_1 and n_2 denote the outer and inner sample sizes. In particular, the optimal query complexity of our methods improves the existing best results in the nonconvex nonsmooth problems by a factor of $O(\max(n_1^{2/3}, n_2^{2/3}))$. We apply the proposed algorithms to relationship-guided portfolio management and policy evaluation data mining applications. The empirical results validate that our methods have faster convergence rate than the existing nonconvex composition stochastic algorithms.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

ADMM, Stochastic composition optimization

ACM Reference Format:
. 2019. Fast Stochastic ADMMs for Large-Scale Nonconvex Composition Data Mining. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Stochastic composition optimization has been successful in many applications of data mining and machine learning such as portfolio management[19, 29], policy evaluation, recommendation systems and sparse additive models [29]. Recently, Wang et al. [29] have formally proposed stochastic composition problem as follows:

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_\xi\big[F(\mathbb{E}_\zeta[G(x,\zeta)],\xi)\big], \tag{1}$$

where $G(x):\mathbb{R}^d \to \mathbb{R}^p$ is inner component function parameterized by random variable ζ, $F(y)\in\mathbb{R}^p \to \mathbb{R}$ is outer component function parameterized by random variable ξ. In fact, we generally solve the finite-sum setting for the composition problem (1), which can be represented as follows:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n_1}\sum_{i=1}^{n_1} F_i\big(\frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x)\big) + \psi(x), \tag{2}$$

where $\psi(x)$ is a regularized penalty.

The above composition problems have many data mining applications. For example, the following three data mining tasks:

- **Risk-averse Learning Problem** can be formulated into the following mean-variance minimization problem:
$$\min_x \mathbb{E}_\xi[f_\xi(x)] + \tau\mathbb{E}_\xi\big[\big(f_\xi(x)-\mathbb{E}_\xi[f_\xi(x)]\big)^2\big], \tag{3}$$
where $\tau > 0$, $f_\xi(x)$ is the loss function, and the above second term denotes its variance. This problem is applied for portfolio management [1].
- **Policy Evaluation** for Markov Decision Processes can be formulated as a Bellman residual minimization problem:
$$\min_{V\in\mathbb{R}^N} \mathbb{E}\big[\big(V(x_t)-r(x_t)-\gamma\mathbb{E}[V(x_{t+1})|x_t]\big)^2\big], \tag{4}$$
where $\{x_1,\cdots,x_N\}\subset\mathcal{X}$ ($|\mathcal{X}|=N$) denotes a Markov chain with unknown transition operator, $r:\mathcal{X}\to\mathbb{R}$ is a reward function and a discount factor $\gamma\in(0,1)$. Here the goal of policy evaluation is to obtain function value $V=(V(x_1),\cdots,V(x_N)):\mathcal{X}\to\mathbb{R}^N$. This policy evaluation is widely applied in reinforcement learning.
- **Low-Rank Matrix Estimation** has the following form:
$$\min_{U,V} = f(\mathbb{E}(X)-UV^T), \tag{5}$$

Conference'17, July 2017, Washington, DC, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Convergence properties comparison of the nonconvex compositional stochastic algorithms and other ones. (C, NC, S, NS and mNS are the abbreviations of convex, non-convex, smooth, non-smooth and the sum of multiple non-smooth functions, respectively.) Note that $n_1 = n_2$ in [20].

Algorithm	Reference	Problem	Convergence rate	Query Complexity
com-SVR-ADMM	Yu and Huang [34]	C(S) + C(NS)	$O(\frac{\log(T)}{T})$	unkown
ASC-PG	Wang et al. [31]	NC(S)	$O(\frac{1}{T^{4/9}})$	$O(\epsilon^{-9/4})$
SC-SCSG	Liu et al. [20]	NC(S)	$O(\frac{1}{T})$	$O(\min\{\epsilon^{-9/5}, n_1^{4/5}\epsilon^{-1}\})$
VRSC-PG	Huo et al. [11]	NC(S) + C(NS)	$O(\frac{1}{T})$	$O((n_1+n_2)^{2/3}\epsilon^{-1})$
com-SVRG-ADMM	Ours	NC(S) + C(mNS)	$O(\frac{1}{T})$	$O(n_1^{2/3}\epsilon^{-1})$
com-SAGA-ADMM			$O(\frac{1}{T})$	$O(n_2^{2/3}\epsilon^{-1})$

where $f(\cdot)$ denotes the loss function, random matrix X sample from the unknown matrix $\bar{X}$ with $\mathbb{E}(X) = \bar{X}$ and the unknowns U and V are $n \times k$ ($k \ll n$) matrices. Low rank matrix approximation is widely applied in image analysis, topic models, recommendation systems, and graphical models.

Though stochastic gradient descent (SGD) and its variants [3]are well suited for solving the classic single-level stochastic optimization problem (*i.e.,*$G(x) \equiv x$ in (1) and (2)), they are not easily used for the above composition optimization problems due to composition functions in the objective. For solving composition optimization problem (2), Wang et al. [29] have proposed a class of stochastic compositional gradient (SCGD) methods by sampling the inner functions of composition functions. Though its scalability, SCGD still suffers from a slow convergence rate due to the existing of large variance in stochastic gradients. To reduce variance of stochastic gradients, Lian et al. [18], Liu et al. [20, 21] have proposed some fast SCGD methods by using the variance reduced techniques. Moreover, for solving stochastic composition problems with nonsmooth regularization, Wang et al. [30, 31] have proposed the stochastic compositional proximal gradient method. Moreover, Huo et al. [11], Lin et al. [19] have proposed fast stochastic compositional proximal gradient methods by using the variance reduced techniques. In particular, Yu and Huang [34] has proposed a fast stochastic ADMM-based method for nonsmooth stochastic composition problem, based on the stochastic ADMM method [35].

So far, the above stochastic composition methods mainly rely on the convexity of objective functions. In fact, the nonconvex stochastic composition problems are also successful in many applications such as tensor decomposition and deep learning [17]. Recently, some works [11, 20, 29, 31] have begin to studied the nonconvex stochastic composition optimization. However, these nonconvex stochastic composition methods still exist **two main drawbacks**: **1)** These nonconvex methods only deal with some simple nonsmooth penalties or even can not deal with the nonsmooth penalties. Clearly, these methods are limited in data mining applications, which often use nonsmooth penalties to include data knowledge. For example, in portfolio management, we usually access the relationships between assets. To improve investment benefits, we use the relatively complex graph-guided fused lasso [15] penalty to incorporate these relationships. Due to the existing of graph-guided fused lasso, the above compositional proximal gradient methods are not well suited for this problem. **2)** These nonconvex methods still have high query complexity resulting in not being still competent for large-scale problems.

In the paper, thus, we propose a class of fast compositional stochastic ADMM methods for the following *nonconvex nonsmooth* composition problem:

$$\min_{x,\{y\}_{j=1}^k} \frac{1}{n_1}\sum_{i=1}^{n_1} F_i\Big(\frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x)\Big) + \sum_{j=1}^{k} \psi_j(y_j) \qquad (6)$$
$$\text{s.t. } Ax + \sum_{j=1}^{k} B_j y_j = c,$$

where $f(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} F_i(\frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x)) : \mathbb{R}^d \to \mathbb{R}$ is a *nonconvex* and smooth function, and $\psi_j(y_j) : \mathbb{R}^q \to \mathbb{R}$ is a convex and possibly *nonsmooth* function for all $j \in [k]$, $k \geq 1$. The above problem (6) can use many complex regularization penalties such as graph-guided fused lasso and superposition structures penalties (e.g., sparse + low rank). Due to the flexibility in splitting the objective function into loss $f(x)$ and regularizer $\psi_j(y_j)$ for all $j \in [k]$, ADMM is an efficient method for solving the above constricted problem. Considering n_1 and n_2 are large in the problem (6), the classic ADMM and its variants are not suited for this problem, due to evaluating $G(x) = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x)$. Thus, we use the variance reduced sampling techniques such as SVRG [13] and SAGA [7] to estimate both values and gradients of the inner functions $\{G_j(x)\}_{j=1}^{n_2}$ and the gradients of the outer functions $\{F_i(G(x))\}_{i=1}^{n_1}$.

1.1 Challenges and Contributions

Although both the SVRG and SAGA have shown good performances in the classic SGD, applying these techniques to the nonconvex compositional stochastic ADMM methods *is not a trivial task*. There exist the following two main **challenges**:

- Both SVRG and SAGA rely on the assumption that the stochastic gradient is an *unbiased* estimate of the true full gradient, which does not hold in the compositional stochastic methods;
- Due to failure of the Féjer monotonicity of iteration, the convergence analysis of the nonconvex ADMM is generally quite difficult [27]. Obviously, with using the

inexact gradient, this difficulty is greater in the nonconvex compositional stochastic ADMM methods.

In the paper, thus, we will fill this gap between the nonconvex compositional stochastic ADMM and the variance reduction methods. In summary, our main **contributions** are given as follows:

1) We propose a class of fast compositional stochastic ADMM methods (*i.e.,* com-SVRG-ADMM and com-SAGA-ADMM) to solve the nonconvex nonsmooth composition problem (6), based on the sampling variance reduction techniques such as the SVRG and SAGA.
2) Moreover, we prove that both the com-SVRG-ADMM and com-SAGA-ADMM have $O(\frac{1}{T})$ convergence rate, and have the optimal query complexities of $O(n_1^{2/3}\epsilon^{-1})$ and $O(n_2^{2/3}\epsilon^{-1})$ for finding an ϵ-approximate solution, respectively. In particular, the optimal query complexity of our methods improves the existing best results in nonconvex nonsmooth composition problems by a factor of $O(\max(n_1^{2/3}, n_2^{2/3}))$.
3) Finally, we apply the proposed algorithms to relationship-guided portfolio management and policy evaluation, whose experimental results validate that our algorithms have faster convergence rate than the existing nonconvex composition stochastic algorithms.

2 RELATED WORKS

ADMM [4, 9] is a popular optimization tool for solving the composite and constrained problems in data mining. For example, it is able to efficiently solve some complicated structure problems in data mining such as the graph-guided fused lasso [15] and superposition structures penalties, which are too complicated for the other popular optimization methods such as proximal gradient methods [2]. Thus, ADMM has been widely studied in recent years [14, 32]. For large scale optimization, the online and stochastic versions of ADMM [22, 24, 28] have been presented. To accelerate these stochastic ADMMs, some faster stochastic ADMM methods [25, 35] have been proposed by using the variance reduced techniques.

In fact, ADMM is also highly successful in solving various nonconvex problems such as tensor decomposition [16] and learning neural networks [26]. Thus, the nonconvex stochastic ADMM methods [10, 36] are proposed based on the variance reduced techniques. More recently, Yu and Huang [34] have proposed a compositional stochastic ADMM (com-SVR-ADMM) method for stochastic composition problems. However, this com-SVR-ADMM method builds on the convexity of objective function. Clearly, this ADMM method is limited in many applications. In the paper, thus, we present a class of fast nonconvex compositional stochastic ADMM methods.

3 PRELIMINARIES

In the section, we begin with restating the standard ϵ-approximate stationary point of the problem (6), as in [12, 36].

Definition 1. *Given $\epsilon > 0$, the point $(x^*, y^*_{[k]}, \lambda^*)$ is said to be an ϵ-approximate stationary point of the problems* (6), *if it holds that*

$$\mathbb{E}\big[dist(0, \partial L(x^*, y^*_{[k]}, \lambda^*))^2\big] \le \epsilon, \tag{7}$$

where $L(x, y_{[k]}, \lambda) = f(x) + \sum_{j=1}^k \psi_j(y_j) - \langle \lambda, Ax + \sum_{j=1}^k B_j y_j - c\rangle$,

$$\partial L(x, y_{[k]}, \lambda) = \begin{bmatrix} \nabla_x L(x, y_{[k]}, \lambda) \\ \partial_{y_1} L(x, y_{[k]}, \lambda) \\ \cdots \\ \partial_{y_k} L(x, y_{[k]}, \lambda) \\ -Ax - \sum_{j=1}^k B_j y_j + c \end{bmatrix},$$

$dist(0, \partial L) = \min_{L' \in \partial L} \|0 - L'\|$.

Next, we make some mild assumptions regarding problem (6) as follows:

Assumption 1. ***(Lipschitz Gradients)*** *There exist constants L_F, L_G and L_f for $\nabla F_i(x)$, $\nabla G_j(x)$ and $\nabla f(x)$ satisfying that*

$$\|\nabla F_i(x) - \nabla F_i(y)\| \le L_F\|x - y\|, \ \forall x, y \in \mathbb{R}^p, \forall i \in [n_1]$$
$$\|\nabla G_j(x) - \nabla G_j(y)\| \le L_G\|x - y\|, \ \forall x, y \in \mathbb{R}^d, \forall j \in [n_2]$$
$$\|(\nabla G_j(x))^T \nabla F_i(G(x)) - (\nabla G_j(y))^T \nabla F_i(G(y))\| \le L_f\|x - y\|;$$

Following [18], the above last inequality satisfies

$$\|\nabla f(x) - \nabla f(y)\| \le L_f\|x - y\|, \ \forall x, y \in \mathbb{R}^d,$$
$$f(x) \le f(y) + \nabla f(y)^T(x - y) + \frac{L_f}{2}\|x - y\|^2.$$

Assumption 2. ***(Bounded Gradients)*** *Gradients $\nabla F_i(x)$ and $\nabla G_j(x)$ have the upper bounds B_F and B_G, respectively,* i.e.,

$$\|\nabla F_i(x)\| \le \Delta_F, \ \forall x \in \mathbb{R}^d, \forall i \in [n_1],$$
$$\|\nabla G_j(y)\| \le \Delta_G, \ \forall y \in \mathbb{R}^p, \forall j \in [n_2].$$

Assumption 3. *$f(x)$ and $\psi_j(y_j)$ for all $j \in [k]$ are all lower bounded, and denote $f^* = \inf_x f(x)$ and $\psi_j^* = \inf_{y_j} \psi_j(y_j)$ for $j \in [k]$.*

Assumption 4. *A is a full row rank matrix.*

Assumptions 1 and 2 have been commonly used in the convergence analysis of the composition stochastic algorithms [29, 34]. Assumption 3 has been used in the study of ADMMs [34, 35].

3.1 Notations

Let $y_{[k]} = \{y_1, \cdots, y_k\}$ and $y_{[j:k]} = \{y_j, \cdots, y_k\}$ for $j \in [k] = \{1, 2, \cdots, k\}$. Given a positive definite matrix Q, $\|x\|_Q^2 = x^T Q x$; $\sigma_{\max}(Q)$ and $\sigma_{\min}(Q)$ denote the largest and smallest eigenvalues of matrix Q, respectively; the conditional number $\kappa_Q = \frac{\sigma_{\max}(Q)}{\sigma_{\min}(Q)}$. $\sigma^A_{\max}$ and $\sigma^A_{\min}$ denote the largest and smallest eigenvalues of matrix AA^T, respectively. Given positive definite matrices $\{H_j\}_{j=1}^k$, let $\sigma^H_{\min} = \min_j \sigma_{\min}(H_j)$ and $\sigma^H_{\max} = \max_j \sigma_{\max}(H_j)$.

4 FAST COMPOSITIONAL STOCHASTIC ADMMS

In this section, we propose a class of fast compositional stochastic ADMM methods for solving the problem (6) based on the variance reduced techniques, called as compositional SVRG-ADMM and compositional SAGA-ADMM.

4.1 Nonconvex Compositional SVRG-ADMM

In the subsection, we propose a fast compositional SVRG-ADMM (com-SVRG-ADMM) for the nonconvex stochastic composition problem (6). We begin with defining the augmented Lagrangian function of the problem (6) as follows:

$$\begin{aligned}\mathcal{L}_\rho(x, y_{[k]}, \lambda) = f(x) + \sum_{j=1}^k \psi_j(y_j) - \langle \lambda, Ax + \sum_{j=1}^k B_j y_j - c\rangle \\ + \frac{\rho}{2}\|Ax + \sum_{j=1}^k B_j y_j - c\|^2, \qquad (8)\end{aligned}$$

where λ denotes the dual variable; $\rho > 0$ denotes the penalty parameter.

In the com-SVRG-ADMM, updating variables $\{y_j\}_{j=1}^k$ and λ is the same as the classic ADMM. Due to using the sampling technique to estimate gradient of composition function $f(x)$, we define an approximated function over x_k as follows:

$$\begin{aligned}&\hat{\mathcal{L}}_\rho(x, y_{[k]}^{s,t+1}, \lambda_t^s, \hat{g}_t^s) = f(x_t^s) + (\hat{g}_t^s)^T(x - x_t^s) \\ &+ \frac{1}{2\eta}\|x - x_t^s\|_Q^2 + \sum_{j=1}^k \psi_j(y_j^{s,t+1}) - (\lambda_t^s)^T(Ax + \sum_{j=1}^k B_j y_j^{s,t+1} - c) \\ &+ \frac{\rho}{2}\|Ax + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2, \qquad (9)\end{aligned}$$

where $\eta > 0$ is a step size and $\hat{g}_t^s$ is a stochastic gradient over x_t^s. To avoid computing inverse of matrix $\frac{Q}{\eta} + A^TA$, set $Q = rI - \rho\eta A^TA \succ I$ with $r > \rho\eta\sigma_{\max}^A + 1$ to linearize term $\frac{\rho}{2}\|Ax + \sum_{j=1}^k B_j y_j - c\|^2$. To use the proximal operator to update y_j:

$$y_j^{s,t+1} = \arg\min_{y_j \in \mathbb{R}^q} \frac{1}{2}\|y_j - y_j^{s,t}\|^2 + \psi_j(y_j), \ \forall j \in [k] \qquad (10)$$

set $H_j = \gamma_j I - \rho B_j^T B_j \succ I$ with $\gamma_j > \rho\sigma_{\max}(B_j^T B_j) + 1$ for all $j \in [k]$ to linearize term $\frac{\rho}{2}\|Ax + \sum_{j=1}^k B_j y_j - c\|^2$.

The com-SVRG-ADMM algorithm is described in Algorithm 1. Similar to the algorithmic framework of SVRG [13], the com-SVRG-ADMM also has two-layer loops. At the s-th outer loop, we remain a snapshot $\tilde{x}^s$ and compute its full gradient

$$\nabla f(\tilde{x}^s) = \frac{1}{n_2}\sum_{j=1}^{n_2}(\nabla G_j(\tilde{x}^s))^T \frac{1}{n_1}\sum_{i=1}^{n_1}\nabla F_i(G(\tilde{x}^s)), \qquad (11)$$

where $G(\tilde{x}^s) = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j(\tilde{x}^s)$. In fact, computing a full gradient of $f(x)$ needs $(n_1 + 2n_2)$ queries.

Algorithm 1 com-SVRG-ADMM for Nonconvex Stochastic Composition Optimization

```
Input: $m$, $T$, $S = [T/m]$, $\eta > 0$ and $\rho > 0$;
Initialize: $x_0^1$ and $\lambda_0^1$;
for $s = 1, 2, \cdots, S$ do
    $\tilde{x}^s = x_0^s$;
    $G(\tilde{x}^s) = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j(\tilde{x}^s)$;
    $\nabla G(\tilde{x}^s) = \sum_{j=1}^{n_2} \nabla G_j(\tilde{x}^s)$;
    $\nabla f(\tilde{x}^s) = (\nabla G(\tilde{x}^s))^T \frac{1}{n_1}\sum_{i=1}^{n_1} \nabla F_i(G^s)$;
    for $t = 0, 1, \cdots, m-1$ do
        Uniformly randomly pick a mini-batch $\mathcal{J}_1^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_1^t| = b_1$, then update $\hat{G}_t^s$ using (12);
        Uniformly randomly pick a mini-batch $\mathcal{J}_2^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_2^t| = b_2$, then update $\nabla\hat{G}_t^s$ using (13);
        Uniformly randomly pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n_1\}$, and $|\mathcal{I}_t| = b$, then update $\hat{g}_t^s$ using (14);
        ♠ $y_j^{s,t+1} = \arg\min_{y_j} \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_j, y_{[j+1:k]}^{s,t}, \lambda_t^s) + \frac{1}{2}\|y_j - y_j^{s,t}\|_{H_j}^2$, for all $j \in [k]$;
        ♠ $x_{t+1}^s = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[k]}^{s,t+1}, \lambda_t^s, \hat{g}_t^s\big)$;
        ♠ $\lambda_{t+1}^s = \lambda_t^s - \rho(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c)$;
    end for
    $x_0^{s+1} = x_m^s$, $y_j^{s+1,0} = y_j^{s,m}$ for $j \in [k]$, $\lambda_0^{s+1} = \lambda_m^s$;
end for
Output: Iterate $\{x, y_{[k]}, \lambda\}$ chosen uniformly random from $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$.
```

At the t-th inner loop, to reduce the number of queries, we first uniformly randomly pick a mini-batch $\mathcal{J}_1^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_1^t| = b_1$, then $\hat{G}_{t+1}^s$ is estimated as follows:

$$\hat{G}_t^s = \frac{1}{b_1}\sum_{j \in \mathcal{J}_1^t}(G_j(x_t^s) - G_j(\tilde{x}^s)) + G(\tilde{x}^s). \qquad (12)$$

Similarly, we uniformly randomly pick a mini-batch $\mathcal{J}_2^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_2^t| = b_2$, then $\nabla\hat{G}_{t+1}^s$ is estimated as follows:

$$\nabla\hat{G}_t^s = \frac{1}{b_2}\sum_{j \in \mathcal{J}_2^t}(\nabla G_j(x_t^s) - \nabla G_j(\tilde{x}^s)) + \nabla G(\tilde{x}^s). \qquad (13)$$

It is important to note that in Algorithm 1, $\mathcal{J}_1^t$ and $\mathcal{J}_2^t$ are independent.

Finally, based on the above estimated $\hat{G}_{t+1}^s$ and $\nabla\hat{G}_{t+1}^s$, we uniformly randomly pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n_1\}$ with $|\mathcal{I}_t| = b$, then update $\hat{g}_t^s$ as follows:

$$\begin{aligned}\hat{g}_t^s =& \frac{1}{b}\sum_{i \in I_t}\big((\nabla\hat{G}_t^s)^T \nabla F_i(\hat{G}_t^s) - \nabla G(\tilde{x}^s)^T \nabla F_i(G(\tilde{x}^s))\big) \\ &+ \nabla f(\tilde{x}^s). \qquad (14)\end{aligned}$$

4.2 Nonconvex Compositional SAGA-ADMM

In the subsection, we propose a fast compositional SAGA-ADMM (com-SAGA-ADMM) for the nonconvex stochastic composition problem (6).

Considering the fact that computing a full gradient of the composition function $f(x)$ needs $(n_1 + 2n_2)$ queries, we use memory space in exchange for computation. Specifically, we use some old queried values (e.g., both values and gradients of the inner functions and gradients of the outer functions) stead of computing the full gradient of the composition function $f(x)$.

Algorithm 2 com-SAGA-ADMM for Nonconvex Stochastic Composition Optimization

```
**Input:** $T$, $\eta > 0$ and $\rho > 0$;
**Initialize:** For $j \in [n_2]$, $z_j^0 = x_0$, $u_j^0 = G_j(z_j^0)$, $U_0 = \frac{1}{n_2}\sum_{j=1}^{n_2} u_j^0$, $\hat{z}_j^0 = x_0$, $v_j^0 = \nabla G_j(\hat{z}_j^0)$, and $V_0 = \frac{1}{n_2}\sum_{j=1}^{n_2} v_j^0$; For $i \in [n_1]$, $w_i^0 = (V_0)^T \nabla F_i(U_0)$, and $W_0 = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i^0$;
**for** $t = 0, 1, \cdots, T-1$ **do**
    Uniformly randomly pick a mini-batch $\mathcal{J}_1^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_1^t| = b_1$, then update $\hat{G}_t$ using (15);
    Uniformly randomly pick a mini-batch $\mathcal{J}_2^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_2^t| = b_2$, then update $\nabla\hat{G}_t$ using (16);
    Uniformly randomly pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n_1\}$ with $|\mathcal{I}_t| = b$ then update $\hat{g}_t$ using (17);
    ♠ $y_j^{t+1} = \arg\min_{y_j} \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_j, y_{[j+1:k]}^t, \lambda_t) + \frac{1}{2}\|y_j - y_j^t\|_{H_j}^2$, for all $j \in [k]$;
    ♠ $x_{t+1} = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[k]}^{t+1}, \lambda_t, \hat{g}_t\big)$;
    ♠ $\lambda_{t+1} = \lambda_t - \rho(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c)$;
    For $j \in \mathcal{J}_1^t$, $z_j^{t+1} = x_t$, $u_j^{t+1} = G_j(x_t)$, for $j \notin \mathcal{J}_1^t$, $z_j^{t+1} = z_j^t$, $u_j^{t+1} = u_j^t$, then update $U_{t+1}$ using $U_{t+1} = U_t - \frac{1}{n_2}\sum_{j\in\mathcal{J}_1^t}(u_j^t - u_j^{t+1})$;
    For $j \in \mathcal{J}_2^t$, $\hat{z}_j^{t+1} = x_t$ and $v_j^{t+1} = \nabla G_j(x_t)$, for $j \notin \mathcal{J}_2^t$, $\hat{z}_j^{t+1} = \hat{z}_j^t$ and $v_j^{t+1} = v_j^t$, then update $V_{t+1}$ using $V_{t+1} = V_t - \frac{1}{n_2}\sum_{j\in\mathcal{J}_2^t}(v_j^t - v_j^{t+1})$;
    For $i \in \mathcal{I}_t$, $w_i^{t+1} = (V_{t+1})^T \nabla F_i(U_{t+1})$, for $i \notin \mathcal{I}^t$, $w_i^{t+1} = w_i^t$; then update $W_{t+1}$ using $W_{t+1} = W_t - \frac{1}{n_1}\sum_{i\in\mathcal{I}^t}(w_i^t - w_i^{t+1})$;
**end for**
**Output:** Iterate $\{x, y_{[k]}, \lambda\}$ chosen uniformly random from $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$.
```

The com-SAGA-ADMM algorithm is described in Algorithm 2. Similar to the algorithmic framework of SAGA [7], we first store both values and gradients of the inner functions $\{G_j(x)\}_{j=1}^{n_2}$, and gradients of the outer functions $\{F_i(G(x))\}_{i=1}^{n_1}$. Thus, we need store $(n_1 + 2n_2)$ queried values. First, we uniformly randomly pick a mini-batch $\mathcal{J}_1^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_1^t| = b_1$, then $\hat{G}_{t+1}$ is estimated as follows:

$$\hat{G}_t = \frac{1}{b_1}\sum_{j\in\mathcal{J}_1^t}(G_j(x_t) - u_j^t) + U_t, \tag{15}$$

where $U_t = \frac{1}{n_2}\sum_{j=1}^{n_2} u_j^t$. Similarly, we uniformly randomly pick a mini-batch $\mathcal{J}_2^t$ (with replacement) from $\{1, 2, \cdots, n_2\}$ with $|\mathcal{J}_2^t| = b_2$, then $\nabla\hat{G}_{t+1}$ is estimated as follows:

$$\nabla\hat{G}_t = \frac{1}{b_2}\sum_{j\in\mathcal{J}_2^t}(\nabla G_j(x_t) - v_j^t) + V_t, \tag{16}$$

where $V_t = \frac{1}{n_2}\sum_{j=1}^{n_2} v_j^t$. Here $\mathcal{J}_1^t$ and $\mathcal{J}_2^t$ are independent.

Finally, we uniformly randomly pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n_1\}$ with $|\mathcal{I}_t| = b$, then $\hat{g}_t$ is estimated as follows:

$$\hat{g}_t = \frac{1}{b}\sum_{i\in I_t}\big((\nabla\hat{G}_t)^T \nabla F_i(\hat{G}_t) - w_i^t\big) + W_t, \tag{17}$$

where $w_i^t = (V_t)^T \nabla F_i(U_t)$ and $W_t = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i^t$.

In Algorithm 2, it is important to note that the auxiliary variables $\{z_j^t\}_{j=1}^{n_2}$ and $\{\hat{z}_j^t\}_{j=1}^{n_2}$ are only useful for in the following convergence analysis, while these variables are not necessary introduced in the algorithmic implementation.

5 THEORETICAL ANALYSIS

In this section, we study the convergence analysis of both com-SVRG-ADMM and com-SAGA-ADMM. Specifically, we give the convergence rates and the optimal IFO complexity of the proposed algorithms.

5.1 Convergence Analysis of com-SVRG-ADMM

In the subsection, we give convergence analysis of the com-SVRG-ADMM. First, we introduce some useful lemmas as follows:

LEMMA 1. *Suppose the sequence $\big\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\big\}_{s=1}^S$ is generated by Algorithm 1, the stochastic gradient $\hat{g}_t^s$ satisfies the following inequality*

$$\mathbb{E}\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 \leq (\frac{L_f^2}{b} + \frac{2\triangle_F^2 L_G^2}{b_2} + \frac{2\triangle_G^4 L_F^2}{b_1})\mathbb{E}\|x_t^s - \tilde{x}^s\|^2, \tag{18}$$

where $\nabla f(x_t^s) = (\nabla G(x_t^s))^T \frac{1}{n_1}\sum_{i=1}^{n_1} F_i(G(x_t^s))$ with $\nabla G(x_t^s) = \frac{1}{n_2}\sum_{j=1}^{n_2} \nabla G_j(x_t^s)$ and $G(x_t^s) = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x_t^s)$.

LEMMA 2. *Suppose the sequence $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ is generated from Algorithm 1, and define a* Lyapunov *function:*

$$\begin{aligned}\Phi_t^s =& \mathbb{E}\big[\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} + \frac{9L_f^2}{\sigma_{\min}^A \rho})\|x_t^s - x_{t-1}^s\|^2 \\ &+ (\frac{9L_f^2}{b\sigma_{\min}^A \rho} + \frac{18\triangle_F^2 L_G^2}{b_2 \sigma_{\min}^A \rho} + \frac{18\triangle_G^4 L_F^2}{b_1 \sigma_{\min}^A \rho})\|x_{t-1}^s - \tilde{x}^s\|^2 \\ &+ c_t\|x_t^s - \tilde{x}^s\|^2\big],\end{aligned} \tag{19}$$

where the positive sequence $\{c_t\}$ satisfies, for $s = 1, 2, \cdots, S$

$$c_t = \begin{cases} \frac{36L_f^2}{b\sigma_{\min}^A\rho} + \frac{36\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{36\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho} + \frac{L_f}{b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1} \\ \quad + (1+\beta)c_{t+1}, \ 1 \le t \le m, \\ 0, \ t \ge m+1. \end{cases}$$

Further let $b = m^2$, $b_1 \ge \frac{8\triangle_G^4 L_F^2 m^2}{L_f^2}$, $b_2 \ge \frac{8\triangle_F^2 L_G^2 m^2}{L_f^2}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{11L_f}$ $(0 < \alpha \le 1)$ and $\rho = \frac{2\sqrt{735}\kappa_Q L_f}{\sigma_{\min}^A \alpha}$, we have

$$\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m}(\|x_t^s - x_{t+1}^s\|^2 + \frac{1}{b}\|x_t^s - \tilde{x}^s\|^2 + \sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2) \\ \le \frac{\Phi_0^1 - \Phi^*}{\gamma T}, \tag{20}$$

where $\gamma = \min(\sigma_{\min}^H, \frac{18L_f^2}{\sigma_{\min}^A\rho} + \frac{L_f}{2}, \chi_t)$ with $\chi_t \ge \frac{\sqrt{735}\kappa_Q L_f}{2\sigma_{\min}^A\alpha^2}$, and Φ^* denotes a low bound of Φ_t^s.

Next, based on the above lemmas, we give the convergence analysis of com-SVRG-ADMM. For notational simplicity, let

$$\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big), \ \nu_2 = 5L_f^2 + \frac{3\sigma_{\max}^2(Q)}{\eta^2}$$

$$\nu_3 = \frac{15L_f^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho^2}.$$

THEOREM 1. *Suppose the sequence $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ is generated from Algorithm 1. Let $b = m^2$, $b_1 \ge \frac{8\triangle_G^4 L_F^2 m^2}{L_f^2}$, $b_2 \ge \frac{8\triangle_F^2 L_G^2 m^2}{L_f^2}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{11L_f}$ $(0 < \alpha \le 1)$ and $\rho = \frac{2\sqrt{735}\kappa_Q L_f}{\sigma_{\min}^A\alpha}$, then we have*

$$\min_{s,t}\mathbb{E}\big[dist(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \le \frac{2\nu_{\max}(\Phi_0^1 - \Phi^*)}{\gamma T} = O(\frac{1}{T}), \tag{21}$$

where $T = mS$, $\gamma = \min(\sigma_{\min}^H, \frac{18L_f^2}{\sigma_{\min}^A\rho} + \frac{L_f}{2}, \chi_t)$ with $\chi_t \ge \frac{\sqrt{735}\kappa_Q L_f}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ and Φ^* is a lower bound of function Φ_t^s. It implies that the whole number of iteration T satisfies

$$T = \frac{4\nu_{\max}(\Phi_0^1 - \Phi^*)}{\epsilon\gamma},$$

then $(x_{t^*}^{s^*}, y_{[k]}^{s^*,t^*}, \lambda_{t^*}^{s^*})$ is an ϵ-approximate solution of (6), where $(t^*, s^*) = \arg\min_{t,s}\theta_t^s$.

REMARK 1. *Theorem 1 shows that given $m = [n_1^{1/3}]$, $b = [n_1^{2/3}]$, $b_1 \ge \frac{8\triangle_G^4 L_F^2 m^2}{L_f^2}$ and $b_2 \ge \frac{8\triangle_F^2 L_G^2 m^2}{L_f^2}$, the com-SVRG-ADMM has convergence rate of $O(\frac{1}{T})$, and has the optimal query complexity of $O(n_1^{2/3}\epsilon^{-1})$ for finding an ϵ-approximate solution of the problem (6).*

5.2 Convergence Analysis of com-SAGA-ADMM

In the subsection, we give convergence analysis of the com-SAGA-ADMM. First, we introduce some useful lemmas as follows:

LEMMA 3. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2, the stochastic gradient $\hat{g}_t$ satisfies the following inequality:*

$$\mathbb{E}\|\hat{g}_t - \nabla f(x_t)\|^2 \le (\frac{2\triangle_F^2 L_G^2}{bn_2} + \frac{2\triangle_F^2 L_G^2}{b_2 n_2})\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 \\ + (\frac{2\triangle_G^4 L_F^2}{bn_2} + \frac{2\triangle_G^4 L_F^2}{b_1 n_2})\sum_{j=1}^{n_2}\mathbb{E}\|x_t - z_j^t\|^2, \tag{22}$$

where $\nabla f(x_t) = (\nabla G(x_t))^T \frac{1}{n_1}\sum_{i=1}^{n_1} F_i(G(x_t))$ with $\nabla G(x_t) = \frac{1}{n_2}\sum_{j=1}^{n_2}\nabla G_j(x_t)$ and $G(x_t) = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x_t)$.

LEMMA 4. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2, and define a generalized* Lyapunov *function*

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\rho\eta^2} + \frac{9L_f^2}{\sigma_{\min}^A\rho})\|x_t - x_{t-1}\|^2 \\ + (\frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b} + \frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b_2})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_{t-1} - \hat{z}_j^{t-1}\|^2 \\ + (\frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b} + \frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b_1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_{t-1} - z_j^{t-1}\|^2 \\ + \frac{c_t}{n_2}\sum_{i=1}^{n_2}\|x_t - z_j^t\|^2) + \frac{d_t}{n_2}\sum_{i=1}^{n_2}\|x_t - \hat{z}_j^t\|^2\big],$$

where $\{c_t\}$ and $\{d_t\}$ are positive sequences satisfy

$$c_t = \begin{cases} \frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b} + \frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b_1} + \frac{2\triangle_G^4 L_F^2}{L_f b} + \frac{2\triangle_G^4 L_F^2}{L_f b_1} \\ \quad + (1-p_1)(1+\beta_1)c_{t+1}, \ 0 \le t \le T-1, \\ 0, \ t \ge T. \end{cases}$$

$$d_t = \begin{cases} \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b} + \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b_2} + \frac{2\triangle_F^2 L_G^2}{L_f b} + \frac{2\triangle_F^2 L_G^2}{L_f b_2} \\ \quad + (1-p_2)(1+\beta_2)d_{t+1}, \ 0 \le t \le T-1, \\ 0, \ t \ge T. \end{cases}$$

where p_1 and p_2 denote the probability of an index j being in $\mathcal{J}_t^1$ and $\mathcal{J}_t^2$, respectively. Given $b_1 = b_2 = [n_2^{2/3}]$, $b \ge \max(b_1, b_2)$, $\hat{L}_f = \frac{96\triangle_G^4 L_F^2 + 96\triangle_F^2 L_G^2 + L_f^2}{L_f}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{\hat{L}_f}$ $(0 < \alpha \le 1)$ and $\rho = \frac{2\sqrt{15}\kappa_Q\hat{L}_f}{\sigma_{\min}^A\alpha}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}(\|x_t - x_{t+1}\|^2 + \sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2 + \frac{1}{b_1 n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2 \\ + \frac{1}{b_2 n_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2) \le \frac{\Omega_0 - \Omega^*}{\gamma T}, \tag{23}$$

where $\gamma = \min(\chi_t, \sigma^H_{\min}, \frac{\triangle^4_G L^2_F b_1}{L_f b} + \frac{\triangle^4_G L^2_F}{L_f}, \frac{\triangle^2_F L^2_G b_2}{L_f b} + \frac{\triangle^2_F L^2_G}{L_f})$ *with* $\chi_t \geq \frac{\sqrt{15}\kappa_Q \hat{L}_f}{2\alpha}$*. and* Ω^* *denotes a low bound of* Ω_t*.*

Theorem 2. *Suppose the sequence* $\{x_t, y^t_{[k]}, \lambda_t\}_{t=1}^T$ *is generated from Algorithm 2. Let* $b_1 = b_2 = [n_2^{2/3}]$, $b \geq \max(b_1, b_2)$, $\hat{L}_f = \frac{96\triangle^4_G L^2_F + 96\triangle^2_F L^2_G + L^2_f}{L_f}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{\hat{L}_f}$ $(0 < \alpha \leq 1)$ *and* $\rho = \frac{2\sqrt{15}\kappa_Q \hat{L}_f}{\sigma^A_{\min}\alpha}$*, we have*

$$\min_{1\leq t\leq T} \mathbb{E}\big[\mathit{dist}(0, \partial L(x_t, y^t_{[k]}, \lambda_t))^2\big] \leq \frac{2\nu_{\max}(\Omega_0 - \Omega^*)}{\gamma T} = O(\frac{1}{T}), \tag{24}$$

where $\gamma = \min(\chi_t, \sigma^H_{\min}, \frac{\triangle^4_G L^2_F b_1}{L_f b} + \frac{\triangle^4_G L^2_F}{L_f}, \frac{\triangle^2_F L^2_G b_2}{L_f b} + \frac{\triangle^2_F L^2_G}{L_f})$ *with* $\chi_t \geq \frac{\sqrt{15}\kappa_Q \hat{L}_f}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ *and* Ω^* *is a lower bound of function* Ω_t*. It implies that the iteration number* T *satisfies*

$$T = \frac{4\kappa_{\max}}{\epsilon\gamma}(\Omega_0 - \Omega^*),$$

then $(x_{t^*}, y^{t^*}_{[k]}, \lambda_{t^*})$ *is an* ϵ*-approximate solution of* (6)*, where* $t^* = \arg\min_{1\leq t\leq T} \theta_t$*.*

Remark 2. *Theorem 2 shows that given* $b_1 = b_2 = [n_2^{2/3}]$ *and* $b \geq \max(b_1, b_2)$*, the com-SAGA-ADMM has convergence rate of* $O(\frac{1}{T})$*, and has the optimal query complexity of* $O(n_2^{2/3}\epsilon^{-1})$ *for finding an* ϵ*-approximate solution of the problem* (6)*.*

6 EXPERIMENTS

In this section, we use two applications to evaluate our proposed methods: 1) relationship-guided portfolio management; 2) policy evaluation in reinforcement learning. In the experiments, we compare our methods (com-SVRG-ADMM and com-SAGA-ADMM) with two other nonconvex nonsmooth composition optimization methods: Accelerated Stochastic Compositional Proximal Gradient (ASC-PG [30]) and Variance Reduced Stochastic Compositional Proximal Gradient (VRSC-PG [11]).

6.1 Relationship-Guided Portfolio Management

In this subsection, we apply the proposed methods to portfolio management with incorporating to known relationships between assets. Given d assets and the reward vectors $\{r_i\}_{i=1}^n \subset \mathbb{R}^d$ observed at n time points, then the goal of portfolio management is to maximize return of the investment as well as to minimize risk of the investment. In fact, we usually can obtain the relationships between assets, and use these relationships to improve investment benefits. Thus, we propose a relationship-guided portfolio management optimization, which is equivalent to the following mean-variance optimization with graph-guided fused lasso:

$$\min_{x\in\mathbb{R}^d} \underbrace{\frac{1}{n}\sum_{i=1}^n \big(\langle x, r_i\rangle - \frac{1}{n}\sum_{i=1}^n \langle x, r_i\rangle\big)^2 - \frac{1}{n}\sum_{i=1}^n \langle x, r_i\rangle}_{f(x)} + \tau_1 \sum_j \kappa(|x_j|) + \tau_2\|\tilde{G}x\|_1, \tag{25}$$

where $\tau_1, \tau_2 \geq 0$ and $\tilde{G}$ denotes relationship matrix of the assets, *e.g.,* if $\tilde{G}_{ij} > 0$, i-th asset and j-th asset have a positive correlation investment. Here $\kappa(|x_j|) = \beta\log(1+\frac{|x_j|}{\alpha})$ is the nonconvex log-sum penalty function [5], which control the sparsity of combination coefficient x. In the experiment, $\tilde{G}$ is obtained by sparse inverse covariance matrix estimation [8]. To use our methods to solve the problem (25), we can change this problem into the following form:

$$\begin{aligned} &\min_{x,y_1,y_2} \bar{f}(x) + \tau_1\kappa_0\|y_1\|_1 + \tau_2\|y_2\|_1 \\ &\text{s.t.} \quad Ax + B_1 y_1 + B_2 y_2 = 0, \end{aligned} \tag{26}$$

where $\bar{f}(x) = f(x) + \tau_1(\sum_{j=1}^d \kappa(|x_j|) - \kappa_0\|x\|_1)$ with $\kappa_0 = \kappa'(0)$, $A = [I; \tilde{G}]$, $B_1 = [-I, 0]$ and $B_2 = [0, -I]$. Following [33], the function $\bar{f}(x)$ is nonconvex and nonsmooth.

In the experiment, we use 21 US Research Returns datasets from the Center for Research in Security Prices website[1], including three large 100-porfolio datasets and 18 medium datasets for Developed Market Factors and Returns, which are at detail described in Table 2. In the problem (25), we fix $\tau_1 = 10^{-4}$ and $\tau_2 = 10^{-5}$. In addition, all algorithm use the same initial solution $x_0 = [1, 1, \cdots, 1]^T$.

Figure 1 shows that in relationship-guided portfolio management, the objective values of our algorithms faster decrease than those of other algorithms, as CPU time consumed increases. These results demonstrate that our algorithms have relatively faster convergence rate than other algorithms. In particular, the objective values of com-SVRG-ADMM faster decrease than that of com-SAGA-ADMM, because com-SAGA-ADMM need much time to search old objective values and gradients from the corresponding tables.

6.2 Policy Evaluation in Reinforcement Learning

In this subsection, we apply the proposed methods to policy evaluation in reinforcement learning based on Bellman equations. Assume that there are N states and the policy of interest π. Let $V^\pi(s) \in \mathbb{R}$ denotes the value of state s under policy π, then the Bellman equation of this problem is given as follows: for all $s_1 \in \{1, 2, \cdots, N\}$

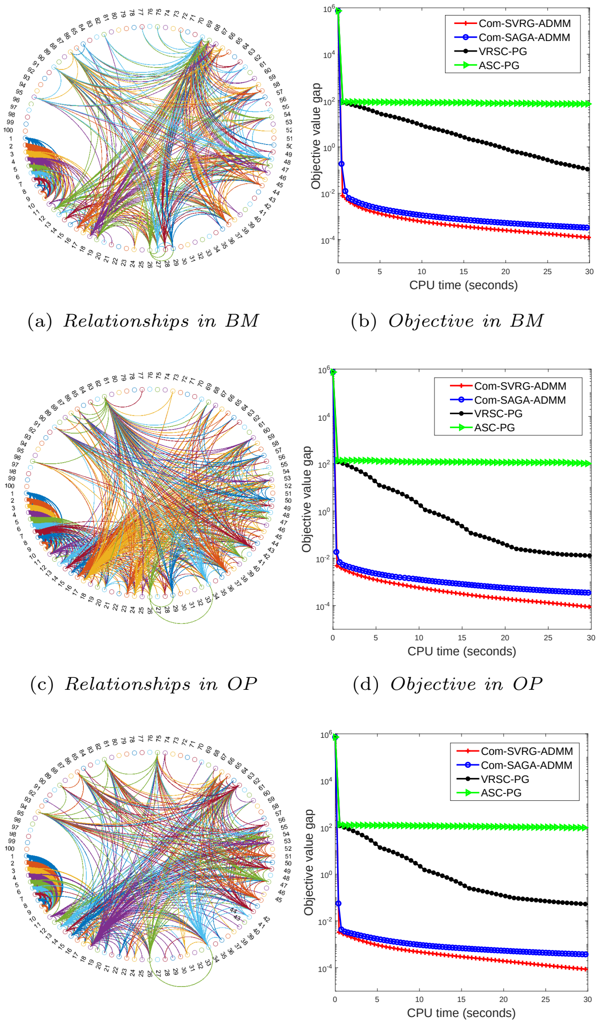
$$V^\pi(s_1) = \mathbb{E}_\pi\{r_{s_1,s_2} + \gamma V^\pi(s_2)\}, \ \forall s_2 \in \{1, 2, \cdots, N\}, \tag{27}$$

where r_{s_1,s_2} denotes the reward of changing from state s_1 to s_2, and $0 < \gamma < 1$ is a discount factor. The goal of policy evaluation is to obtain $V^\pi = (V^\pi(s_1), \cdots, V^\pi(s_N))$. Considering curse of dimensionality, we usually approximate the value of

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data Library/changes crsp.html

Table 2: The Statistics of 21 CRSP Real Datasets.

Type	#Assets	#Time points	Datasets
100-Portfolios	100	13781	Book-to-Market (BM), Operating Profitability (OP), Investment (INV)
Market Factors	25	7240	BM: Asia-Pacific-ex-Japan, Europe, Global-ex-US, Global, Japan, North-America
			OP: Asia-Pacific-ex-Japan, Europe, Global-ex-US, Global, Japan, North-America
			INV: Asia-Pacific-ex-Japan, Europe, Global-ex-US, Global, Japan, North-America



(a) *Relationships in BM* (b) *Objective in BM*

(c) *Relationships in OP* (d) *Objective in OP*

(e) *Relationships in INV* (f) *Objective in INV*

Figure 1: The relationships between 100 assets and the performance of all algorithms on three 100-Portfolio Datasets.

each state by some linear map of its feature $\phi_s \in \mathbb{R}^d$ $(d < N)$, then assume that $V^\pi(s) \approx \phi_s^T w$ for some $w \in \mathbb{R}^d$. Thus, we formulate this policy evaluation problem as a Bellman residual minimization problem:

$$\min_{w\in\mathbb{R}^d} \underbrace{\sum_{s=1}^{N}(\phi_s^T w - q_{\pi,s}(w))^2}_{f(w)} + \tau_1 \sum_j \kappa(|w_j|) + \tau_2\|\hat{G}w\|_1, \quad (28)$$

where $q_{\pi,s}(w) = \mathbb{E}_\pi[r_{s,s'} + \gamma\phi_{s'}^T w] = \sum_{s'} P^\pi_{s,s'}(r_{s,s'} + \gamma\phi_{s'}^T w)$ and $P^\pi_{s,s'}$ denotes the transition probabilities under a policy π. Here, we use a prior matrix $\hat{G}$ to encode the structure relationships of parameter w. Similarly, we can change the above problem into the following form:

$$\min_{w,y_1,y_2} \bar{f}(w) + \tau_1\kappa_0\|y_1\|_1 + \tau_2\|y_2\|_1$$
$$\text{s.t.} \quad Aw + B_1y_1 + B_2y_2 = 0, \quad (29)$$

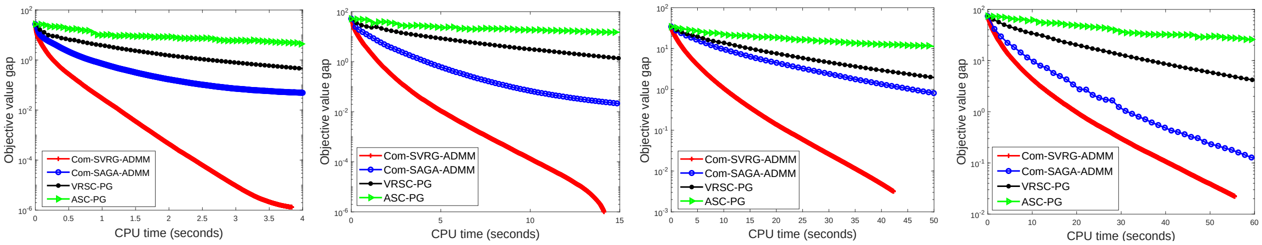
where $\bar{f}(w) = f(w) + \tau_1(\sum_{j=1}^d \kappa(|w_j|) - \kappa_0\|w\|_1)$ with $\kappa_0 = \kappa'(0)$, $A = [I; \tilde{G}]$, $B_1 = [-I, 0]$ and $B_2 = [0, -I]$.

Following [6], we generate some Markov decision processes, which includes 100, 400, 800 and 1000 states, respectively. Specifically, the transition probabilities are generated randomly from the uniform distribution in $[0, 1]$, and the rewards r_{s_1,s_2} from state s_1 to state s_2 are also sampled uniformly in $[0, 1]$. In the experiment, we fix a discount factor $\gamma = 10^{-5}$. In addition, we fix $\tau_1 = 10^{-4}$ and $\tau_2 = 10^{-5}$ in the problem (28).

Figure 2 shows that in policy evaluation, the the objective values of our algorithms faster decrease than those of other algorithms, as CPU time consumed increases. These results demonstrate that our algorithms have relatively faster convergence rate than other algorithms. In particular, the objective values of com-SVRG-ADMM faster decrease than that of com-SAGA-ADMM, because com-SAGA-ADMM need much time to search old objective values and gradients from the corresponding tables.

7 CONCLUSIONS

In this paper, we have proposed two fast composition stochastic ADMM methods (*i.e.*, com-SVRG-ADMM and com-SAGA-ADMM) for nonconvex stochastic composition optimization with multiple nonsmooth regularized penalties. Moreover, we have proved that both the com-SVRG-ADMM and com-SAGA-ADMM have convergence rate of $O(\frac{1}{T})$, and the com-SVRG-ADMM and com-SAGA-ADMM have the optimal query complexities of $O(n_1^{2/3}\epsilon^{-1})$ and $O(n_2^{2/3}\epsilon^{-1})$ for finding an ϵ-approximate solution, respectively, In particular, the optimal query complexity of our methods improves the existing best results in the nonconvex nonsmooth problems by a factor of $O(\max(n_1^{2/3}, n_2^{2/3}))$. Finally, we have applied the proposed methods to relationship-guided portfolio management and policy evaluation, whose experimental results validate that our methods have faster convergence rate than the existing nonconvex composition stochastic algorithms.



(a) $N = 100, d = 20$ (b) $N = 400, d = 50$ (c) $N = 800, d = 100$ (d) $N = 1000, d = 120$

Figure 2: Objective value gaps of policy evaluation in reinforcement learning.

REFERENCES

[1] Gordon J Alexander and Alexandre M Baptista. 2004. A comparison of VaR and CVaR constraints on portfolio selection with the mean-variance model. *Management science* 50, 9 (2004), 1261–1273.

[2] Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2, 1 (2009), 183–202.

[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. 2018. Optimization methods for large-scale machine learning. *SIAM Rev.* 60, 2 (2018), 223–311.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3, 1 (2011), 1–122.

[5] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. 2008. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications* 14, 5-6 (2008), 877–905.

[6] Christoph Dann, Gerhard Neumann, and Jan Peters. 2014. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research* 15, 1 (2014), 809–883.

[7] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. 2014. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*. 1646–1654.

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 3 (2008), 432–441.

[9] Daniel Gabay and Bertrand Mercier. 1976. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications* 2, 1 (1976), 17–40.

[10] Feihu Huang, Songcan Chen, and Zhaosong Lu. 2016. Stochastic Alternating Direction Method of Multipliers with Variance Reduction for Nonconvex Optimization. *arXiv preprint arXiv:1610.02758* (2016).

[11] Zhouyuan Huo, Bin Gu, Ji Liu, and Heng Huang. 2018. Accelerated method for stochastic composition optimization with nonsmooth regularization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[12] Bo Jiang, Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. 2016. Structured Nonconvex and Nonsmooth Optimization: Algorithms and Iteration Complexity Analysis. *arXiv preprint arXiv:1605.02408* (2016).

[13] Rie Johnson and Tong Zhang. 2013. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*. 315–323.

[14] Mojtaba Kadkhodaie, Konstantina Christakopoulou, Maziar Sanjabi, and Arindam Banerjee. 2015. Accelerated alternating direction method of multipliers. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 497–506.

[15] Seyoung Kim, Kyung-Ah Sohn, and Eric P Xing. 2009. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* 25, 12 (2009), i204–i212.

[16] Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM review* 51, 3 (2009), 455–500.

[17] Xiangru Lian and Ji Liu. 2018. Revisit Batch Normalization: New Understanding from an Optimization View and a Refinement via Composition Optimization. *arXiv preprint arXiv:1810.06177* (2018).

[18] Xiangru Lian, Mengdi Wang, and Ji Liu. 2016. Finite-sum composition optimization via variance reduced gradient descent. *arXiv preprint arXiv:1610.04674* (2016).

[19] Tianyi Lin, Chenyou Fan, Mengdi Wang, and Michael I Jordan. 2018. Improved Oracle Complexity for Stochastic Compositional Variance Reduced Gradient. *arXiv preprint arXiv:1806.00458* (2018).

[20] Liu Liu, Ji Liu, Cho-Jui Hsieh, and Dacheng Tao. 2018. Stochastically Controlled Stochastic Gradient for the Convex and Non-convex Composition problem. *arXiv preprint arXiv:1809.02505* (2018).

[21] Liu Liu, Ji Liu, and Dacheng Tao. 2018. Dualityfree Methods for Stochastic Composition Optimization. *IEEE transactions on neural networks and learning systems* 99 (2018), 1–13.

[22] Hua Ouyang, Niao He, Long Tran, and Alexander G Gray. 2013. Stochastic Alternating Direction Method of Multipliers. *ICML* 28 (2013), 80–88.

[23] Sashank Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. 2016. Proximal Stochastic Methods for Nonsmooth Non-convex Finite-Sum Optimization. In *Advances in Neural Information Processing Systems*. 1145–1153.

[24] Taiji Suzuki. 2013. Dual Averaging and Proximal Gradient Descent for Online Alternating Direction Multiplier Method.. In *ICML*. 392–400.

[25] Taiji Suzuki. 2014. Stochastic Dual Coordinate Ascent with Alternating Direction Method of Multipliers. In *ICML*. 736–744.

[26] Gavin Taylor, Ryan Burmeister, Zheng Xu, Bharat Singh, Ankit Patel, and Tom Goldstein. 2016. Training neural networks without gradients: a scalable ADMM approach. In *ICML*. 2722–2731.

[27] Fenghui Wang, Wenfei Cao, and Zongben Xu. 2015. Convergence of multi-block Bregman ADMM for nonconvex composite problems. *arXiv preprint arXiv:1505.03063* (2015).

[28] Huahua Wang and Arindam Banerjee. 2012. Online Alternating Direction Method. In *ICML*. 1119–1126.

[29] Mengdi Wang, Ethan X Fang, and Han Liu. 2017. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming* 161, 1-2 (2017), 419–449.

[30] Mengdi Wang, Ji Liu, and Ethan Fang. 2016. Accelerating stochastic composition optimization. In *Advances in Neural Information Processing Systems*. 1714–1722.

[31] Mengdi Wang, Ji Liu, and Ethan X Fang. 2017. Accelerating Stochastic Composition Optimization. *Journal of Machine Learning Research* 18 (2017), 1–23.

[32] Sen Yang, Jie Wang, Wei Fan, Xiatian Zhang, Peter Wonka, and Jieping Ye. 2013. An efficient ADMM algorithm for multidimensional anisotropic total variation regularization problems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 641–649.

[33] Quanming Yao and James Kwok. 2016. Efficient Learning with a Family of Nonconvex Regularizers by Redistributing Nonconvexity. In *ICML*. 2645–2654.

[34] Yue Yu and Longbo Huang. 2017. Fast stochastic variance reduced ADMM for stochastic composition optimization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 3364–3370.

[35] Shuai Zheng and James T Kwok. 2016. Fast and Light Stochastic ADMM. In *IJCAI*.

[36] Shuai Zheng and James T Kwok. 2016. Stochastic Variance-Reduced ADMM. *arXiv preprint arXiv:1604.07070* (2016).

A SUPPLEMENTARY MATERIALS

In this section, we will at detail give the proof of the above lemmas and theorems. First, to make the paper easier to follow, we give the following notations:

Notations:

- $[k] = \{1, 2, \cdots, k\}$ and $[j : k] = \{j, j+1, \cdots, k\}$ for all $1 \leq j \leq k$.
- $\|\cdot\|$ denotes the vector ℓ_2 norm and the matrix spectral norm, respectively.
- $\|x\|_Q = \sqrt{x^T Q x}$, where Q is a positive definite matrix.
- $\sigma_{\min}^A$ and $\sigma_{\max}^A$ denotes the minimum and maximum eigenvalues of $A^T A$, respectively.
- $\sigma_{\max}^{B_j}$ denotes the maximum eigenvalues of $B_j^T B_j$ for all $j \in [k]$, and $\sigma_{\max}^B = \max_{j=1}^k \sigma_{\max}^{B_j}$.
- $\sigma_{\min}(Q)$ and $\sigma_{\max}(Q)$ denote the minimum and maximum eigenvalues of matrix Q, respectively; the conditional number $\kappa_Q = \frac{\sigma_{\max}(Q)}{\sigma_{\min}(Q)}$.
- $\sigma_{\min}(H_j)$ and $\sigma_{\max}(H_j)$ denote the minimum and maximum eigenvalues of matrix H_j for all $j \in [k]$, respectively; $\sigma_{\min}(H) = \min_{j=1}^k \sigma_{\min}(H_j)$ and $\sigma_{\max}(H) = \max_{j=1}^k \sigma_{\max}(H_j)$.
- η denotes the step size of updating variable x.
- L_f denotes the Lipschitz constant of $\nabla f(x)$.
- b_1 and b_2 denote the mini-batch sizes for objective values and gradients of inner function, respectively; b denotes the mini-batch size for gradients of outer function.
- T, m and S are the total number of iterations, the number of iterations in the inner loop, and the number of iterations in the outer loop, respectively.

Next, we give an useful lemma in the following.

Lemma 5. *[23] For random variables $z_1, \cdots, z_r$ are independent and mean 0, we have*

$$\mathbb{E}[\|z_1 + \cdots + z_r\|^2] = \mathbb{E}[\|z_1\|^2 + \cdots + \|z_r\|^2]. \tag{30}$$

A.1 Theoretical Analysis of the com-SVRG-ADMM

In this subsection, we in detail give the convergence analysis of the com-SVRG-ADMM. First, we give some useful lemmas as follows:

Lemma 6. *Suppose the sequence $\big\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\big\}_{s=1}^S$ is generated by Algorithm 1, the stochastic gradient $\hat{g}_t^s$ satisfies the following inequality*

$$\mathbb{E}\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 \leq (\frac{L_f^2}{b} + \frac{2\triangle_F^2 L_G^2}{b_2} + \frac{2\triangle_G^4 L_F^2}{b_1})\mathbb{E}\|x_t^s - \tilde{x}^s\|^2, \tag{31}$$

where $\nabla f(x_t^s) = (\nabla G(x_t^s))^T \frac{1}{n_1}\sum_{i=1}^{n_1} F_i(G(x_t^s))$ with $\nabla G(x_t^s) = \frac{1}{n_2}\sum_{j=1}^{n_2} \nabla G_j(x_t^s)$ and $G(x_t^s) = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x_t^s)$.

Proof. We begin with defining h_t^s to be an unbiased estimate for $\nabla f(x_t^s)$ (*i.e.*, $\mathbb{E}[h_t^s] = \nabla f(x_t^s)$) as follows:

$$h_t^s = \frac{1}{b}\sum_{i\in\mathcal{I}_t}\big((\nabla G(x_t^s))^T \nabla F_i(G(x_t^s)) - (\nabla G(\tilde{x}^s))^T \nabla F_i(G(\tilde{x}^s))\big) + \nabla f(\tilde{x}^s), \tag{32}$$

where $\mathcal{I}_t$ is uniformly sampled from $\{1, 2, \cdots, n_1\}$ with $|\mathcal{I}_t| = b$. Then we have

$$\begin{aligned}
\mathbb{E}\|h_t^s - \nabla f(x_t^s)\|^2 &= \mathbb{E}\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}\big((\nabla G(x_t^s))^T \nabla F_i(G(x_t^s)) - (\nabla G(\tilde{x}^s))^T \nabla F_i(G(\tilde{x}^s))\big) + \nabla f(\tilde{x}^s) - \nabla f(x_t^s)\|^2 \\
&= \frac{1}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t^s))^T \nabla F_i(G(x_t^s)) - (\nabla G(\tilde{x}^s))^T \nabla F_i(G(\tilde{x}^s)) + \nabla f(\tilde{x}^s) - \nabla f(x_t^s)\|^2 \\
&\leq \frac{1}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t^s))^T \nabla F_i(G(x_t^s)) - (\nabla G(\tilde{x}^s))^T \nabla F_i(G(\tilde{x}^s))\|^2 \\
&\leq \frac{L_f^2}{b}\mathbb{E}\|x_t^s - \tilde{x}^s\|^2,
\end{aligned} \tag{33}$$

where the second equality follows Lemma 5.

Next, considering the upper bound of $\mathbb{E}\|\hat{g}_t^s - h_t^s\|^2$, we have

$$\begin{aligned}
\mathbb{E}\|\hat{g}_t^s - h_t^s\|^2 &= \mathbb{E}\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}\big((\nabla\hat{G}_t^s)^T\nabla F_i(\hat{G}_t^s) - (\nabla G(x_t^s))^T\nabla F_i(G(x_t^s))\big)\|^2 \\
&\leq \frac{1}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla\hat{G}_t^s)^T\nabla F_i(\hat{G}_t^s) - (\nabla G(x_t^s))^T\nabla F_i(G(x_t^s))\|^2 \\
&= \frac{1}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla\hat{G}_t^s)^T\nabla F_i(\hat{G}_t^s) - (\nabla G(x_t^s))^T\nabla F_i(\hat{G}_t^s) + (\nabla G(x_t^s))^T\nabla F_i(\hat{G}_t^s) - (\nabla G(x_t^s))^T\nabla F_i(G(x_t^s))\|^2 \\
&\leq \underbrace{\frac{2\triangle_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\nabla\hat{G}_t^s - \nabla G(x_t^s)\|^2}_{L_1} + \underbrace{\frac{2\triangle_G^2 L_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\hat{G}_t^s - G(x_t^s)\|^2}_{L_2},
\end{aligned} \tag{34}$$

where the second inequality holds by Assumptions 1 and 2.

Using Assumptions 1 and 2, we have

$$\begin{aligned}
L_1 &= \frac{2\triangle_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\nabla\hat{G}_t^s - \nabla G(x_t^s)\|^2 \\
&= \frac{2\triangle_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\frac{1}{b_2}\sum_{j\in\mathcal{J}_2^t}(\nabla G_{j_t}(x_t^s) - \nabla G_{j_t}(\tilde{x}^s)) + \nabla G(\tilde{x}^s) - \nabla G(x_t^s)\|^2 \\
&= \frac{2\triangle_F^2}{b_2^2}\sum_{j\in\mathcal{J}_2^t}\mathbb{E}\|\nabla G_{j_t}(x_t^s) - \nabla G_{j_t}(\tilde{x}^s) + \nabla G(\tilde{x}^s) - \nabla G(x_t^s)\|^2 \\
&\leq \frac{2\triangle_F^2}{b_2^2}\sum_{j\in\mathcal{J}_2^t}\mathbb{E}\|\nabla G_{j_t}(x_t^s) - \nabla G_{j_t}(\tilde{x}^s)\|^2 \\
&\leq \frac{2\triangle_F^2 L_G^2}{b_2}\mathbb{E}\|x_t^s - \tilde{x}^s\|^2,
\end{aligned} \tag{35}$$

where the second equality follows Lemma 5, and the first inequality holds by the equality $\mathbb{E}\|\xi - \mathbb{E}[\xi]\|^2 = \mathbb{E}\|\xi\|^2 - \|\mathbb{E}[\xi]\|^2$.

Similarly, we have

$$\begin{aligned}
L_2 &= \frac{2\triangle_G^2 L_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\hat{G}_t^s - G(x_t^s)\|^2 \\
&= \frac{2\triangle_G^2 L_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\frac{1}{b_1}\sum_{j\in\mathcal{J}_1^t}(G_{j_t}(x_t^s) - G_{j_t}(\tilde{x}^s)) + G(\tilde{x}^s) - G(x_t^s)\|^2 \\
&= \frac{2\triangle_G^2 L_F^2}{b_1^2}\sum_{j\in\mathcal{J}_1^t}\mathbb{E}\|G_{j_t}(x_t^s) - G_{j_t}(\tilde{x}^s) + G(\tilde{x}^s) - G(x_t^s)\|^2 \\
&\leq \frac{2\triangle_G^2 L_F^2}{b_1^2}\sum_{j\in\mathcal{J}_1^t}\mathbb{E}\|G_{j_t}(x_t^s) - G_{j_t}(\tilde{x}^s)\|^2 \\
&\leq \frac{2\triangle_G^4 L_F^2}{b_1}\mathbb{E}\|x_t^s - \tilde{x}^s\|^2.
\end{aligned} \tag{36}$$

Thus, we have

$$\mathbb{E}\|\hat{g}_t^s - h_t^s\|^2 \leq (\frac{2\triangle_F^2 L_G^2}{b_2} + \frac{2\triangle_G^4 L_F^2}{b_1})\mathbb{E}\|x_t^s - \tilde{x}^s\|^2. \tag{37}$$

Finally, combining the inequalities (33) with (37), we have

$$\mathbb{E}\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 \leq (\frac{L_f^2}{b} + \frac{2\triangle_F^2 L_G^2}{b_2} + \frac{2\triangle_G^4 L_F^2}{b_1})\mathbb{E}\|x_t^s - \tilde{x}^s\|^2. \tag{38}$$

□

Lemma 7. *Suppose the sequence $\big\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\big\}_{s=1}^S$ is generated by Algorithm 1, the following inequality holds*

$$\begin{aligned}\mathbb{E}\|\lambda_{t+1}^s - \lambda_t^s\|^2 \leq& (\frac{9L_f^2}{b\sigma_{\min}^A} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A})\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2}\|x_{t+1}^s - x_t^s\|^2 \\ &+ (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2} + \frac{9L_f^2}{\sigma_{\min}^A})\|x_t^s - x_{t-1}^s\|^2. \end{aligned} \tag{39}$$

Proof. Using the optimal condition for the step 13 of Algorithm 1, we have

$$\hat{g}_t^s + \frac{Q}{\eta}(x_{t+1}^s - x_t^s) - A^T\lambda_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) = 0, \tag{40}$$

By the step 14 of Algorithm 1, we have

$$A^T\lambda_{t+1}^s = \hat{g}_t^s + \frac{Q}{\eta}(x_{t+1}^s - x_t^s). \tag{41}$$

Since

$$A^T(\lambda_{t+1}^s - \lambda_t^s) = \hat{g}_t^s - \hat{g}_{t-1}^s + \frac{Q}{\eta}(x_{t+1}^s - x_t^s) - \frac{Q}{\eta}(x_t^s - x_{t-1}^s), \tag{42}$$

then we have

$$\|\lambda_{t+1}^s - \lambda_t^s\|^2 \leq \frac{1}{\sigma_{\min}^A}\big[3\|\hat{g}_t^s - \hat{g}_{t-1}^s\|^2 + \frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_{t+1}^s - x_t^s\|^2 + \frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_t^s - x_{t-1}^s\|^2\big]. \tag{43}$$

Next, considering the upper bound of $\|\hat{g}_t^s - \hat{g}_{t-1}^s\|^2$, we have

$$\begin{aligned}\|\hat{g}_t^s - \hat{g}_{t-1}^s\|^2 &= \|\hat{g}_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) - \nabla f(x_{t-1}^s) + \nabla f(x_{t-1}^s) - \hat{g}_{t-1}^s\|^2 \\ &\leq 3\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 + 3\|\nabla f(x_t^s) - \nabla f(x_{t-1}^s)\|^2 + 3\|\nabla f(x_{t-1}^s) - \hat{g}_{t-1}^s\|^2 \\ &\leq (\frac{3L_f^2}{b} + \frac{6\triangle_F^2 L_G^2}{b_2} + \frac{6\triangle_G^4 L_F^2}{b_1})(\mathbb{E}\|x_t^s - \tilde{x}^s\|^2 + \mathbb{E}\|x_{t-1}^s - \tilde{x}^s\|^2) + 3\|\nabla f(x_t^s) - \nabla f(x_{t-1}^s)\|^2 \\ &\leq (\frac{3L_f^2}{b} + \frac{6\triangle_F^2 L_G^2}{b_2} + \frac{6\triangle_G^4 L_F^2}{b_1})(\mathbb{E}\|x_t^s - \tilde{x}^s\|^2 + \mathbb{E}\|x_{t-1}^s - \tilde{x}^s\|^2) + 3L_f^2\|x_t^s - x_{t-1}^s\|^2, \end{aligned} \tag{44}$$

where the second inequality holds by the above Lemma 6 and the third inequality holds by Assumption 1. Finally, combining (43) and (44), we obtain the above result.

□

Lemma 8. *Suppose the sequence $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ is generated from Algorithm 1, and define a* Lyapunov *function:*

$$\begin{aligned}\Phi_t^s =& \mathbb{E}\big[\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L_f^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + (\frac{9L_f^2}{b\sigma_{\min}^A\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} \\ &+ \frac{18\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho})\|x_{t-1}^s - \tilde{x}^s\|^2 + c_t\|x_t^s - \tilde{x}^s\|^2\big], \end{aligned} \tag{45}$$

where the positive sequence $\{c_t\}$ satisfies, for $s = 1, 2, \cdots, S$

$$c_t = \begin{cases} \frac{36L_f^2}{b\sigma_{\min}^A\rho} + \frac{36\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{36\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho} + \frac{L_f}{b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1} + (1+\beta)c_{t+1},\ 1 \leq t \leq m, \\ 0,\ t \geq m+1. \end{cases}$$

Further let $b = m^2$, $b_1 \geq \frac{8\triangle_G^4 L_F^2 m^2}{L_f^2}$, $b_2 \geq \frac{8\triangle_F^2 L_G^2 m^2}{L_f^2}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{11L_f}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{735}\kappa_Q L_f}{\sigma_{\min}^A\alpha}$, we have

$$\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^m(\|x_t^s - x_{t+1}^s\|^2 + \frac{1}{b}\|x_t^s - \tilde{x}^s\|^2 + \sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2) \leq \frac{\Phi_0^1 - \Phi^*}{\gamma T}, \tag{46}$$

where $\gamma = \min(\sigma_{\min}^H, \frac{18L_f^2}{\sigma_{\min}^A\rho} + \frac{L_f}{2}, \chi_t)$ with $\chi_t \geq \frac{\sqrt{735}\kappa_Q L_f}{2\sigma_{\min}^A\alpha^2}$, and Φ^* denotes a low bound of Φ_t^s.

Proof. By the optimal condition of step 8 in Algorithm 1, we have, for $j \in [k]$

$$
\begin{aligned}
0 = & (y_j^{s,t} - y_j^{s,t+1})^T \big(\partial\psi_j(y_j^{s,t+1}) - B_j^T\lambda_t^s + \rho B_j^T(Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c) + H_j(y_j^{s,t+1} - y_j^{s,t})\big) \\
\leq & \psi_j(y_j^{s,t}) - \psi_j(y_j^{s,t+1}) - (\lambda_t^s)^T(B_j y_j^{s,t} - B_j y_j^{s,t+1}) + \rho(B_j y_j^{s,t} - B_j y_j^{s,t+1})^T(Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c) \\
& - \|y_j^{s,t+1} - y_j^{s,t}\|_{H_j}^2 \\
= & \psi_j(y_j^{s,t}) - \psi_j(y_j^{s,t+1}) - (\lambda_t^s)^T(Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^{k} B_i y_i^{s,t} - c) + (\lambda_t^s)^T(Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c) \\
& + \frac{\rho}{2}\|Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^{k} B_i y_i^{s,t} - c\|^2 - \frac{\rho}{2}\|Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c\|^2 - \|y_j^{s,t+1} - y_j^{s,t}\|_{H_j}^2 \\
& - \frac{\rho}{2}\|B_j y_j^{s,t} - B_j y_j^{s,t+1}\|^2 \\
= & \underbrace{f(x_t^s) + \sum_{i=1}^{j-1}\psi_i(y_i^{s,t+1}) + \sum_{i=j}^{k}\psi_i(y_i^{s,t}) - (\lambda_t^s)^T(Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^{k} B_i y_i^{s,t} - c) + \frac{\rho}{2}\|Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^{k} B_i y_i^{s,t} - c\|^2}_{\mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:k]}^{s,t}, \lambda_t^s)} \\
& - \underbrace{(f(x_t^s) + \sum_{i=1}^{j}\psi_i(y_i^{s,t+1}) + \sum_{i=j+1}^{k}\psi_i(y_i^{s,t}) - (\lambda_t^s)^T(Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c) + \frac{\rho}{2}\|Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c\|^2}_{\mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:k]}^{s,t}, \lambda_t^s)} \\
& - \|y_j^{s,t+1} - y_j^{s,t}\|_{H_j}^2 - \frac{\rho}{2}\|B_j y_j^{s,t} - B_j y_j^{s,t+1}\|^2 \\
\leq & \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:k]}^{s,t}, \lambda_t^s) - \mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:k]}^{s,t}, \lambda_t^s) - \sigma_{\min}(H_j)\|y_j^{s,t} - y_j^{s,t+1}\|^2,
\end{aligned}
\tag{47}
$$

where the first inequality holds by the convexity of function $\psi_j(y)$, and the second equality follows by applying the equality $(a-b)^T b = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ on the term $(By_j^{s,t} - By_j^{s,t+1})^T(Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c)$. Thus, we have, for all $j \in [k]$

$$
\mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:k]}^{s,t}, \lambda_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:k]}^{s,t}, \lambda_t^s) - \sigma_{\min}(H_j)\|y_j^{s,t} - y_j^{s,t+1}\|^2. \tag{48}
$$

Telescoping inequality (48) over j from 1 to k, we obtain

$$
\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) - \sigma_{\min}^H \sum_{j=1}^{k} \|y_j^{s,t} - y_j^{s,t+1}\|^2, \tag{49}
$$

where $\sigma_{\min}^H = \min_{j\in[k]} \sigma_{\min}(H_j)$.

By Assumption 1, we have

$$
0 \leq f(x_t^s) - f(x_{t+1}^s) + \nabla f(x_t^s)^T(x_{t+1}^s - x_t^s) + \frac{L_f}{2}\|x_{t+1}^s - x_t^s\|^2. \tag{50}
$$

Using optimal condition of the step 9 in Algorithm 1, we have

$$
0 = (x_t^s - x_{t+1}^s)^T\big(\hat{g}_t^s - A^T\lambda_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^{k} B_j y_j^{s,t+1} - c) + \frac{Q}{\eta}(x_{t+1}^s - x_t^s)\big). \tag{51}
$$

Combining (50) and (51), we have

$$\begin{aligned}
0 \leq\, & f(x_t^s) - f(x_{t+1}^s) + \nabla f(x_t^s)^T(x_{t+1}^s - x_t^s) + \frac{L_f}{2}\|x_{t+1}^s - x_t^s\|^2 \\
& + (x_t^s - x_{t+1}^s)^T\big(\hat{g}_t^s - A^T\lambda_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{Q}{\eta}(x_{t+1}^s - x_t^s)\big) \\
=\, & f(x_t^s) - f(x_{t+1}^s) + \frac{L_f}{2}\|x_t^s - x_{t+1}^s\|^2 - \frac{1}{\eta}\|x_t^s - x_{t+1}^s\|_Q^2 + (x_t^s - x_{t+1}^s)^T(\hat{g}_t^s - \nabla f(x_t^s)) \\
& - (\lambda_t^s)^T(Ax_t^s - Ax_{t+1}^s) + \rho(Ax_t^s - Ax_{t+1}^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) \\
\overset{(i)}{=}\, & f(x_t^s) - f(x_{t+1}^s) + \frac{L_f}{2}\|x_t^s - x_{t+1}^s\|^2 - \frac{1}{\eta}\|x_t^s - x_{t+1}^s\|_Q^2 + (x_t^s - x_{t+1}^s)^T(\hat{g}_t^s - \nabla f(x_t^s)) - (\lambda_t^s)^T(Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) \\
& + (\lambda_t^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\big(\|Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 - \|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 - \|Ax_t^s - Ax_{t+1}^s\|^2\big) \\
=\, & \underbrace{f(x_t^s) + \sum_{j=1}^k \psi_j(y_j^{s,t+1}) - (\lambda_t^s)^T(Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2}_{\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s)} \\
& - \underbrace{f(x_{t+1}^s) + \sum_{j=1}^k \psi_j(y_j^{s,t+1}) - (\lambda_t^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2}_{\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s)} \\
& + \frac{L_f}{2}\|x_t^s - x_{t+1}^s\|^2 + (x_t^s - x_{t+1}^s)^T(\hat{g}_t^s - \nabla f(x_t^s)) - \frac{1}{\eta}\|x_t^s - x_{t+1}^s\|_Q^2 - \frac{\rho}{2}\|Ax_t^s - Ax_{t+1}^s\|^2 \\
\leq\, & \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L_f}{2})\|x_t^s - x_{t+1}^s\|^2 + (x_t^s - x_{t+1}^s)^T(\hat{g}_t^s - \nabla f(x_t^s)) \\
\overset{(ii)}{\leq}\, & \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f)\|x_t^s - x_{t+1}^s\|^2 + \frac{1}{2L_f}\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 \\
\overset{(iii)}{\leq}\, & \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f)\|x_t^s - x_{t+1}^s\|^2 + (\frac{L_f}{2b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1})\mathbb{E}\|x_t^s - \tilde{x}^s\|^2,
\end{aligned} \tag{52}$$

where the equality (i) holds by applying the equality $(a-b)^T b = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ on the term $(Ax_t^s - Ax_{t+1}^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c)$, the inequality (ii) holds by the inequality $a^T b \leq \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|b\|^2$, and the inequality (iii) holds by Lemma 6. Thus, we obtain

$$\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f)\|x_t^s - x_{t+1}^s\|^2 + (\frac{L_f}{2b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1})\mathbb{E}\|x_t^s - \tilde{x}^s\|^2. \tag{53}$$

By the step 10 in Algorithm 1, we have

$$\begin{aligned}
\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) &= \frac{1}{\rho}\|\lambda_{t+1}^s - \lambda_t^s\|^2 \\
&\leq (\frac{9L_f^2}{b\sigma_{\min}^A\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho})\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) \\
&\quad + \frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2 + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L_f^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2.
\end{aligned} \tag{54}$$

Combining (49), (53) and (54), we have

$$\begin{aligned}\mathcal{L}_\rho(x^s_{t+1}, y^{s,t+1}_{[k]}, \lambda^s_{t+1}) \leq \mathcal{L}_\rho(x^s_t, y^{s,t}_{[k]}, \lambda^s_t) - \sigma^H_{\min}\sum_{j=1}^k \|y^{s,t}_j - y^{s,t+1}_j\|^2 - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L_f)\|x^s_t - x^s_{t+1}\|^2 \\ + (\frac{9L_f^2}{b\sigma^A_{\min}\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma^A_{\min}\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma^A_{\min}\rho})\big(\|x^s_t - \tilde{x}^s\|^2 + \|x^s_{t-1} - \tilde{x}^s\|^2\big) + \frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho}\|x^s_{t+1} - x^s_t\|^2 \\ + (\frac{L_f}{2b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1})\mathbb{E}\|x^s_t - \tilde{x}^s\|^2 + (\frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x^s_t - x^s_{t-1}\|^2. \qquad (55)\end{aligned}$$

Next, we define a *Lyapunov* function Φ^s_t as follows:

$$\Phi^s_t = \mathbb{E}\big[\mathcal{L}_\rho(x^s_t, y^{s,t}_{[k]}, \lambda^s_t) + (\frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L_f^2}{\sigma^A_{\min}\rho})\|x^s_t - x^s_{t-1}\|^2 + (\frac{9L_f^2}{b\sigma^A_{\min}\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma^A_{\min}\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma^A_{\min}\rho})\|x^s_{t-1} - \tilde{x}^s\|^2 + c_t\|x^s_t - \tilde{x}^s\|^2\big]. \qquad (56)$$

Considering the upper bound of $\|x^s_{t+1} - \tilde{x}^s\|^2$, we have

$$\begin{aligned}\|x^s_{t+1} - x^s_t + x^s_t - \tilde{x}^s\|^2 &= \|x^s_{t+1} - x^s_t\|^2 + 2(x^s_{t+1} - x^s_t)^T(x^s_t - \tilde{x}^s) + \|x^s_t - \tilde{x}^s\|^2 \\ &\leq \|x^s_{t+1} - x^s_t\|^2 + 2\big(\frac{1}{2\beta}\|x^s_{t+1} - x^s_t\|^2 + \frac{\beta}{2}\|x^s_t - \tilde{x}^s\|^2\big) + \|x^s_t - \tilde{x}^s\|^2 \\ &= (1 + 1/\beta)\|x^s_{t+1} - x^s_t\|^2 + (1 + \beta)\|x^s_t - \tilde{x}^s\|^2, \qquad (57)\end{aligned}$$

where the above inequality holds by the Cauchy-Schwarz inequality with $\beta > 0$. Combining (56) with (57), then we obtain

$$\begin{aligned}\Phi^s_{t+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x^s_{t+1}, y^{s,t+1}_{[k]}, \lambda^s_{t+1}) + (\frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L_f^2}{\sigma^A_{\min}\rho})\|x^s_{t+1} - x^s_t\|^2 + (\frac{9L_f^2}{b\sigma^A_{\min}\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma^A_{\min}\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma^A_{\min}\rho})\|x^s_t - \tilde{x}^s\|^2 + c_{t+1}\|x^s_{t+1} - \tilde{x}^s\|^2\big] \\ &\leq \mathcal{L}_\rho(x^s_t, y^{s,t}_{[k]}, \lambda^s_t) + (\frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L_f^2}{\sigma^A_{\min}\rho})\|x^s_t - x^s_{t-1}\|^2 + (\frac{9L_f^2}{b\sigma^A_{\min}\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma^A_{\min}\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma^A_{\min}\rho})\|x^s_{t-1} - \tilde{x}^s\|^2 \\ &\quad + \big(\frac{36L_f^2}{b\sigma^A_{\min}\rho} + \frac{36\triangle_F^2 L_G^2}{b_2\sigma^A_{\min}\rho} + \frac{36\triangle_G^4 L_F^2}{b_1\sigma^A_{\min}\rho} + \frac{L_f}{b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1} + (1 + \beta)c_{t+1}\big)\|x^s_t - \tilde{x}^s\|^2 - \sigma^H_{\min}\sum_{j=1}^k \|y^{s,t}_j - y^{s,t+1}_j\|^2 \\ &\quad - \big(\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L_f - \frac{6\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L_f^2}{\sigma^A_{\min}\rho} - (1 + 1/\beta)c_{t+1}\big)\|x^s_t - x^s_{t+1}\|^2 - (\frac{18L_f^2}{b\sigma^A_{\min}\rho} + \frac{L_f}{2b})\|x^s_t - \tilde{x}^s\|^2 \\ &\leq \Phi^s_t - \sigma^H_{\min}\sum_{j=1}^k \|y^{s,t}_j - y^{s,t+1}_j\|^2 - (\frac{18L_f^2}{b\sigma^A_{\min}\rho} + \frac{L_f}{2b})\|x^s_t - \tilde{x}^s\|^2 - \chi_t\|x^s_t - x^s_{t+1}\|^2, \qquad (58)\end{aligned}$$

where $c_t = \frac{36L_f^2}{b\sigma^A_{\min}\rho} + \frac{36\triangle_F^2 L_G^2}{b_2\sigma^A_{\min}\rho} + \frac{36\triangle_G^4 L_F^2}{b_1\sigma^A_{\min}\rho} + \frac{L_f}{b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1} + (1 + \beta)c_{t+1}$ and $\chi_t = \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L_f - \frac{6\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - (1 + 1/\beta)c_{t+1}$.

Next, we will prove the relationship between Φ^{s+1}_1 and Φ^s_m. Due to $x^{s+1}_0 = x^s_m = \tilde{x}^{s+1}$, we have $\hat{G}^{s+1}_0 = \frac{1}{b_1}\sum_{j\in\mathcal{J}^t_1}(G_{j_t}(x^{s+1}_0) - G_{j_t}(\tilde{x}^{s+1})) + G(\tilde{x}^{s+1}) = G(x^{s+1}_0)$. Similarly, we have $\nabla\hat{G}^{s+1}_0 = \nabla G(x^{s+1}_0)$. Then we obtain

$$\hat{g}^{s+1}_0 = \frac{1}{b}\sum_{i_t\in\mathcal{I}_t}\big((\nabla\hat{G}^{s+1}_0)^T\nabla F_{i_t}(\hat{G}^{s+1}_0) - (\nabla G(\tilde{x}^{s+1}))^T\nabla F_{i_t}(G(\tilde{x}^{s+1}))\big) + \nabla f(\tilde{x}^{s+1}) = \nabla f(x^{s+1}_0) = \nabla f(x^s_m). \qquad (59)$$

It follows that

$$\mathbb{E}\|\hat{g}^{s+1}_0 - \hat{g}^s_m\|^2 = \mathbb{E}\|\nabla f(x^s_m) - \hat{g}^s_m\|^2 \leq (\frac{L_f^2}{b} + \frac{2\triangle_F^2 L_G^2}{b_2} + \frac{2\triangle_G^4 L_F^2}{b_1})\mathbb{E}\|x^s_m - \tilde{x}^s\|^2. \qquad (60)$$

Thus, we have

$$\begin{aligned}\mathbb{E}\|\lambda_1^{s+1}-\lambda_m^s\|^2 &\le \frac{1}{\sigma_{\min}^A}\|\hat{g}_0^{s+1}-\hat{g}_m^s+\frac{Q}{\eta}(x_1^{s+1}-x_0^{s+1})+\frac{Q}{\eta}(x_m^s-x_{m-1}^s)\|^2\\ &= \frac{1}{\sigma_{\min}^A}\|\nabla f(x_m^s)-\hat{g}_m^s+\frac{Q}{\eta}(x_1^{s+1}-x_m^s)+\frac{Q}{\eta}(x_m^s-x_{m-1}^s)\|^2\\ &\le \frac{1}{\sigma_{\min}^A}\big(3\|\nabla f(x_m^s)-\hat{g}_m^s\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_1^{s+1}-x_m^s\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_m^s-x_{m-1}^s\|^2\big)\\ &\le \frac{1}{\sigma_{\min}^A}\big((\frac{L_f^2}{b}+\frac{2\triangle_F^2 L_G}{b_2}+\frac{2\triangle_G^4 L_F^2}{b_1})\|x_m^s-\tilde{x}^s\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_1^{s+1}-x_m^s\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_m^s-x_{m-1}^s\|^2\big).\end{aligned} \tag{61}$$

Since $x_m^s = x_0^{s+1}$, $y_j^{s,m} = y_j^{s+1,0}$ for all $j \in [k]$ and $\lambda_m^s = \lambda_0^{s+1}$, by (49), we have

$$\begin{aligned}\mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) &\le \mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,0}, \lambda_0^{s+1}) - \sigma_{\min}^H\sum_{j=1}^k\|y_j^{s+1,0}-y_j^{s+1,1}\|^2\\ &= \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) - \sigma_{\min}^H\sum_{j=1}^k\|y_j^{s,m}-y_j^{s+1,1}\|^2.\end{aligned} \tag{62}$$

By (53), we have

$$\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) \le \mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) - (\frac{\sigma_{\min}(Q)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L_f)\|x_0^{s+1}-x_1^{s+1}\|^2. \tag{63}$$

By (54), we have

$$\begin{aligned}\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) &\le \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) + \frac{1}{\rho}\|\lambda_1^{s+1}-\lambda_0^{s+1}\|^2\\ &\le \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) + \frac{1}{\sigma_{\min}^A\rho}\big((\frac{L_f^2}{b}+\frac{2\triangle_F^2 L_G^2}{b_2}+\frac{2\triangle_G^4 L_F^2}{b_1})\|x_m^s-\tilde{x}^s\|^2\\ &\quad + \frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_1^{s+1}-x_m^s\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_m^s-x_{m-1}^s\|^2\big).\end{aligned} \tag{64}$$

where the second inequality holds by (61).

Combining (62), (63) with (64), we have

$$\begin{aligned}\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) &\le \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) - \sigma_{\min}^H\sum_{j=1}^k\|y_j^{s,m}-y_j^{s+1,1}\|^2 - (\frac{\sigma_{\min}(Q)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L_f)\|x_0^{s+1}-x_1^{s+1}\|^2\\ &\quad + \frac{1}{\sigma_{\min}^A\rho}\big((\frac{L_f^2}{b}+\frac{2\triangle_F^2 L_G^2}{b_2}+\frac{2\triangle_G^4 L_F^2}{b_1})\|x_m^s-\tilde{x}^s\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_1^{s+1}-x_m^s\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_m^s-x_{m-1}^s\|^2\big).\end{aligned} \tag{65}$$

Therefore, we have

$$\begin{aligned}
\Phi_1^{s+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L_f^2}{\sigma_{\min}^A\rho})\|x_1^{s+1} - x_0^{s+1}\|^2 + (\frac{9L_f^2}{b\sigma_{\min}^A\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho})\|x_0^{s+1} - \tilde{x}^{s+1}\|^2 \\
&\quad + c_1\|x_1^{s+1} - \tilde{x}^{s+1}\|^2\big] \\
&= \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) + \big(\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L_f^2}{\sigma_{\min}^A\rho} + c_1\big)\|x_1^{s+1} - x_0^{s+1}\|^2 \\
&\le \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L_f^2}{\sigma_{\min}^A\rho})\|x_m^s - x_{m-1}^s\|^2 + (\frac{9L_f^2}{b\sigma_{\min}^A\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho})\|x_{m-1}^s - \tilde{x}^s\|_2^2 \\
&\quad + (\frac{36L_f^2}{b\sigma_{\min}^A\rho} + \frac{36\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{36\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho} + \frac{L_f}{b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1})\|x_m^s - \tilde{x}^s\|_2^2 \\
&\quad - \sigma_{\min}^H \sum_{j=1}^k \|y_j^{s,m} - y_j^{s+1,1}\|^2 - \big(\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - c_1\big)\|x_1^{s+1} - x_m^s\|_2^2 \\
&\quad - \frac{9L_f^2}{\sigma_{\min}^A\rho}\|x_m^s - x_{m-1}^s\|_2^2 - (\frac{9L_f^2}{b\sigma_{\min}^A\rho} + \frac{18\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{18\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho})\|x_{m-1}^s - \tilde{x}^s\|_2^2 - (\frac{18L_f^2}{b\sigma_{\min}^A\rho} + \frac{L_f}{2b})\|x_m^s - \tilde{x}^s\|_2^2 \\
&\le \Phi_m^s - \sigma_{\min}^H \sum_{j=1}^k \|y_j^{s,m} - y_j^{s+1,1}\|^2 - (\frac{18L_f^2}{b\sigma_{\min}^A\rho} + \frac{L_f}{2b})\|x_m^s - \tilde{x}^s\|_2^2 - \chi_m\|x_1^{s+1} - x_m^s\|^2, \qquad (66)
\end{aligned}$$

where $c_m = \frac{36L_f^2}{b\sigma_{\min}^A\rho} + \frac{36\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{36\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho} + \frac{L_f}{b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1}$, and $\chi_m = \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L_f^2}{\sigma_{\min}^A\rho} - c_1$.

Let $c_{m+1} = 0$ and $\beta = \frac{1}{m}$, recursing on t, we have

$$\begin{aligned}
c_{t+1} &= (\frac{36L_f^2}{b\sigma_{\min}^A\rho} + \frac{36\triangle_F^2 L_G^2}{b_2\sigma_{\min}^A\rho} + \frac{36\triangle_G^4 L_F^2}{b_1\sigma_{\min}^A\rho} + \frac{L_f}{b} + \frac{\triangle_F^2 L_G^2}{L_f b_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1})\frac{(1+\beta)^{m-t} - 1}{\beta} \\
&\le \frac{2m}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1})\big((1+\frac{1}{m})^{m-t} - 1\big) \\
&\le \frac{2m}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1})(e - 1) \\
&\le \frac{4m}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}), \qquad (67)
\end{aligned}$$

where the first inequality holds by $\rho \ge \frac{36L_f}{\sigma_{\min}^A}$ and the second inequality holds by $(1+\frac{1}{m})^m$ is an increasing function and $\lim_{m\to\infty}(1+\frac{1}{m})^m = e$. It follows that, for $t = 1, 2, \cdots, m$

$$\begin{aligned}
\chi_t &\ge \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1 + 1/\beta)\frac{4m}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}) \\
&= \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1 + m)\frac{4m}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}) \\
&\ge \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{8m^2}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}) \\
&= \underbrace{\frac{\sigma_{\min}(Q)}{\eta} - L_f - \frac{8m^2}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1})}_{M_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L_f^2}{\sigma_{\min}^A\rho}}_{M_2}. \qquad (68)
\end{aligned}$$

Given $b = m^2$, $b_1 \ge \frac{8\triangle_G^4 L_F^2 m^2}{L_f^2}$ and $b_2 \ge \frac{8\triangle_F^2 L_G^2 m^2}{L_f^2}$, we have

$$\begin{aligned}
M_1 &= \frac{\sigma_{\min}(Q)}{\eta} - L_f - \frac{8m^2}{b}(L_f + \frac{\triangle_F^2 L_G^2 b}{L_f b_2} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}) \\
&\ge \frac{\sigma_{\min}(Q)}{\eta} - L_f - 10L_f \ge 0, \qquad (69)
\end{aligned}$$

where the second inequality holds by $0 < \eta \leq \frac{\sigma_{\min}(Q)}{11L_f}$. Further, given $\eta = \frac{\alpha\sigma_{\min}(Q)}{11L_f}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{735}\kappa_Q L_f}{\sigma^A_{\min}\alpha^2} > \frac{36L_f}{\sigma^A_{\min}}$, we have

$$\begin{aligned} M_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L_f^2}{\sigma^A_{\min}\rho} \\ &= \frac{\rho\sigma^A_{\min}}{2} - \frac{726\kappa_Q^2 L_f^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9L_f^2}{\sigma^A_{\min}\rho} \\ &\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{726\kappa_Q^2 L_f^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9\kappa_Q^2 L_f^2}{\sigma^A_{\min}\rho\alpha^2} \\ &= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{735\kappa_Q^2 L_f^2}{\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\ &\geq \frac{\sqrt{735}\kappa_Q L_f}{2\sigma^A_{\min}\alpha^2} > 9L_f. \end{aligned} \tag{70}$$

Then we have $\chi_t \geq \frac{\sqrt{735}\kappa_Q L_f}{2\sigma^A_{\min}\alpha^2} > 9L_f$.

Thus, by (58) and (66), we obtain that the sequence $\{(\Phi^s_t)^m_{t=1}\}^S_{s=1}$ is decreasing. Next, by the definition of Φ^s_t, we have

$$\begin{aligned} \Phi^s_t &\geq \mathbb{E}\big[\mathcal{L}_\rho(x^s_t, y^{s,t}_{[k]}, \lambda^s_t)\big] \\ &= f(x^s_t) + \sum_{j=1}^k \psi_j(y^t_j) - (\lambda^s_t)^T(Ax^s_t + \sum_{j=1}^m B_j y^{s,t}_j - c) + \frac{\rho}{2}\|Ax^s_t + \sum_{j=1}^k B_j y^{s,t}_j - c\| \\ &= f(x^s_t) + \sum_{j=1}^k \psi_j(y^t_j) - \frac{1}{\rho}(\lambda^s_t)^T(z_{t-1} - z_t) + \frac{1}{2\rho}\|\lambda^s_t - \lambda^s_{t-1}\|^2 \\ &= f(x^s_t) + \sum_{j=1}^k \psi_j(y^t_j) - \frac{1}{2\rho}\|\lambda^s_{t-1}\|^2 + \frac{1}{2\rho}\|\lambda^s_t\|^2 + \frac{1}{\rho}\|\lambda^s_t - \lambda^s_{t-1}\|^2 \\ &\geq f^* + \sum_{j=1}^k \psi^*_j - \frac{1}{2\rho}\|\lambda^s_{t-1}\|^2 + \frac{1}{2\rho}\|\lambda^s_t\|^2. \end{aligned} \tag{71}$$

Summing the inequality (71) over $t = 0, 1\cdots, m$ and $s = 1, 2, \cdots, S$, we have

$$\frac{1}{T}\sum_{t=0}^m\sum_{s=1}^S \Phi^s_t \geq f^* + \sum_{j=1}^k \psi^*_j - \frac{1}{2\rho}\|\lambda^1_0\|^2, \tag{72}$$

where $T = mS$. It follows that the function Φ^s_t has a lower bound. Let Φ^* denotes a low bound of Φ^s_t.

Finally, telescoping the inequalities (58) and (66) over t from 0 to m and s from 1 to S, we have

$$\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^m (\|x^s_t - x^s_{t+1}\|^2 + \frac{1}{b}\|x^s_t - \tilde{x}^s\|^2 + \sum_{j=1}^k \|y^{s,t}_j - y^{s,t+1}_j\|^2) \leq \frac{\Phi^1_0 - \Phi^*}{\gamma T}, \tag{73}$$

where $\gamma = \min(\sigma^H_{\min}, \frac{18L_f^2}{\sigma^A_{\min}\rho} + \frac{L_f}{2}, \chi_t)$ with $\chi_t \geq \frac{\sqrt{735}\kappa_Q L_f}{2\sigma^A_{\min}\alpha^2}$.

□

Next, based on the above lemmas, we give the convergence analysis of com-SVRG-ADMM. For notational simplicity, let

$$\begin{aligned} &\nu_1 = k\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(H)\big), \ \nu_2 = 5L_f^2 + \frac{3\sigma^2_{\max}(Q)}{\eta^2} \\ &\nu_3 = \frac{15L_f^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho^2}. \end{aligned}$$

Theorem 3. *Suppose the sequence* $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ *is generated from Algorithm 1. Let* $b = m^2$, $b_1 \geq \frac{8\triangle_G^4 L_F^2 m^2}{L_f^2}$, $b_2 \geq \frac{8\triangle_F^2 L_G^2 m^2}{L_f^2}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{11L_f}$ $(0 < \alpha \leq 1)$ *and* $\rho = \frac{2\sqrt{735}\kappa_Q L_f}{\sigma_{\min}^A \alpha}$, *then we have*

$$\min_{s,t} \mathbb{E}\big[dist(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \leq \frac{2\nu_{\max}(\Phi_0^1 - \Phi^*)}{\gamma T} = O(\frac{1}{T}), \tag{74}$$

where $T = mS$, $\gamma = \min(\sigma_{\min}^H, \frac{18L_f^2}{\sigma_{\min}^A \rho} + \frac{L_f}{2}, \chi_t)$ *with* $\chi_t \geq \frac{\sqrt{735}\kappa_Q L_f}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ *and* Φ^* *is a lower bound of function* Φ_t^s. *It implies that the whole number of iteration* T *satisfies*

$$T = \frac{4\nu_{\max}(\Phi_0^1 - \Phi^*)}{\epsilon\gamma}$$

then $(x_{t^*}^{s^*}, y_{[k]}^{s^*,t^*}, \lambda_{t^*}^{s^*})$ *is an* ϵ*-approximate solution of* (6), *where* $(t^*, s^*) = \arg\min_{t,s} \theta_t^s$.

Proof. First, we define a variable $\theta_t^s = \|x_{t+1}^s - x_t^s\|^2 + \|x_t^s - x_{t-1}^s\|^2 + \frac{1}{b}(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2) + \sum_{j=1}^k \|y_j^{s,t} - y_j^{s,t+1}\|^2$. By the step 12 of Algorithm 1, we have, for all $j \in [k]$

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \partial_{y_j} L(x, y_{[k]}, \lambda))^2\big]_{s,t+1} &= \mathbb{E}\big[\text{dist}(0, \partial g_j(y_j^{s,t+1}) - B_j^T \lambda_{t+1}^s)^2\big] \\
&= \|B_j^T \lambda_t^s - \rho B_j^T (A x_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{k} B_i y_i^{s,t} - c) - H_j(y_j^{s,t+1} - y_j^{s,t}) - B_j^T \lambda_{t+1}^s\|^2 \\
&= \|\rho B_j^T A(x_{t+1}^s - x_t^s) + \rho B_j^T \sum_{i=j+1}^{k} B_i(y_i^{s,t+1} - y_i^{s,t}) - H_j(y_j^{s,t+1} - y_j^{s,t})\|^2 \\
&\leq k\rho^2 \sigma_{\max}^{B_j} \sigma_{\max}^A \|x_{t+1}^s - x_t^s\|^2 + k\rho^2 \sigma_{\max}^{B_j} \sum_{i=j+1}^{k} \sigma_{\max}^{B_i} \|y_i^{s,t+1} - y_i^{s,t}\|^2 \\
&\quad + k\sigma_{\max}^2(H_j)\|y_j^{s,t+1} - y_j^{s,t}\|^2 \\
&\leq k\big(\rho^2 \sigma_{\max}^B \sigma_{\max}^A + \rho^2 (\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big)\theta_t^s,
\end{aligned} \tag{75}$$

where the first inequality follows by the inequality $\|\frac{1}{n}\sum_{i=1}^n \alpha_i\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|\alpha_i\|^2$.

By the step 13 of Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}[\text{dist}(0, \nabla_x L(x, y_{[k]}, \lambda))]_{s,t+1} &= \mathbb{E}\|A^T \lambda_{t+1}^s - \nabla f(x_{t+1}^s)\|^2 \\
&= \mathbb{E}\|\hat{g}_t^s - \nabla f(x_{t+1}^s) - \frac{Q}{\eta}(x_t^s - x_{t+1}^s)\|^2 \\
&= \mathbb{E}\|\hat{g}_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) - \nabla f(x_{t+1}^s) - \frac{Q}{\eta}(x_t^s - x_{t+1}^s)\|^2 \\
&\leq 3\mathbb{E}\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 + 3\mathbb{E}\|\nabla f(x_t^s) - \nabla f(x_{t+1}^s)\|^2 + 3\mathbb{E}\|\frac{Q}{\eta}(x_t^s - x_{t+1}^s)\|^2\| \\
&\leq (\frac{3L_f^2}{b} + \frac{6\triangle_F^2 L_G^2}{b_2} + \frac{6\triangle_G^4 L_F^2}{b_1})\mathbb{E}\|x_t^s - \tilde{x}^s\|^2 + 3(L_f^2 + \frac{\sigma_{\max}^2(Q)}{\eta^2})\|x_t^s - x_{t+1}^s\|^2 \\
&\leq \big(5L_f^2 + \frac{3\sigma_{\max}^2(Q)}{\eta^2}\big)\theta_t^s,
\end{aligned} \tag{76}$$

where the second inequality holds by Lemma 6 and the third inequality holds by $b = m^2$, $b_1 \geq \frac{8\triangle_G^4 L_F^2 m^2}{L_f^2}$ and $b_2 \geq \frac{8\triangle_F^2 L_G^2 m^2}{L_f^2}$.

By the step 14 of Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{dist}(0, \nabla_\lambda L(x, y_{[k]}, \lambda))]_{s,t+1} &= \mathbb{E}\|Ax^s_{t+1} + By^s_{t+1} - c\|^2 \\
&= \frac{1}{\rho^2}\mathbb{E}\|\lambda^s_{t+1} - \lambda^s_t\|^2 \\
&\leq (\frac{9L_f^2}{b\sigma^A_{\min}\rho^2} + \frac{18\triangle^2_F L_G^2}{b_2\sigma^A_{\min}\rho^2} + \frac{18\triangle^4_G L_F^2}{b_1\sigma^A_{\min}\rho^2})\big(\|x^s_t - \tilde{x}^s\|^2 + \|x^s_{t-1} - \tilde{x}^s\|^2\big) + \frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho^2}\|x^s_{t+1} - x^s_t\|^2 \\
&\quad + (\frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho^2} + \frac{9L_f^2}{\sigma^A_{\min}\rho^2})\|x^s_t - x^s_{t-1}\|^2 \\
&\leq (\frac{15L_f^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho^2})\theta^s_t,
\end{aligned} \tag{77}$$

where the first inequality holds by Lemma 7 and the second inequality holds by $b = m^2$, $b_1 \geq \frac{8\triangle^4_G L_F^2 m^2}{L_f^2}$ and $b_2 \geq \frac{8\triangle^2_F L_G^2 m^2}{L_f^2}$.

By (73), we have

$$\min_{s,t} \mathbb{E}\big[\mathrm{dist}(0, \partial L(x^s_t, y^{s,t}_{[k]}, \lambda^s_t))^2\big] \leq \frac{\nu_{\max}}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\theta^s_t \leq \frac{2\nu_{\max}(\Phi^1_0 - \Phi^*)}{\gamma T}, \tag{78}$$

where $\gamma = \min(\sigma^H_{\min}, \frac{18L_f^2}{\sigma^A_{\min}\rho} + \frac{L_f}{2}, \chi_t)$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$,

$$\begin{aligned}
&\nu_1 = k\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(H)\big), \ \nu_2 = 5L_f^2 + \frac{3\sigma^2_{\max}(Q)}{\eta^2} \\
&\nu_3 = \frac{15L_f^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho^2}.
\end{aligned}$$

Given $\eta = \frac{\alpha\sigma_{\min}(Q)}{11L_f}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{735}\kappa_Q L_f}{\sigma^A_{\min}\alpha}$, it is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(1)$, which are independent on n_1 and n_2. Thus, we obtain

$$\min_{s,t} \mathbb{E}\big[\mathrm{dist}(0, \partial L(x^s_t, y^{s,t}_{[k]}, \lambda^s_t))^2\big] \leq O(\frac{1}{T}). \tag{79}$$

□

A.2 Theoretical Analysis of the com-SAGA-ADMM

In this subsection, we in detail give the convergence analysis of the com-SAGA-ADMM. We begin with giving some useful lemmas as follows:

Lemma 9. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2, the stochastic gradient $\hat{g}_t$ satisfies the following inequality:*

$$\mathbb{E}\|\hat{g}_t - \nabla f(x_t)\|^2 \leq (\frac{2\triangle_F^2 L_G^2}{bn_2} + \frac{2\triangle_F^2 L_G^2}{b_2 n_2})\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 + (\frac{2\triangle_G^4 L_F^2}{bn_2} + \frac{2\triangle_G^4 L_F^2}{b_1 n_2})\sum_{j=1}^{n_2}\mathbb{E}\|x_t - z_j^t\|^2, \tag{80}$$

where $\nabla f(x_t) = (\nabla G(x_t))^T \frac{1}{n_1}\sum_{i=1}^{n_1} F_i(G(x_t))$ with $\nabla G(x_t) = \frac{1}{n_2}\sum_{j=1}^{n_2}\nabla G_j(x_t)$ and $G(x_t) = \frac{1}{n_2}\sum_{j=1}^{n_2} G_j(x_t)$.

Proof. We begin with defining h_t to be an unbiased estimate for $\nabla f(x_t)$ (*i.e.*, $\mathbb{E}[h_t] = \nabla f(x_t)$) as follows:

$$h_t = \frac{1}{b}\sum_{i\in\mathcal{I}_t}\big((\nabla G(x_t))^T\nabla F_i(G(x_t)) - w_i^t\big) + W_t, \tag{81}$$

where $\mathcal{I}_t$ is uniformly sampled from $\{1, 2, \cdots, n_1\}$ with $|\mathcal{I}_t| = b$ and $W_t = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i^t$. Then we have

$$\begin{aligned}
\mathbb{E}\|h_t - \nabla f(x_t)\|^2 &= \mathbb{E}\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}\big((\nabla G(x_t))^T\nabla F_i(G(x_t)) - w_i^t\big) + W_t - (\nabla G(x_t))^T\frac{1}{n_1}\sum_{i=1}^{n_1}F_i(G(x_t))\|^2 \\
&= \frac{1}{b^2}\mathbb{E}\|\sum_{i\in\mathcal{I}_t}\big((\nabla G(x_t))^T\nabla F_i(G(x_t)) - w_i^t + W_t - (\nabla G(x_t))^T\frac{1}{n_1}\sum_{i=1}^{n_1}F_i(G(x_t))\big)\|^2 \\
&= \frac{1}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - w_i^t + W_t - (\nabla G(x_t))^T\frac{1}{n_1}\sum_{i=1}^{n_1}F_i(G(x_t))\|^2 \\
&\leq \frac{1}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - w_i^t\|^2 \\
&= \frac{1}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - (V_t)^T\nabla F_i(G(x_t)) + (V_t)^T\nabla F_i(G(x_t)) - (V_t)^T\nabla F_i(U_t)\|^2 \\
&\leq \underbrace{\frac{2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - (V_t)^T\nabla F_i(G(x_t))\|^2}_{L_1} + \underbrace{\frac{2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(V_t)^T\nabla F_i(G(x_t)) - (V_t)^T\nabla F_i(U_t)\|^2}_{L_2},
\end{aligned} \tag{82}$$

where the third equality holds by Lemma 5, and the first inequality holds by the equality $\mathbb{E}\|\xi - \mathbb{E}[\xi]\|^2 = \mathbb{E}\|\xi\|^2 - \|\mathbb{E}[\xi]\|^2$.

By Assumptions 1 and 2, we have

$$\begin{aligned}
L_1 &= \frac{2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - (V_t)^T\nabla F_i(G(x_t))\|^2 \\
&\leq \frac{2\triangle_F^2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\nabla G(x_t) - V_t\|^2 \\
&= \frac{2\triangle_F^2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\nabla G(x_t) - \frac{1}{n_2}\sum_{j=1}^{n_2}\nabla G(\hat{z}_j^t)\|^2 \\
&= \frac{2\triangle_F^2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\frac{1}{n_2}\sum_{j=1}^{n_2}(\nabla G(x_t) - \nabla G(\hat{z}_j^t))\|^2 \\
&\leq \frac{2\triangle_F^2}{bn_2}\sum_{j=1}^{n_2}\|\nabla G(x_t) - \nabla G(\hat{z}_j^t)\|^2 \\
&\leq \frac{2\triangle_F^2 L_G^2}{bn_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2,
\end{aligned} \tag{83}$$

where the second inequality holds by the inequality $\|\sum_{i=1}^r \alpha_i\|^2 \leq r\sum_{i=1}^r\|\alpha_i\|^2$.

Since $V_t = \frac{1}{n_2}\sum_{j=1}^{n_2}\nabla G(z_j^t)$, by Assumption 2, we have $\|V_t\|^2 \leq \triangle_G^2$. Then, we have

$$\begin{aligned}
L_2 &= \frac{2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(V_t)^T\nabla F_i(G(x_t)) - (V_t)^T\nabla F_i(U_t)\|^2 \\
&\leq \frac{2\triangle_G^2 L_F^2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|G(x_t) - U_t\|^2 \\
&= \frac{2\triangle_G^2 L_F^2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|G(x_t) - \frac{1}{n_2}\sum_{j=1}^{n_2}G(z_j^t)\|^2 \\
&= \frac{2\triangle_G^2 L_F^2}{b^2}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\frac{1}{n_2}\sum_{j=1}^{n_2}(G(x_t) - G(z_j^t))\|^2 \\
&\leq \frac{2\triangle_G^2 L_F^2}{bn_2}\sum_{j=1}^{n_2}\mathbb{E}\|G(x_t) - G(z_j^t)\|^2 \\
&\leq \frac{2\triangle_G^4 L_F^2}{bn_2}\sum_{j=1}^{n_2}\mathbb{E}\|x_t - z_j^t\|^2, \qquad (84)
\end{aligned}$$

where the first inequality holds by $\|V_t\|^2 \leq \triangle_G^2$.

Combining inequalities (82), (83) and (84), we obtain

$$\mathbb{E}\|h_t - \nabla f(x_t)\|^2 \leq \frac{2\triangle_F^2 L_G^2}{bn_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 + \frac{2\triangle_G^4 L_F^2}{bn_2}\sum_{j=1}^{n_2}\mathbb{E}\|x_t - z_j^t\|^2. \qquad (85)$$

Next, considering the upper bound of $\mathbb{E}\|h_t - \hat{g}_t\|^2$, we have

$$\begin{aligned}
&\mathbb{E}\|h_t - \hat{g}_t\|^2 = \mathbb{E}\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}\big((\nabla G(x_t))^T\nabla F_i(G(x_t)) - (\nabla\hat{G}_t)^T\nabla F_i(\hat{G}_t)\big)\|^2 \\
&\leq \frac{1}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - (\nabla\hat{G}_t)^T\nabla F_i(\hat{G}_t)\|^2 \\
&= \frac{1}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - (\nabla G(x_t))^T\nabla F_i(\hat{G}_t) + (\nabla G(x_t))^T\nabla F_i(\hat{G}_t) - (\nabla\hat{G}_t)^T\nabla F_i(\hat{G}_t)\|^2 \\
&\leq \underbrace{\frac{2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - (\nabla G(x_t))^T\nabla F_i(\hat{G}_t)\|^2}_{L_3} + \underbrace{\frac{2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(\hat{G}_t) - (\nabla\hat{G}_t)^T\nabla F_i(\hat{G}_t)\|^2}_{L_4}. \qquad (86)
\end{aligned}$$

Using Assumptions 1 and 2, we have

$$\begin{aligned}
L_3 &= \frac{2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(G(x_t)) - (\nabla G(x_t))^T\nabla F_i(\hat{G}_t)\|^2 \\
&\leq \frac{2\triangle_G^2 L_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|G(x_t) - \hat{G}_t\|^2 \\
&= \frac{2\triangle_G^2 L_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|G(x_t) - \frac{1}{b_1}\sum_{j\in\mathcal{J}_1^t}(G_j(x_t) - u_j^t) - U_t\|^2 \\
&= \frac{2\triangle_G^2 L_F^2}{bb_1^2}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_1^t}\mathbb{E}\|G(x_t) - U_t - (G_j(x_t) - u_j^t)\|^2 \\
&\leq \frac{2\triangle_G^2 L_F^2}{bb_1^2}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_1^t}\mathbb{E}\|G_j(x_t) - u_j^t\|^2 \\
&= \frac{2\triangle_G^2 L_F^2}{b_1 n_2}\sum_{j=1}^{n_2}\|G_j(x_t) - G_j(z_j^t)\|^2 \leq \frac{2\triangle_G^4 L_F^2}{b_1 n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2, \qquad (87)
\end{aligned}$$

where the third equality follows by Lemma 5 and the second inequality holds by the equality $\mathbb{E}\|\xi-\mathbb{E}[\xi]\|^2=\mathbb{E}\|\xi\|^2-\|\mathbb{E}[\xi]\|^2$. Similarly, we have

$$\begin{aligned}L_4 &= \frac{2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\nabla G(x_t))^T\nabla F_i(\hat{G}_t)-(\nabla\hat{G}_t)^T\nabla F_i(\hat{G}_t)\|^2 \\ &\leq \frac{2\triangle_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\nabla G(x_t)-\nabla\hat{G}_t\|^2 \\ &= \frac{2\triangle_F^2}{b}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\nabla G(x_t)-\frac{1}{b_2}\sum_{j\in\mathcal{J}_2^t}(\nabla G_j(x_t)-v_j^t)-V_t\|^2 \\ &\leq \frac{2\triangle_F^2}{bb_2^2}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_2^t}\mathbb{E}\|\nabla G_j(x_t)-v_j^t\|^2 \\ &= \frac{2\triangle_F^2}{b_2^2}\sum_{j\in\mathcal{J}_2^t}\mathbb{E}\|\nabla G_j(x_t)-\nabla G_j(\hat{z}_j^t)\|^2 \\ &\leq \frac{2\triangle_F^2L_G^2}{b_2n_2}\sum_{j=1}^{n_2}\|x_t-\hat{z}_j^t\|^2. \end{aligned} \tag{88}$$

Combining the inequalities (86), (87) and (88), we obtain

$$\mathbb{E}\|h_t-\hat{g}_t\|^2 \leq \frac{2\triangle_G^4L_F^2}{b_1n_2}\sum_{j=1}^{n_2}\|x_t-z_j^t\|^2+\frac{2\triangle_F^2L_G^2}{b_2n_2}\sum_{j=1}^{n_2}\|x_t-\hat{z}_j^t\|^2. \tag{89}$$

Finally, combining the inequalities (85) and (89), we have

$$\begin{aligned}\mathbb{E}\|\hat{g}_t-\nabla f(x_t)\|^2 &= \mathbb{E}\|\hat{g}_t-h_t+h_t-\nabla f(x_t)\|^2 \\ &\leq 2\mathbb{E}\|\hat{g}_t-h_t\|^2+2\mathbb{E}\|h_t-\nabla f(x_t)\|^2 \\ &\leq (\frac{2\triangle_F^2L_G^2}{bn_2}+\frac{2\triangle_F^2L_G^2}{b_2n_2})\sum_{j=1}^{n_2}\|x_t-\hat{z}_j^t\|^2+(\frac{2\triangle_G^4L_F^2}{bn_2}+\frac{2\triangle_G^4L_F^2}{b_1n_2})\sum_{j=1}^{n_2}\mathbb{E}\|x_t-z_j^t\|^2. \end{aligned} \tag{90}$$

□

Lemma 10. *Suppose the sequence $\{x_t,y_{[k]}^t,\lambda_t\}_{t=1}^T$ is generated by Algorithm 2, the following inequality holds:*

$$\begin{aligned}\mathbb{E}\|\lambda_{t+1}-\lambda_t\|^2 \leq &(\frac{18\triangle_F^2L_G^2}{\sigma_{\min}^Abn_2}+\frac{18\triangle_F^2L_G^2}{\sigma_{\min}^Ab_2n_2})\sum_{j=1}^{n_2}(\mathbb{E}\|x_t-\hat{z}_j^t\|^2+\mathbb{E}\|x_{t-1}-\hat{z}_j^{t-1}\|^2) \\ &+(\frac{18\triangle_G^4L_F^2}{\sigma_{\min}^Abn_2}+\frac{18\triangle_G^4L_F^2}{\sigma_{\min}^Ab_1n_2})\sum_{j=1}^{n_2}(\mathbb{E}\|x_t-z_j^t\|^2+\mathbb{E}\|x_{t-1}-z_j^{t-1}\|^2)+\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2}\|x_{t+1}-x_t\|^2 \\ &+(\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2}+\frac{9L_f^2}{\sigma_{\min}^A})\|x_t-x_{t-1}\|^2. \end{aligned} \tag{91}$$

Proof. By the optimize condition of the the step 8 in Algorithm 2, we have

$$\hat{g}_t+\frac{Q}{\eta}(x_{t+1}-x_t)-A^T\lambda_t+\rho A^T(Ax_{t+1}+\sum_{j=1}^kB_jy_j^{t+1}-c)=0. \tag{92}$$

Using the step 9 of Algorithm 2, then we have

$$A^T\lambda_{t+1}=\hat{g}_t+\frac{Q}{\eta}(x_{t+1}-x_t). \tag{93}$$

It follows that

$$A^T(\lambda_{t+1}-\lambda_t) = \hat{g}_t-\hat{g}_{t-1}+\frac{Q}{\eta}(x_{t+1}-x_t)-\frac{Q}{\eta}(x_t-x_{t-1}). \tag{94}$$

By Assumption 4, we have

$$\|\lambda_{t+1}-\lambda_t\|^2 \leq \frac{1}{\sigma_{\min}^A}\big[3\|\hat{g}_t-\hat{g}_{t-1}\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_{t+1}-x_t\|^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}\|x_t-x_{t-1}\|^2\big]. \tag{95}$$

Considering the upper bound of $\|\hat{g}_t^s - \hat{g}_{t-1}^s\|^2$, we have

$$\begin{aligned}\|\hat{g}_t - \hat{g}_{t-1}\|^2 &= \|\hat{g}_t - \nabla f(x_t) + \nabla f(x_t) - \nabla f(x_{t-1}) + \nabla f(x_{t-1}) - \hat{g}_{t-1}\|^2 \\ &\leq 3\|\hat{g}_t - \nabla f(x_t)\|^2 + 3\|\nabla f(x_t) - \nabla f(x_{t-1})\|^2 + 3\|\nabla f(x_{t-1}) - \hat{g}_{t-1}\|^2 \\ &\leq (\frac{6\triangle_F^2 L_G^2}{bn_2} + \frac{6\triangle_F^2 L_G^2}{b_2 n_2})\sum_{j=1}^{n_2}(\mathbb{E}\|x_t - \hat{z}_j^t\|^2 + \mathbb{E}\|x_{t-1} - \hat{z}_j^{t-1}\|^2) \\ &\quad + (\frac{6\triangle_G^4 L_F^2}{bn_2} + \frac{6\triangle_G^4 L_F^2}{b_1 n_2})\sum_{j=1}^{n_2}(\mathbb{E}\|x_t - z_j^t\|^2 + \mathbb{E}\|x_{t-1} - z_j^{t-1}\|^2) + 3L_f^2\|x_t - x_{t-1}\|^2, \end{aligned} \tag{96}$$

where the second inequality holds by lemma 9 and Assumption 1. Finally, combining the inequalities (95) and (96), we can obtain the above result. □

Lemma 11. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2, and define a* Lyapunov *function*

$$\begin{aligned}\Omega_t =& \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \rho\eta^2} + \frac{9L_f^2}{\sigma_{\min}^A\rho})\|x_t - x_{t-1}\|^2 + (\frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b} + \frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b_2})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_{t-1} - \hat{z}_j^{t-1}\|^2 \\ &+ (\frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b} + \frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b_1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_{t-1} - z_j^{t-1}\|^2 + \frac{c_t}{n_2}\sum_{i=1}^{n_2}\|x_t - z_j^t\|^2) + \frac{d_t}{n_2}\sum_{i=1}^{n_2}\|x_t - \hat{z}_j^t\|^2\big],\end{aligned}$$

where $\{c_t\}$ and $\{d_t\}$ are positive sequences satisfy

$$c_t = \begin{cases} \frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b} + \frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho b_1} + \frac{2\triangle_G^4 L_F^2}{L_f b} + \frac{2\triangle_G^4 L_F^2}{L_f b_1} + (1-p_1)(1+\beta_1)c_{t+1}, \ 0 \leq t \leq T-1, \\ 0, \ t \geq T. \end{cases}$$

$$d_t = \begin{cases} \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b} + \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho b_2} + \frac{2\triangle_F^2 L_G^2}{L_f b} + \frac{2\triangle_F^2 L_G^2}{L_f b_2} + (1-p_2)(1+\beta_2)d_{t+1}, \ 0 \leq t \leq T-1, \\ 0, \ t \geq T. \end{cases}$$

where p_1 and p_2 denote the probability of an index j being in $\mathcal{J}_t^1$ and $\mathcal{J}_t^2$, respectively. Given $b_1 = b_2 = [n_2^{2/3}]$, $b \geq \max(b_1, b_2)$, $\hat{L}_f = \frac{96\triangle_G^4 L_F^2 + 96\triangle_F^2 L_G^2 + L_f^2}{L_f}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{\hat{L}_f}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{15}\kappa_Q\hat{L}_f}{\sigma_{\min}^A\alpha}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}(\|x_t - x_{t+1}\|^2 + \sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2 + \frac{1}{b_1 n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2 + \frac{1}{b_2 n_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2) \leq \frac{\Omega_0 - \Omega^*}{\gamma T}, \tag{97}$$

where $\gamma = \min(\chi_t, \sigma_{\min}^H, \frac{\triangle_G^4 L_F^2 b_1}{L_f b} + \frac{\triangle_G^4 L_F^2}{L_f}, \frac{\triangle_F^2 L_G^2 b_2}{L_f b} + \frac{\triangle_F^2 L_G^2}{L_f})$ with $\chi_t \geq \frac{\sqrt{15}\kappa_Q\hat{L}_f}{2\alpha}$. and Ω^* denotes a low bound of Ω_t.

Proof. By optimal condition of the step 7 in Algorithm 2, we have, for $j \in [k]$

$$
\begin{aligned}
0 &= (y_j^t - y_j^{t+1})^T \big(\partial\psi_j(y_j^{t+1}) - B_j^T\lambda_t + \rho B_j^T(Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c) + H_j(y_j^{t+1} - y_j^t)\big) \\
&\leq \psi_j(y_j^t) - \psi_j(y_j^{t+1}) - (\lambda_t)^T(B_j y_j^t - B_j y_j^{t+1}) + \rho(B_j y_j^t - B_j y_j^{t+1})^T(Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c) \\
&\quad - \|y_j^{t+1} - y_j^t\|_{H_j}^2 \\
&= \psi_j(y_j^t) - \psi_j(y_j^{t+1}) - (\lambda_t)^T(Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{k} B_i y_i^t - c) + (\lambda_t)^T(Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c) \\
&\quad + \frac{\rho}{2}\|Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{k} B_i y_i^t - c\|^2 - \frac{\rho}{2}\|Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c\|^2 - \|y_j^{t+1} - y_j^t\|_{H_j}^2 \\
&\quad - \frac{\rho}{2}\|B_j y_j^t - B_j y_j^{t+1}\|^2 \\
&= \underbrace{f(x_t) + \sum_{i=1}^{j-1}\psi_i(y_i^t) + \sum_{i=j}^{k}\psi_i(y_i^{t+1}) - (\lambda_t)^T(Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{k} B_i y_i^t - c) + \frac{\rho}{2}\|Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{k} B_i y_i^t - c\|^2}_{\mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:k]}^t, \lambda_t)} \\
&\quad - \underbrace{(f(x_t) + \sum_{i=1}^{j}\psi_i(y_i^t) + \sum_{i=j+1}^{k}\psi_i(y_i^{t+1}) - (\lambda_t)^T(Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c) + \frac{\rho}{2}\|Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c\|^2}_{\mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:k]}^t, \lambda_t)} \\
&\quad - \|y_j^{t+1} - y_j^t\|_{H_j}^2 - \frac{\rho}{2}\|B_j y_j^t - B_j y_j^{t+1}\|^2 \\
&\leq \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:k]}^t, \lambda_t) - \mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:k]}^t, \lambda_t) - \sigma_{\min}(H_j)\|y_j^t - y_j^{t+1}\|^2, \qquad (98)
\end{aligned}
$$

where the first inequality holds by the convexity of function $\psi_j(y)$, and the second equality follows by applying the equality $(a-b)^T b = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ on the term $(By_j^t - By_j^{t+1})^T(Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c)$. Thus we have, for all $j \in [k]$

$$
\mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:k]}^t, \lambda_t) \leq \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:k]}^t, \lambda_t) - \sigma_{\min}(H_j)\|y_j^t - y_j^{t+1}\|^2. \qquad (99)
$$

Telescoping inequality (99) over j from 1 to k, we obtain

$$
\mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) \leq \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) - \sigma_{\min}^H \sum_{j=1}^{k} \|y_j^t - y_j^{t+1}\|^2, \qquad (100)
$$

where $\sigma_{\min}^H = \min_{j\in[k]} \sigma_{\min}(H_j)$.

By Assumption 1, we have

$$
0 \leq f(x_t) - f(x_{t+1}) + \nabla f(x_t)^T(x_{t+1} - x_t) + \frac{L_f}{2}\|x_{t+1} - x_t\|^2. \qquad (101)
$$

Using the step 8 of Algorithm 2, we have

$$
0 = (x_t - x_{t+1})^T\big(\hat{g}_t - A^T\lambda_t + \rho A^T(Ax_{t+1} + \sum_{j=1}^{k} B_j y_j^{t+1} - c) + \frac{Q}{\eta}(x_{t+1} - x_t)\big). \qquad (102)
$$

Combining (101) and (102), we have

$$
\begin{aligned}
0 \le\; & f(x_t) - f(x_{t+1}) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L_f}{2}\|x_{t+1} - x_t\|^2 \\
& + (x_t - x_{t+1})^T \big(\hat{g}_t - A^T\lambda_t + \rho A^T(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c) + \frac{Q}{\eta}(x_{t+1} - x_t)\big) \\
=\; & f(x_t) - f(x_{t+1}) + \frac{L_f}{2}\|x_t - x_{t+1}\|^2 - \frac{1}{\eta}\|x_t - x_{t+1}\|_Q^2 + (x_t - x_{t+1})^T(\hat{g}_t - \nabla f(x_t)) \\
& - (\lambda_t)^T(Ax_t - Ax_{t+1}) + \rho(Ax_t - Ax_{t+1})^T(Ax_t + \sum_{j=1}^k B_j y_j^{t+1} - c) \\
\overset{(i)}{=}\; & f(x_t) - f(x_{t+1}) + \frac{L_f}{2}\|x_t - x_{t+1}\|^2 - \frac{1}{\eta}\|x_t - x_{t+1}\|_Q^2 + (x_t - x_{t+1})^T(\hat{g}_t - \nabla f(x_t)) - (\lambda_t)^T(Ax_t + \sum_{j=1}^k B_j y_j^{t+1} - c) \\
& + (\lambda_t)^T(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c) + \frac{\rho}{2}\big(\|Ax_t + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 - \|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 - \|Ax_t - Ax_{t+1}\|^2\big) \\
=\; & \underbrace{f(x_t) + \sum_{j=1}^k \psi_j(y_j^{t+1}) - (\lambda_t)^T(Ax_t + \sum_{j=1}^k B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax_t + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2}_{\mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t)} \\
& - \underbrace{(f(x_{t+1}) + \sum_{j=1}^k \psi_j(y_j^{t+1}) - (\lambda_t)^T(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2}_{\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t)} \\
& + \frac{L_f}{2}\|x_t - x_{t+1}\|^2 + (x_t - x_{t+1})^T(\hat{g}_t - \nabla f(x_t)) - \frac{1}{\eta}\|x_t - x_{t+1}\|_Q^2 - \frac{\rho}{2}\|Ax_t - Ax_{t+1}\|^2 \\
\le\; & \mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) - \mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L_f}{2})\|x_t - x_{t+1}\|^2 + (x_t - x_{t+1})^T(\hat{g}_t - \nabla f(x_t)) \\
\overset{(ii)}{\le}\; & \mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) - \mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f)\|x_t - x_{t+1}\|^2 + \frac{1}{2L_f}\|\hat{g}_t - \nabla f(x_t)\|^2 \\
\overset{(iii)}{\le}\; & \mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) - \mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f)\|x_t - x_{t+1}\|^2 + (\frac{\triangle_F^2 L_G^2}{L_f b n_2} + \frac{\triangle_F^2 L_G^2}{L_f b_2 n_2})\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 \\
& + (\frac{\triangle_G^4 L_F^2}{L_f b n_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1 n_2})\sum_{j=1}^{n_2}\mathbb{E}\|x_t - z_j^t\|^2,
\end{aligned}
\tag{103}
$$

where the equality (i) holds by applying the equality $(a-b)^T b = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ on the term $(Ax_t - Ax_{t+1})^T(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c)$; the inequality (ii) follows by the inequality $a^T b \le \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|a\|^2$, and the inequality (iii) holds by Lemma 9. Thus, we obtain

$$
\begin{aligned}
\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) \le & \mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f)\|x_t - x_{t+1}\|^2 \\
& + (\frac{\triangle_F^2 L_G^2}{L_f b n_2} + \frac{\triangle_F^2 L_G^2}{L_f b_2 n_2})\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 + (\frac{\triangle_G^4 L_F^2}{L_f b n_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1 n_2})\sum_{j=1}^{n_2}\mathbb{E}\|x_t - z_j^t\|^2.
\end{aligned}
\tag{104}
$$

By the step 8 in Algorithm 2, we have

$$\begin{aligned}
\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) - \mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) &= \frac{1}{\rho}\|\lambda_{t+1} - \lambda_t\|^2 \\
&\leq (\frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b n_2} + \frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b_2 n_2}) \sum_{j=1}^{n_2} (\mathbb{E}\|x_t - \hat{z}_j^t\|^2 + \mathbb{E}\|x_{t-1} - \hat{z}_j^{t-1}\|^2) \\
&\quad + (\frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b n_2} + \frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b_1 n_2}) \sum_{j=1}^{n_2} (\mathbb{E}\|x_t - z_j^t\|^2 + \mathbb{E}\|x_{t-1} - z_j^{t-1}\|^2) + \frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho}\|x_{t+1} - x_t\|^2 \\
&\quad + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} + \frac{9L_f^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2. \qquad (105)
\end{aligned}$$

Combining (100), (104) and (105), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) &\leq \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) - (\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f)\|x_t - x_{t+1}\|^2 - \sigma_{\min}^H \sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2 \\
&\quad + (\frac{\triangle_F^2 L_G^2}{L_f b n_2} + \frac{\triangle_F^2 L_G^2}{L_f b_2 n_2}) \sum_{j=1}^{n_2} \|x_t - \hat{z}_j^t\|^2 + (\frac{\triangle_G^4 L_F^2}{L_f b n_2} + \frac{\triangle_G^4 L_F^2}{L_f b_1 n_2}) \sum_{j=1}^{n_2} \mathbb{E}\|x_t - z_j^t\|^2 \\
&\quad + (\frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b n_2} + \frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b_2 n_2}) \sum_{j=1}^{n_2} (\mathbb{E}\|x_t - \hat{z}_j^t\|^2 + \mathbb{E}\|x_{t-1} - \hat{z}_j^{t-1}\|^2) \\
&\quad + (\frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b n_2} + \frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b_1 n_2}) \sum_{j=1}^{n_2} (\mathbb{E}\|x_t - z_j^t\|^2 + \mathbb{E}\|x_{t-1} - z_j^{t-1}\|^2) + \frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho}\|x_{t+1} - x_t\|^2 \\
&\quad + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} + \frac{9L_f^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2. \qquad (106)
\end{aligned}$$

Next, we define a generalized *Lyapunov* function as follows:

$$\begin{aligned}
\Omega_t =& \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \rho \eta^2} + \frac{9L_f^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + (\frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b} + \frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b_2})\frac{1}{n_2}\sum_{j=1}^{n_2} \|x_{t-1} - \hat{z}_j^{t-1}\|^2 \\
&+ (\frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b} + \frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b_1})\frac{1}{n_2}\sum_{j=1}^{n_2} \|x_{t-1} - z_j^{t-1}\|^2 + \frac{c_t}{n_2}\sum_{i=1}^{n_2} \|x_t - z_j^t\|^2) + \frac{d_t}{n_2}\sum_{i=1}^{n_2} \|x_t - \hat{z}_j^t\|^2\big],
\end{aligned}$$

where $\{c_t\}$ and $\{d_t\}$ are positive sequences.

By the step 10 of Algorithm 2, we have

$$\begin{aligned}
\frac{1}{n_2}\sum_{i=1}^{n_2} \|x_{t+1} - z_i^{t+1}\|^2 &= \frac{1}{n_2}\sum_{i=1}^{n_2} \big(p_1\|x_{t+1} - x_t\|^2 + (1 - p_1)\|x_{t+1} - z_i^t\|^2\big) \\
&= \frac{p_1}{n_2}\sum_{i=1}^{n_2} \|x_{t+1} - x_t\|^2 + \frac{1 - p_1}{n_2}\sum_{i=1}^{n_2} \|x_{t+1} - z_i^t\|^2 \\
&= p_1\|x_{t+1} - x_t\|^2 + \frac{1 - p_1}{n_2}\sum_{i=1}^{n_2} \|x_{t+1} - z_i^t\|^2, \qquad (107)
\end{aligned}$$

where p_1 denotes probability of an index j being in $\mathcal{J}_t^1$. Here, we have

$$p_1 = 1 - (1 - \frac{1}{n_2})^{b_1} \geq 1 - \frac{1}{1 + b_1/n_2} = \frac{b_1/n_2}{1 + b_1/n_2} \geq \frac{b_1}{2n_2}, \qquad (108)$$

where the first inequality follows from $(1-a)^b \leq \frac{1}{1+ab}$, and the second inequality holds by $b_1 \leq n_2$. Considering the upper bound of $\|x_{t+1} - z_i^t\|^2$, we have

$$\begin{aligned} \|x_{t+1} - z_i^t\|^2 &= \|x_{t+1} - x_t + x_t - z_i^t\|^2 \\ &= \|x_{t+1} - x_t\|^2 + 2(x_{t+1} - x_t)^T(x_t - z_i^t) + \|x_t - z_i^t\|^2 \\ &\leq \|x_{t+1} - x_t\|^2 + 2\big(\frac{1}{2\beta_1}\|x_{t+1} - x_t\|^2 + \frac{\beta_1}{2}\|x_t - z_i^t\|^2\big) + \|x_t - z_i^t\|^2 \\ &= (1 + \frac{1}{\beta_1})\|x_{t+1} - x_t\|^2 + (1+\beta_1)\|x_t - z_i^t\|^2, \end{aligned} \tag{109}$$

where $\beta_1 > 0$.

Combining (107) with (109), we have

$$\frac{1}{n_2}\sum_{i=1}^{n_2}\|x_{t+1} - z_i^{t+1}\|^2 \leq (1 + \frac{1-p_1}{\beta_1})\|x_{t+1} - x_t\|^2 + \frac{(1-p_1)(1+\beta_1)}{n_2}\sum_{i=1}^{n_2}\|x_t - z_i^t\|^2. \tag{110}$$

Similarly, by the step 11 of Algorithm 2, we have

$$\frac{1}{n_2}\sum_{i=1}^{n_2}\|x_{t+1} - \hat{z}_i^{t+1}\|^2 \leq (1 + \frac{1-p_2}{\beta_2})\|x_{t+1} - x_t\|^2 + \frac{(1-p_2)(1+\beta_2)}{n_2}\sum_{i=1}^{n_2}\|x_t - \hat{z}_i^t\|^2, \tag{111}$$

where $\beta_2 > 0$, p_2 denotes probability of an index j being in $\mathcal{J}_t^2$ and $p_2 \geq \frac{b_2}{2n_2}$.

It follows that

$$\begin{aligned} \Omega_{t+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) + (\frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\rho\eta^2} + \frac{9L_f^2}{\sigma^A_{\min}\rho})\|x_{t+1} - x_t\|^2 + (\frac{18\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b} + \frac{18\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b_2})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 \\ &\quad + (\frac{18\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b} + \frac{18\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b_1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2 + \frac{c_t}{n_2}\sum_{i=1}^{n_2}\|x_{t+1} - z_j^{t+1}\|^2 + \frac{d_t}{n_2}\sum_{i=1}^{n_2}\|x_{t+1} - \hat{z}_j^{t+1}\|^2\big] \\ &\leq \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\sigma^2_{\max}(Q)}{\sigma^A_{\min}\rho\eta^2} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x_t - x_{t-1}\|^2 + (\frac{18\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b} + \frac{18\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b_2})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_{t-1} - \hat{z}_j^{t-1}\|^2 \\ &\quad + (\frac{18\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b} + \frac{18\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b_1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_{t-1} - z_j^{t-1}\|^2 \\ &\quad + (\frac{36\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b} + \frac{36\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b_1} + \frac{2\triangle_G^4 L_F^2}{L_f b} + \frac{2\triangle_G^4 L_F^2}{L_f b_1} + (1-p_1)(1+\beta_1)c_{t+1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2 \\ &\quad + (\frac{36\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b} + \frac{36\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b_2} + \frac{2\triangle_F^2 L_G^2}{L_f b} + \frac{2\triangle_F^2 L_G^2}{L_f b_2} + (1-p_2)(1+\beta_2)d_{t+1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 \\ &\quad - (\frac{\triangle_F^2 L_G^2}{L_f b} + \frac{\triangle_F^2 L_G^2}{L_f b_2})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 - (\frac{\triangle_G^4 L_F^2}{L_f b} + \frac{\triangle_G^4 L_F^2}{L_f b_1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2 - \sigma^H_{\min}\sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2 \\ &\quad - \big(\frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L_f - \frac{6\kappa_A\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - (1 + \frac{1-p_1}{\beta_1})c_{t+1} - (1 + \frac{1-p_2}{\beta_2})d_{t+1}\big)\|x_t - x_{t+1}\|^2 \\ &= \Omega_t - \chi_t\|x_t - x_{t+1}\|^2 - \sigma^H_{\min}\sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2 - (\frac{\triangle_F^2 L_G^2}{L_f b} + \frac{\triangle_F^2 L_G^2}{L_f b_2})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 \\ &\quad - (\frac{\triangle_G^4 L_F^2}{L_f b} + \frac{\triangle_G^4 L_F^2}{L_f b_1})\frac{1}{n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2, \end{aligned} \tag{112}$$

where $c_{t+1} = \frac{36\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b} + \frac{36\triangle_G^4 L_F^2}{\sigma^A_{\min}\rho b_1} + \frac{2\triangle_G^4 L_F^2}{L_f b} + \frac{2\triangle_G^4 L_F^2}{L_f b_1} + (1-p_1)(1+\beta_1)c_{t+1}$, $d_t = \frac{36\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b} + \frac{36\triangle_F^2 L_G^2}{\sigma^A_{\min}\rho b_2} + \frac{2\triangle_F^2 L_G^2}{L_f b} + \frac{2\triangle_F^2 L_G^2}{L_f b_2} + (1-p_2)(1+\beta_2)d_{t+1}$, and $\chi_t = \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L_f - \frac{6\sigma^2_{\max}(Q)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - (1 + \frac{1-p_1}{\beta_1})d_{t+1} - (1 + \frac{1-p_2}{\beta_2})c_{t+1}$.

Let $c_T = 0$ and $\beta_1 = \frac{b_1}{4n_2}$. Since $(1-p_1)(1+\beta_1) = 1+\beta_1-p_1-p_1\beta_1 \leq 1+\beta_1-p_1$ and $p_1 \geq \frac{b_1}{2n_2}$, it follows that

$$\begin{aligned} c_t &\leq c_{t+1}(1-\theta_1) + \frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b} + \frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho b_1} + \frac{2\triangle_G^4 L_F^2}{L_f b} + \frac{2\triangle_G^4 L_F^2}{L_f b_1} \\ &\leq c_{t+1}(1-\theta_1) + \frac{3\triangle_G^4 L_F^2}{L_f b} + \frac{3\triangle_G^4 L_F^2}{L_f b_1} \end{aligned} \tag{113}$$

where $\theta_1 = p_1 - \beta_1 \geq \frac{b_1}{4n_2}$ and the second inequality holds by $\rho \geq \frac{36L_f}{\sigma_{\min}^A}$. Then recursing on t, for $0 \leq t \leq T-1$, we have

$$c_t \leq \frac{3}{b_1}(\frac{\triangle_G^4 L_F^2}{L_f} + \frac{\triangle_G^4 L_F^2 b_1}{L_f b})\frac{1-\theta_1^{T-t}}{\theta_1} \leq \frac{12n_2}{b_1^2}(\frac{\triangle_G^4 L_F^2}{L_f} + \frac{\triangle_G^4 L_F^2 b_1}{L_f b}). \tag{114}$$

Similarly, let $d_T = 0$ and and $\beta_2 = \frac{b_2}{4n_2}$. Since $(1-p_2)(1+\beta_2) = 1+\beta_2-p_2-p_2\beta_2 \leq 1+\beta_2-p_2$ and $p_2 \geq \frac{b_2}{2n_2}$, it follows that

$$\begin{aligned} d_t &\leq d_{t+1}(1-\theta_2) + \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b} + \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho b_2} + \frac{2\triangle_F^2 L_G^2}{L_f b} + \frac{2\triangle_F^2 L_G^2}{L_f b_2} \\ &\leq d_{t+1}(1-\theta_2) + \frac{3\triangle_F^2 L_G^2}{L_f b} + \frac{3\triangle_F^2 L_G^2}{L_f b_2}, \end{aligned} \tag{115}$$

where $\theta_2 = p_2 - \beta_2 \geq \frac{b_2}{4n_2}$ and the second inequality holds by $\rho \geq \frac{36L_f}{\sigma_{\min}^A}$. Then recursing on t, for $0 \leq t \leq T-1$, we have

$$d_t \leq \frac{3}{b_2}(\frac{\triangle_F^2 L_G^2}{L_f} + \frac{\triangle_F^2 L_G^2 b_2}{L_f b})\frac{1-\theta_2^{T-t}}{\theta_2} \leq \frac{12n_2}{b_2^2}(\frac{\triangle_F^2 L_G^2}{L_f} + \frac{\triangle_F^2 L_G^2 b_2}{L_f b}). \tag{116}$$

Thus, we have

$$\begin{aligned} \chi_t &= \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L_f^2}{\sigma_{\min}^A \rho} - (1+\frac{1-p_1}{\beta_1})c_{t+1} - (1+\frac{1-p_2}{\beta_2})d_{t+1} \\ &\geq \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L_f^2}{\sigma_{\min}^A \rho} - (1+\frac{4n_2-2b_1}{b_1})\frac{12n_2}{b_1^2}(\frac{\triangle_G^4 L_F^2}{L_f} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}) \\ &\quad - (1+\frac{4n_2-2b_2}{b_2})\frac{12n_2}{b_2^2}(\frac{\triangle_F^2 L_G^2}{L_f} + \frac{\triangle_F^2 L_G^2 b_2}{L_f b}) \\ &= \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L_f^2}{\sigma_{\min}^A \rho} - (\frac{4n_2}{b_1}-1)\frac{12n_2}{b_1^2}(\frac{\triangle_G^4 L_F^2}{L_f} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}) \\ &\quad - (\frac{4n_2}{b_2}-1)\frac{12n_2}{b_2^2}(\frac{\triangle_F^2 L_G^2}{L_f} + \frac{\triangle_F^2 L_G^2 b_2}{L_f b}) \\ &\geq \frac{\sigma_{\min}(Q)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_f - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L_f^2}{\sigma_{\min}^A \rho} - \frac{48n_2^2}{b_1^3}(\frac{\triangle_G^4 L_F^2}{L_f} + \frac{\triangle_G^4 L_F^2 b}{L_f b_1}) \\ &\quad - \frac{48n_2^2}{b_2^3}(\frac{\triangle_F^2 L_G^2}{L_f} + \frac{\triangle_F^2 L_G^2 b_2}{L_f b}) \\ &= \underbrace{\frac{\sigma_{\min}(Q)}{\eta} - L_f - \frac{48n_2^2}{b_1^3}(\frac{\triangle_G^4 L_F^2}{L_f} + \frac{\triangle_G^4 L_F^2 b_1}{L_f b}) - \frac{48n_2^2}{b_2^3}(\frac{\triangle_F^2 L_G^2}{L_f} + \frac{\triangle_F^2 L_G^2 b_2}{L_f b})}_{M_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L_f^2}{\sigma_{\min}^A \rho}}_{M_2} \end{aligned} \tag{117}$$

Let $b_1 = b_2 = [n_2^{2/3}]$, $b \geq \max(b_1, b_2)$ and $0 < \eta \leq \frac{L_f \sigma_{\min}(Q)}{96\triangle_G^4 L_F^2 + 96\triangle_F^2 L_G^2 + L_f^2}$, we have

$$\begin{aligned} M_1 &= \frac{\sigma_{\min}(Q)}{\eta} - L_f - \frac{48n_2^2}{b_1^3}(\frac{\triangle_G^4 L_F^2}{L_f} + \frac{\triangle_G^4 L_F^2 b_1}{L_f b}) - \frac{48n_2^2}{b_2^3}(\frac{\triangle_F^2 L_G^2}{L_f} + \frac{\triangle_F^2 L_G^2 b_2}{L_f b}) \\ &\geq \frac{\sigma_{\min}(Q)}{\eta} - L_f - \frac{96\triangle_G^4 L_F^2}{L_f} - \frac{96\triangle_F^2 L_G^2}{L_f} \geq 0. \end{aligned} \tag{118}$$

For notational simplicity, let $\hat{L}_f = \frac{96\triangle_G^4 L_F^2 + 96\triangle_F^2 L_G^2 + L_f^2}{L_f}$. Further let $\eta = \frac{\alpha\sigma_{\min}(Q)}{\hat{L}_f}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{15}\kappa_Q \hat{L}_f}{\sigma_{\min}^A \alpha}$, we have

$$\begin{aligned}
M_2 &= \frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L_f^2}{\sigma_{\min}^A \rho} \\
&= \frac{\rho\sigma_{\min}^A}{2} - \frac{6\kappa_Q^2 \hat{L}_f^2}{\sigma_{\min}^A \rho\alpha^2} - \frac{9L_f^2}{\sigma_{\min}^A \rho} \\
&\geq \frac{\rho\sigma_{\min}^A}{2} - \frac{6\kappa_Q^2 \hat{L}_f^2}{\sigma_{\min}^A \rho\alpha^2} - \frac{9\kappa_Q^2 \hat{L}_f^2}{\sigma_{\min}^A \rho\alpha^2} \\
&= \frac{\rho\sigma_{\min}^A}{4} + \underbrace{\frac{\rho\sigma_{\min}^A}{4} - \frac{15\kappa_Q^2 \hat{L}_f^2}{\sigma_{\min}^A \rho\alpha^2}}_{\geq 0}, \\
&\geq \frac{\sqrt{15}\kappa_Q \hat{L}_f}{2\alpha}
\end{aligned} \tag{119}$$

where the first inequality holds by $\hat{L}_f > L_f$. Then we have $\chi_t \geq \frac{\sqrt{15}\kappa_Q \hat{L}_f}{2\alpha}$.

Thus, by (112), we have the sequence $\{\Omega_t\}$ is decreasing. Next, by the definition of Ω_t, we have

$$\begin{aligned}
\Omega_t &\geq \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^{s,t}, \lambda_t)\big] \\
&= f(x_t) + \sum_{j=1}^k \psi_j(y_j^t) - (\lambda_t)^T(Ax_t + \sum_{j=1}^m B_j y_j^{s,t} - c) + \frac{\rho}{2}\|Ax_t + \sum_{j=1}^k B_j y_j^{s,t} - c\| \\
&= f(x_t) + \sum_{j=1}^k \psi_j(y_j^t) - \frac{1}{\rho}(\lambda_t)^T(z_{t-1} - z_t) + \frac{1}{2\rho}\|\lambda_t - \lambda_{t-1}\|^2 \\
&= f(x_t) + \sum_{j=1}^k \psi_j(y_j^t) - \frac{1}{2\rho}\|\lambda_{t-1}\|^2 + \frac{1}{2\rho}\|\lambda_t\|^2 + \frac{1}{\rho}\|\lambda_t - \lambda_{t-1}\|^2 \\
&\geq f^* + \sum_{j=1}^k \psi_j^* - \frac{1}{2\rho}\|\lambda_{t-1}\|^2 + \frac{1}{2\rho}\|\lambda_t\|^2.
\end{aligned} \tag{120}$$

Summing the inequality (120) over $t = 0, 1 \cdots, T$, we have

$$\frac{1}{T}\sum_{t=1}^T \Omega_t^s \geq f^* + \sum_{j=1}^k \psi_j^* - \frac{1}{2\rho}\|\lambda_0^1\|^2. \tag{121}$$

Thus, the function Ω_t is bounded from below. Let Ω^* denotes a low bound of Ω_t.

Finally, telescoping inequality (112) over t from 0 to T, we have

$$\frac{1}{T}\sum_{t=1}^T (\|x_t - x_{t+1}\|^2 + \sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2 + \frac{1}{b_1 n_2}\sum_{j=1}^{n_2}\|x_t - z_j^t\|^2 + \frac{1}{b_2 n_2}\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2) \leq \frac{\Omega_0 - \Omega^*}{\gamma T}, \tag{122}$$

where $\gamma = \min(\chi_t, \sigma_{\min}^H, \frac{\triangle_G^4 L_F^2 b_1}{L_f b} + \frac{\triangle_G^4 L_F^2}{L_f}, \frac{\triangle_F^2 L_G^2 b_2}{L_f b} + \frac{\triangle_F^2 L_G^2}{L_f})$ with $\chi_t \geq \frac{\sqrt{15}\kappa_Q \hat{L}_f}{2\alpha}$.

□

Next, based on the above lemmas, we give convergence analysis of the com-SAGA-ADMM. First, let

$$\begin{aligned}
&\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big), \ \nu_2 = 12\triangle_F^2 L_G^2 + 12\triangle_G^4 L_F^2 + \frac{3\sigma_{\max}^2(Q)}{\eta^2} \\
&\nu_3 = \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A \rho^2} + \frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A \rho^2} + \frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2\rho^2} + \frac{9L_f^2}{\sigma_{\min}^A \rho^2}.
\end{aligned}$$

Theorem 4. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2. Let $b_1 = b_2 = [n_2^{2/3}]$, $b \geq \max(b_1, b_2)$, $\hat{L}_f = \frac{96\triangle_G^4 L_F^2 + 96\triangle_F^2 L_G^2 + L_f^2}{L_f}$, $\eta = \frac{\alpha\sigma_{\min}(Q)}{\hat{L}_f}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{15}\kappa_Q \hat{L}_f}{\sigma_{\min}^A \alpha}$, we have*

$$\min_{1 \leq t \leq T} \mathbb{E}\big[dist(0, \partial L(x_t, y_{[k]}^t, \lambda_t))^2\big] \leq \frac{2\nu_{\max}(\Omega_0 - \Omega^*)}{\gamma T} = O(\frac{1}{T}), \tag{123}$$

where $\gamma = \min(\chi_t, \sigma_{\min}^H, \frac{\triangle_G^4 L_F^2 b_1}{L_f b} + \frac{\triangle_G^4 L_F^2}{L_f}, \frac{\triangle_F^2 L_G^2 b_2}{L_f b} + \frac{\triangle_F^2 L_G^2}{L_f})$ *with* $\chi_t \geq \frac{\sqrt{15}\kappa_Q \hat{L}_f}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ *and* Ω^* *is a lower bound of function* Ω_t*. It implies that the iteration number* T *satisfies*

$$T = \frac{4\kappa_{\max}}{\epsilon\gamma}(\Omega_0 - \Omega^*),$$

then $(x_{t^*}, y_{[k]}^{t^*}, \lambda_{t^*})$ *is an* ϵ*-approximate solution of* (6)*, where* $t^* = \arg\min_{1\leq t\leq T}\theta_t$*.*

Proof. We begin with defining an useful variable $\theta_t = \|x_{t+1} - x_t\|^2 + \|x_t - x_{t-1}\|^2 + \frac{1}{b_1 n_2}\sum_{j=1}^{n_2}(\|x_t - z_j^t\|^2 + \|x_{t-1} - z_j^{t-1}\|^2) + \frac{1}{b_2 n_2}\sum_{j=1}^{n_2}(\|x_t - \hat{z}_j^t\|^2 + \|x_{t-1} - \hat{z}_j^{t-1}\|^2) + \sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2)$. By the optimal condition of the step 7 in Algorithm 2, we have, for all $j \in [k]$

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \partial_{y_j} L(x, y_{[k]}, \lambda))^2\big]_{t+1} &= \mathbb{E}\big[\text{dist}(0, \partial g_j(y_j^{t+1}) - B_j^T\lambda_{t+1})^2\big] \\
&= \|B_j^T\lambda_t - \rho B_j^T(Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{k} B_i y_i^t - c) - H_j(y_j^{t+1} - y_j^t) - B_j^T\lambda_{t+1}\|^2 \\
&= \|\rho B_j^T A(x_{t+1} - x_t) + \rho B_j^T \sum_{i=j+1}^{k} B_i(y_i^{t+1} - y_i^t) - H_j(y_j^{t+1} - y_j^t)\|^2 \\
&\leq k\rho^2\sigma_{\max}^{B_j}\sigma_{\max}^A\|x_{t+1} - x_t\|^2 + k\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^{k}\sigma_{\max}^{B_i}\|y_i^{t+1} - y_i^t\|^2 \\
&\quad + k\sigma_{\max}^2(H_j)\|y_j^{t+1} - y_j^t\|^2 \\
&\leq k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big)\theta_t, \qquad (124)
\end{aligned}$$

where the first inequality follows by the inequality $\|\frac{1}{n}\sum_{i=1}^n z_i\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|z_i\|^2$.

By the step 8 in Algorithm 2, we have

$$\begin{aligned}
\mathbb{E}[\text{dist}(0, \nabla_x L(x, y_{[k]}, \lambda))]_{t+1} &= \mathbb{E}\|A^T\lambda_{t+1} - \nabla f(x_{t+1})\|^2 \\
&= \mathbb{E}\|\hat{g}_t - \nabla f(x_{t+1}) - \frac{G}{\eta}(x_t - x_{t+1})\|^2 \\
&= \mathbb{E}\|\hat{g}_t - \nabla f(x_t) + \nabla f(x_t) - \nabla f(x_{t+1}) - \frac{G}{\eta}(x_t - x_{t+1})\|^2 \\
&\leq 3\mathbb{E}\|\hat{g}_t - \nabla f(x_t)\|^2 + 3\mathbb{E}\|\nabla f(x_t) - \nabla f(x_{t+1})\|^2 + 3\mathbb{E}\|\frac{G}{\eta}(x_t - x_{t+1})\|^2 \\
&\leq (\frac{6\triangle_F^2 L_G^2}{bn_2} + \frac{6\triangle_F^2 L_G^2}{b_2 n_2})\sum_{j=1}^{n_2}\|x_t - \hat{z}_j^t\|^2 + (\frac{6\triangle_G^4 L_F^2}{bn_2} + \frac{6\triangle_G^4 L_F^2}{b_1 n_2})\sum_{j=1}^{n_2}\mathbb{E}\|x_t - z_j^t\|^2 \\
&\quad + 3(L_f^2 + \frac{\sigma_{\max}^2(Q)}{\eta^2})\|x_t - x_{t+1}\|^2 \\
&\leq \big(12\triangle_F^2 L_G^2 + 12\triangle_G^4 L_F^2 + 3L_f^2 + \frac{3\sigma_{\max}^2(Q)}{\eta^2}\big)\theta_t, \qquad (125)
\end{aligned}$$

where the second inequality holds by $b \geq \max(b_1, b_2)$.

By the step 9 of Algorithm 2, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{dist}(0,\nabla_\lambda L(x,y_{[k]},\lambda))]_{t+1} &= \mathbb{E}\|Ax_{t+1}+\sum_{j=1}^k B_j y_j^{t+1}-c\|^2 \\
&= \frac{1}{\rho^2}\mathbb{E}\|\lambda_{t+1}-\lambda_t\|^2 \\
&\leq (\frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A b n_2 \rho^2}+\frac{18\triangle_F^2 L_G^2}{\sigma_{\min}^A b_2 n_2 \rho^2})\sum_{j=1}^{n_2}(\mathbb{E}\|x_t-\hat{z}_j^t\|^2+\mathbb{E}\|x_{t-1}-\hat{z}_j^{t-1}\|^2) \\
&\quad + (\frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A b n_2 \rho^2}+\frac{18\triangle_G^4 L_F^2}{\sigma_{\min}^A b_1 n_2 \rho^2})\sum_{j=1}^{n_2}(\mathbb{E}\|x_t-z_j^t\|^2+\mathbb{E}\|x_{t-1}-z_j^{t-1}\|^2)+\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2\rho^2}\|x_{t+1}-x_t\|^2 \\
&\quad + (\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A \eta^2\rho^2}+\frac{9L_f^2}{\sigma_{\min}^A\rho^2})\|x_t-x_{t-1}\|^2 \\
&\leq (\frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho^2}+\frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho^2}+\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho^2}+\frac{9L_f^2}{\sigma_{\min}^A\rho^2})\theta_t,
\end{aligned} \tag{126}$$

where the second inequality holds by $b \geq \max(b_1, b_2)$.

Using (122), we have

$$\min_{1\leq t\leq T}\mathbb{E}\big[\mathrm{dist}(0,\partial L(x_t,y_{[k]}^t,\lambda_t))^2\big] \leq \frac{\nu_{\max}}{T}\sum_{t=1}^T\theta_t \leq \frac{2\nu_{\max}(\Omega_0-\Omega^*)}{\gamma T}, \tag{127}$$

where $\gamma=\min(\chi_t,\sigma_{\min}^H,\frac{\triangle_G^4 L_F^2 b_1}{L_f b}+\frac{\triangle_G^4 L_F^2}{L_f},\frac{\triangle_F^2 L_G^2 b_2}{L_f b}+\frac{\triangle_F^2 L_G^2}{L_f})$ with $\chi_t \geq \frac{\sqrt{15}\kappa_Q\hat{L}_f}{2\alpha}$, $\nu_{\max}=\max(\nu_1,\nu_2,\nu_3)$,

$$\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A+\rho^2(\sigma_{\max}^B)^2+\sigma_{\max}^2(H)\big),\ \nu_2 = 12\triangle_F^2 L_G^2+12\triangle_G^4 L_F^2+\frac{3\sigma_{\max}^2(Q)}{\eta^2}$$

$$\nu_3 = \frac{36\triangle_F^2 L_G^2}{\sigma_{\min}^A\rho^2}+\frac{36\triangle_G^4 L_F^2}{\sigma_{\min}^A\rho^2}+\frac{3\sigma_{\max}^2(Q)}{\sigma_{\min}^A\eta^2\rho^2}+\frac{9L_f^2}{\sigma_{\min}^A\rho^2}.$$

Given $\eta=\frac{\alpha\sigma_{\min}(Q)}{\hat{L}_f}$ $(0<\alpha\leq 1)$ and $\rho=\frac{2\sqrt{15}\kappa_Q\hat{L}_f}{\sigma_{\min}^A\alpha}$, since k is relatively small, it is easy verifies that $\gamma=O(1)$ and $\nu_{\max}=O(1)$, which are independent on n and d. Thus, we obtain

$$\min_{1\leq t\leq T}\mathbb{E}\big[\mathrm{dist}(0,\partial L(x_t,y_{[k]}^t,\lambda_t))^2\big] \leq O(\frac{1}{T}). \tag{128}$$

□