Asynchronous Distributed Stochastic ADMMs for Nonconvex Optimization

Abstract

Asynchronous parallel optimization has been playing a key role in solving the large-scale machine learning problems. Due to the flexibility in splitting the objective into multiple parts, alternating direction method of multipliers (ADMM) is a popular distributed optimization tool in machine learning. Recently, some works have devoted to studying asynchronous distributed ADMM methods, but few focus on large-scale stochastic optimization. In addition, these methods only focus on machine learning problems with a simple nonsmooth regularization. Clearly, these methods are limited in many applications. In the paper, thus, we focus on the large-scale nonconvex problems with multiple nonsmooth regularization penalties, and propose a class of fast asynchronous distributed stochastic ADMM methods (*i.e.*, AsyDS-ADMM and AsyDS-ADMM+). Moreover, we prove that both the AsyDS-ADMM and AsyDS-ADMM+ have convergence rate of $O(\frac{1}{T})$, where T denotes the number of iterations. Our theoretical analysis also shows that the AsyDS-ADMM and AsyDS-ADMM+ have the optimal incremental first-order oracle (IFO) complexities of $O(\epsilon^{-2})$ and $O(n + n^{2/3}\epsilon^{-1})$ for finding an ϵ-approximate solution, respectively, where n denotes the whole sample size. In particular, this is the first analysis on both the convergence rate and the optimal IFO of asynchronous distributed ADMM for the nonconvex nonsmooth optimization. Finally, the experimental results on benchmark datasets validate efficiency of the proposed algorithms.

1 Introduction

Asynchronous parallel optimization has been playing a pivotal role in solving the large-scale machine learning problems, such as training deep neural networks [Zhang *et al.*, 2016], kernel methods [You *et al.*, 2016] and matrix factorization [Zhang and Kwok, 2014]. In general, the existing asynchronous algorithms are roughly divided into two categories: one is on shared-memory architecture and the other is on distributed-memory architecture. For example, [Recht *et al.*, 2011; Reddi *et al.*, 2015; Davis *et al.*, 2016] proposed some fast asynchronous parallel stochastic methods based on shared-memory machines, which use the inconsistent read setting, implying it can does not guarantee the atomic operation on the whole parameters. At the same time, [Zhang and Kwok, 2014; Meng *et al.*, 2017; Hong, 2018] presented some fast asynchronous parallel/distributed stochastic methods based on distributed-memory machines, which use the consistent read setting, implying it guarantees the atomic operation on the whole parameters. In this paper, we mainly focus on the asynchronous distributed algorithms.

Alternating direction method of multipliers (ADMM [Gabay and Mercier, 1976; Boyd *et al.*, 2011]) is a popular distributed optimization tool in machine learning, due to the flexibility in splitting the objective into multiple parts. For example, in the large-scale distributed optimization, ADMM generally considers the following global variable consensus optimization:

$$\min_{\{x_i,z\}_{i=1}^n} \sum_{i=1}^n f_i(x_i), \ \text{s.t.}\ x_i = z,\ i = 1, \cdots, n. \quad (1)$$

where x_i is i-th node local copy of the parameter to be learned, and z is the so-called consensus variable. Recently, [Boyd *et al.*, 2011] proposed a synchronized distributed ADMM algorithm, where the master needs to wait for all the workers to finish their parameters' update before it can proceed. Clearly, this synchronized distributed ADMM method is slow due to waiting for the slowest worker, and is sensitive to failure of individual worker. To handle this drawback, [Iutzeler *et al.*, 2013; Zhang and Kwok, 2014] proposed an efficient asynchronous distributed ADMM method for the problem (1). Further, [Chang *et al.*, 2016] proposed the asynchronous distributed ADMM method to solve large-scale (non)convex problems with the nonsmooth regularization:

$$\min_{\{x_i\}_{i=0}^n} \sum_{i=1}^n f_i(x_i) + h(x_0), \ \text{s.t.}\ x_i = x_0,\ i = 1, \cdots, n, \quad (2)$$

where $h(x_0) : \mathbb{R}^d \to \mathbb{R}$ is a non-smooth regularization penalty such as sparsity . At the same time, [Hong, 2018] proposed an asynchronous distributed ADMM method specially for the above nonconvex nonsmooth problem (2).

Table 1: Convergence property comparison of the **nonconvex asynchronous** distributed 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. T is the whole iteration number; n denotes the sample size.)

Algorithm	Reference	Problem	Convergence rate	IFO
async-ADMM	[Zhang and Kwok, 2014]	C(S)	$O(\frac{1}{T})$	$O(n\epsilon^{-1})$
ASYSG-CON	[Lian *et al.*, 2015]	NC(S)	$O(\sqrt{\frac{1}{T}})$	$O(\epsilon^{-2})$
Distributed-AsySVRG	[Huo and Huang, 2017]	NC(S)	$O(\frac{1}{T})$	$O(n+n^{\frac{2}{3}}\epsilon^{-1})$
AD-ADMM	[Chang *et al.*, 2016; Hong, 2018]	NC(S) + C(NS)	unknown	unknown
Asyn-ProxSGD	[Zhu *et al.*, 2018]	NC(S) + C(NS)	$O(\sqrt{\frac{1}{T}})$	$O(\epsilon^{-2})$
Asyn-ProxSVRG	[Yu and Huang, 2018]	NC(S) + C(NS)	$O(\frac{1}{T})$	$O(n+n^{\frac{2}{3}}\epsilon^{-1})$
AsyDS-ADMM	Ours	NC(S) + C(mNS)	$O(\frac{1}{T})$	$O(\epsilon^{-2})$
AsyDS-ADMM+			$O(\frac{1}{T})$	$O(n+n^{\frac{2}{3}}\epsilon^{-1})$

Although the above asynchronous distributed ADMM methods are effective in solving the large-scale machine learning problems, there still exist three major **drawbacks**:

- In these algorithms, each worker need to calculate local full gradient at the respective local dataset, then use this gradient to update local parameters or send it to master. In the big data problems, even each worker may store lots of data. In this case, such a large computation slows down the whole algorithm.
- These methods only consider a simple nonsmooth regularization or even do not, which are clearly not competent to many machine learning problems such as those with multiple nonsmooth regularization penalties or some complex regularization penalties, *e.g.*, overlapping group lasso.
- Although these asynchronous ADMM methods are proposed for solving the nonconvex nonsmooth problems, the *convergence rate* and the optimal *incremental first-order oracle* (IFO) of these methods for the nonconvex optimization still are not studied.

In the paper, to circumvent these drawbacks, we propose a class of fast asynchronous distributed stochastic ADMM methods to solve the following nonconvex problem:

$$\min_{x,\{y_j\}_{j=1}^k} \frac{1}{n}\sum_{i=1}^n f_i(x) + \sum_{j=1}^k h_j(y_j) \quad (3)$$
$$\text{s.t. } Ax + \sum_{j=1}^k B_j y_j = c,$$

where $y_{[k]} = \{y_1, \cdots, y_k\}$, $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) : \mathbb{R}^d \to \mathbb{R}$ is a *nonconvex* and smooth function, and $h_j(y_j) : \mathbb{R}^q \to \mathbb{R}$ is a convex and possibly *nonsmooth* function for all $j \in [k]$, $k \geq 1$. In machine learning, function $f(x)$ can be used for the empirical loss, $\sum_{j=1}^k h_j(y_j)$ can be used not only for single structure penalty (*e.g.,* sparse) but also for superposition structure penalties (*e.g.*, sparse + group sparse). Specifically, we propose a fast asynchronous distributed stochastic ADMM (AsyDS-ADMM) method based on the stochastic ADMM method [Ouyang *et al.*, 2013]. To accelerate AsyDS-ADMM, we further propose a faster asynchronous distributed stochastic ADMM (AsyDS-ADMM+) method by using the variance reduced method of SVRG [Johnson and Zhang, 2013].

1.1 Challenges and Contributions

Although the exsiting stochastic ADMM methods [Ouyang *et al.*, 2013; Zheng and Kwok, 2016a; Zheng and Kwok, 2016b] have shown good performances in the large-scale problems, applying these stochastic methods to the asynchronous distributed algorithm is *not trivial*. There are at least two **challenges**:

- These existing stochastic ADMM methods rely on the assumption that the stochastic gradient is an *unbiased* estimate of true full gradient, which does not hold in the asynchronous algorithms due to using the *delayed* gradients;
- Due to failure of the Féjer monotonicity of iteration, convergence analysis of the nonconvex ADMM is generally quite difficult [Wang *et al.*, 2015]. Clearly, due to using the *delayed* stochastic gradients, this difficulty is more serious in nonconvex asynchronous ADMM method.

In the paper, thus, we will fill this gap between the stochastic ADMM methods and the asynchronous distributed algorithms. In summary, our main **contributions** are given below:

1) For solving the problem (3), we propose a class of fast asynchronous distributed stochastic ADMM methods (*i.e.*, AsyDS-ADMM and AsyDS-ADMM+) based on the existing stochastic ADMM methods.
2) We prove that both the AsyDS-ADMM and AsyDS-ADMM+ have convergence rate of $O(\frac{1}{T})$, and the optimal IFO complexities of $O(\epsilon^{-2})$ and $O(n+n^{2/3}\epsilon^{-1})$ for finding an ϵ-approximate solution, respectively. In particular, if set the delay number $\tau = 0$ in our theoretical analysis, we can obtain the optimal IFO complexity of the existing nonconvex SVRG-ADMM method, which is very important but has not been solved by the previous works [Zheng and Kwok, 2016b; Huang *et al.*, 2016].
3) Extensive experimental results and theoretical analysis demonstrate the scalability of our methods.

2 Related Works

The nonconvex asynchronous distributed stochastic optimization is increasingly embraced for solving many machine

learning problems such as training the deep neural networks [Zhang *et al.*, 2016], matrix factorization and matrix completion. Thus, these nonconvex asynchronous algorithms are widely studied in recent years. For example, [Lian *et al.*, 2015] proposed an asynchronous distributed stochastic gradient method for the nonconvex optimization. [Huo and Huang, 2017] proposed an accelerated asynchronous distributed stochastic gradient method based on the SVRG method. More recently, [Zhu *et al.*, 2018; Yu and Huang, 2018] proposed the asynchronous distributed proximal stochastic gradient methods for the nonconvex problems with the nonsmooth regularization. However, these asynchronous proximal stochastic methods only deal with some simple regularization, so it is difficult to apply them to to many machine learning problems including multiple nonsmooth penalties or some complex regularization penalties, such as the overlapping group lasso [Mosci *et al.*, 2010] or graph-guided fused lasso [Kim *et al.*, 2009]. Thus, we propose a class of fast asynchronous distributed stochastic ADMM methods for these complex problems. Moreover, we study the convergence properties of the proposed methods. Table 1 summarizes the convergence properties of the proposed methods and other related ones.

2.1 Notations

Let $y_{[k]} = \{y_1, \cdots, y_k\}$ and $y_{[j:k]} = \{y_j, \cdots, y_k\}$ for $j \in [k]$. Given a positive definite matrix G, $\|x\|_G^2 = x^T G x$; $\sigma_{\max}(G)$ and $\sigma_{\min}(G)$ denote the largest and smallest eigenvalues of G, respectively; the conditional number $\kappa_G = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)}$. $\sigma^A_{\max}$ and $\sigma^A_{\min}$ denote the largest and smallest eigenvalues of matrix AA^T, respectively, and $\kappa_A = \frac{\sigma^A_{\max}}{\sigma^A_{\min}}$.

3 Preliminaries

In the section, we begin with restating the standard ϵ-approximate stationary point of the problem (3), as in [Jiang *et al.*, 2016; Zheng and Kwok, 2016b].

Definition 1. *Given $\epsilon > 0$, the point $(x^*, y^*_{[k]}, \lambda^*)$ is said to be an ϵ-approximate stationary point of the problems* (3)*, if it holds that*

$$\mathbb{E}\big[dist(0, \partial L(x^*, y^*_{[k]}, \lambda^*))^2\big] \leq \epsilon, \tag{4}$$

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

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

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

Next, we make some mild assumptions regarding problem (3) as follows:

Assumption 1. *Each function $f_i(x)$ is L-smooth for $\forall i \in \{1, 2, \cdots, n\}$ such that*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^d,$$

which is equivalent to

$$f_i(x) \leq f_i(y) + \nabla f_i(y)^T(x - y) + \frac{L}{2}\|x - y\|^2.$$

Assumption 2. *Gradient of each function $f_i(x)$ is bounded, i.e., there exists a constant $\delta > 0$ such that for all x, it follows that $\|\nabla f_i(x)\|^2 \leq \delta^2$.*

Assumption 3. *$f(x)$ and $h_j(y_j)$ for all $j \in [k]$ are all lower bounded, and denote $f^* = \inf_x f(x)$ and $h_j^* = \inf_y h_j(y)$ for $j \in [k]$.*

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

Assumption 5. *The maximum gradient delay is upper bounded for all time by a constant $\tau > 0$,* i.e.*, $t - D(t) \leq \tau$, where $D(t) \leq t$ denote the delayed time.*

Assumption 1 has been commonly used in the convergence analysis of nonconvex algorithms. Assumption 2 is widely used for stochastic gradient-based and ADMM-type methods [Boyd *et al.*, 2011]. Assumptions 3 and 4 have been used in the study of ADMM methods [Jiang *et al.*, 2016; Zheng and Kwok, 2016b]. Assumption 5 is standard in asynchronous distributed algorithms [Lian *et al.*, 2015]

4 Asynchronous Distributed Stochastic ADMMs

In this section, we study the asynchronous distributed stochastic ADMM methods, which focus on the consistent read setting, *i.e.*, the algorithm guarantees the atomic operation on the whole parameter x. Specifically, we consider a distributed system including a master and multiple workers. The master receives the gradients under a cyclic-delay architecture, and uses these gradients to update the parameters; While each worker w has a disjoint partition of all data with size n_w, and all workers work and communicate with master independently to access the parameters.

4.1 AsyDS-ADMM for Nonconvex Optimization

In this subsection, we propose an asynchronous distributed stochastic ADMM (AsyDS-ADMM) method to solve the problem (3).

We begin with defining an augmented Lagrangian function of the problem (3) as follows:

$$\mathcal{L}_\rho(x, y_{[k]}, \lambda) = f(x) + \sum_{j=1}^k h_j(y_j) - \langle \lambda, Ax + \sum_{j=1}^k B_j y_j - c\rangle + \frac{\rho}{2}\|Ax + \sum_{j=1}^k B_j y_j - c\|^2, \tag{5}$$

where λ denotes the dual variable; $\rho > 0$ denotes the penalty parameter. To update the parameter x, we introduce an approximated function over x_t as follows:

$$\hat{\mathcal{L}}_\rho\big(x, y^{t+1}_{[k]}, \lambda_t, v_t\big) = f(x_t) + v_t^T(x - x_t) + \frac{1}{2\eta}\|x - x_t\|_G^2 + \sum_{j=1}^k h_j(y_j^{t+1}) - \lambda_t^T(Ax + \sum_{j=1}^k B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2, \tag{6}$$

where v_t is a stochastic gradient over x_t and $\eta > 0$ is a step size. If the matrix A^TA is large, we can set $G = rI - \rho\eta A^TA \succ I$ with $r > \rho\eta\sigma_{\max}(A^TA) + 1$ to linearize the term $\|Ax + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2$.

Specifically, Algorithm 1 describes the algorithmic framework of AsyDS-ADMM. In Algorithm 1, since the workers asynchronously pull parameters and push gradients, the master may use the delayed stochastic gradients $v_t = \nabla f_{\mathcal{I}_t}(x_{D(t)}) = \sum_{i\in\mathcal{I}_t} \nabla f_i(x_{D(t)})$, where $D(t) \leq t$. To adopt the following proximal operator to update y:

$$y_j^{t+1} = \arg\min_{y_j\in\mathbb{R}^d} \frac{1}{2}\|y_j - y_j^t\|^2 + h_j(y_j), \tag{7}$$

we set $Q_j = \tau_j I - \rho B_j^T B_j \succ I$ with $\tau_j > \rho\sigma_{\max}(B_j^TB_j)+1$ to linearize the term $\|Ax_t + \sum_{j=1}^k B_j y_j - c\|^2$.

Algorithm 1 Nonconvex AsyDS-ADMM

1: **Input:** b, T, $\eta > 0$ and $\rho > 0$;
2: **Initialize:** x_1, y_j^1 for $j \in [k]$ and λ_1;
3: **for** $t = 1, 2, \cdots, T$ **do**
4: For each **worker**:
4.1) Retrieve $x_{D(t)}$ from master and uniformly sample a mini-batch $\mathcal{I}_t$ ($|\mathcal{I}_t| = b$) from $\{1, 2, \cdots, n_w\}$;
4.2) Compute $v_t = \nabla f_{\mathcal{I}_t}(x_{D(t)})$ and sent it to master;
5: For the **master**:
5.1) Receive stochastic gradient v_t from a specific worker;
5.2) $y_j^{t+1} = \arg\min_{y_j} \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_j, y_{[j+1:k]}^t, \lambda_t) + \frac{1}{2}\|y_j - y_j^t\|_{Q_j}^2$, for all $j \in [k]$;
5.3) $x_{t+1} = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[k]}^{t+1}, \lambda_t, v_t\big)$;
5.4) $\lambda_{t+1} = \lambda_t - \rho(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c)$;
6: **end for**
7: **Output:** Iterate $\{x, y_{[k]}, \lambda\}$ chosen uniformly random from $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$.

4.2 AsyDS-ADMM+ for Nonconvex Optimization

In this subsection, we propose an accelerated AsyDS-ADMM (AsyDS-ADMM+) by using variance reduced method of the SVRG.

In the above AsyDS-ADMM, we use a stochastic gradient v_t to approximate the full gradient $\nabla f(x_{D(t)})$. Thus, v_t has a large variance due to the random sampling similar to the SGD method [Bottou *et al.*, 2018]. In the paper, to deal with this large variance, we introduce the large mini-batch size b to guarantee the convergence of AsyDS-ADMM, which will be discussed in the below theoretical analysis. In fact, we need lots of time to calculate gradients, as in the deterministic ADMM method, which slows down the whole algorithm. In addition, we can also let the step size decrease gradually as learning proceeds, leading to slower convergence as in the stochastic ADMM [Ouyang *et al.*, 2013].

To achieve higher efficiency, we propose a faster AsyDS-ADMM+ by using the SVRG method. Algorithm 2 describes

Algorithm 2 Nonconvex AsyDS-ADMM+

1: **Input:** b, m, T, $S = [T/m]$, $\eta > 0$ and $\rho > 0$;
2: **Initialize:** $\tilde{x}^1 = x_0^1$, $y_j^{0,1}$ for $j \in [k]$ and λ_0^1;
3: **for** $s = 1, 2, \cdots, S$ **do**
4: For each **worker**:
Receive $\tilde{x}^s$ from master and calculate $\nabla f_w(\tilde{x}^s) = \frac{1}{n_w}\sum_{i\in D_w} \nabla f_i(\tilde{x}^s)$, and send it to master;
5: For the **master**:
Aggregate gradients from all workers and compute $\nabla f(\tilde{x}^s) = \sum_w \frac{n_w}{n}\nabla f_w(\tilde{x}^s)$, and sent it to workers;
6: **for** $t = 0, 1, \cdots, m-1$ **do**
7: For each **worker**:
7.1) Retrieve $x_{D(t)}^s$ from master and uniformly sample a mini-batch $\mathcal{I}_t$ ($|\mathcal{I}_t| = b$) from $\{1, 2, \cdots, n_w\}$;
7.2) Compute $\nabla f_{\mathcal{I}_t}(x_{D(t)}^s)$ and $\nabla f_{\mathcal{I}_t}(\tilde{x}^s)$ and sent $\nabla f_{\mathcal{I}_t}(x_{D(t)}^s) - \nabla f_{\mathcal{I}_t}(\tilde{x}^s)$ to master;
8: For the **master**:
8.1) Receive $\nabla f_{\mathcal{I}_t}(x_{D(t)}^s) - \nabla f_{\mathcal{I}_t}(\tilde{x}^s)$ from a specific worker, then calculate gradient $v_t^s = \nabla f_{\mathcal{I}_t}(x_{D(t)}^s) - \nabla f_{\mathcal{I}_t}(\tilde{x}^s) + \nabla f(\tilde{x}^s)$;
8.2) $y_j^{s,t+1} = \arg\min_{y_j} \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_j, y_{[j+1:k]}^{s,t}, \lambda_t^s) + \frac{1}{2}\|y_j - y_j^{s,t}\|_{Q_j}^2$, for all $j \in [k]$;
8.3) $x_{t+1}^s = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[k]}^{s,t+1}, \lambda_t^s, v_t^s\big)$;
8.4) $\lambda_{t+1}^s = \lambda_t^s - \rho(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c)$;
9: **end for**
10: $\tilde{x}^{s+1} = x_0^{s+1} = x_m^s$, $y_j^{s+1,0} = y_j^{s,m}$ for all $j \in [k]$, $\lambda_0^{s+1} = \lambda_m^s$;
11: **end for**
12: **Output:** Iterate $\{x, y_{[k]}, \lambda\}$ chosen uniformly random from $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$.

the algorithmic framework of AsyDS-ADMM+. Similar to the SVRG algorithm, the AsyDS-ADMM+ also has two-layer loops. In the s-th outer loop, the master broadcasts a snapshot $\tilde{x}^s$ to each worker, then each worker calculates the gradient $\nabla f_w(\tilde{x}^s)$ and sends it back to the master. After that, the master combines these gradient information to compute the full gradient $\nabla f(\tilde{x}^s)$. In the t-th inner loop, each worker uniformly and independently samples a mini-batch $\mathcal{I}_t$ with size b, then calculates the stochastic gradients, and sends it to the master. After that, the master uses these gradients to update the parameters $\{x, y_{[k]}, \lambda\}$.

5 Convergence Analysis

In this section, we will study the convergence properties of the proposed algorithms (*i.e.*, AsyDS-ADMM and AsyDS-ADMM+). For notational simplicity, let

$$\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(Q)\big),$$
$$\nu_2 = 6(1+\tau)L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}, \nu_3 = \frac{18(1+\tau)L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2},$$

and $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$.

5.1 Convergence Analysis of AsyDS-ADMM

In this subsection, we will study the convergence properties of the AsyDS-ADMM.

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

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\tfrac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\rho\eta^2} + \tfrac{9L^2}{\sigma_{\min}^A\rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2\big].$$

Let $\eta = \frac{\sigma_{\min}(G)\alpha}{(1+\tau^2)L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\alpha\sigma_{\min}^A}$, we have

$$\frac{1}{T}\sum_{t=1}^T\big(\|x_t - x_{t+1}\|^2 + \sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2\big) \leq \frac{\Omega_0 - \Omega^*}{\gamma T} + \frac{36\delta^2}{\gamma b\sigma_{\min}^A\rho} + \frac{\delta^2}{\gamma bL},$$

where $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi})$, $\tilde{\chi} \geq \frac{(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$ and Ω^* denotes a low bound of Ω_t.

Let $\theta_t = \|x_{t+1} - x_t\|^2 + \|x_t - x_{t-1}\|^2 + \sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1} - x_l\|^2 + \sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2$.

Theorem 1. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 1. Using the same conditions in Lemma 1, we have*

$$\min_{1\leq t\leq T}\mathbb{E}\big[dist(0, \partial L(x_t, y_{[k]}^t, \lambda_t))^2\big] \leq \frac{2(1+\tau)\nu_{\max}(\Omega_0 - \Omega^*)}{\gamma T} + \frac{36\delta^2}{\gamma b\sigma_{\min}^A\rho} + \frac{\delta^2}{\gamma bL} + \frac{6\delta^2}{b} + \frac{36\delta^2}{b\sigma_{\min}^A\rho^2}. \quad (8)$$

It implies that the iteration number T and mini-batch size b satisfy $T = O(\frac{1}{\epsilon})$ and $b = O(\frac{1}{\epsilon})$, then $(x_{t^*}, y_{[k]}^{t^*}, \lambda_{t^*})$ is an ϵ-approximate solution of (3)*, where $t^* = \arg\min_{1\leq t\leq T}\theta_t$.*

Remark 1. *Theorem 1 shows that given $\eta = \frac{\sigma_{\min}(G)\alpha}{(1+\tau^2)L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\alpha\sigma_{\min}^A}$, the AsyDS-ADMM has an $O(\frac{1}{T})$ convergence rate and the optimal IFO complexity $O(\epsilon^{-2})$ for finding an ϵ-approximate solution. In particular, we can choose $\alpha \in (0, 1]$ according to different problems to obtain an appropriate step-size η and penalty parameter ρ.*

5.2 Convergence Analysis of AsyDS-ADMM+

In the subsection, we will give the convergence analysis of the AsyDS-ADMM+.

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 2, and define a* Lyapunov *function:*

$$\Phi_t^s = \mathbb{E}\big[\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\tfrac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \tfrac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + \frac{18\kappa_A L^2}{\sigma_{\min}^A\rho b}\|x_{t-1}^s - \tilde{x}^s\|^2 + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 + c_t\|x_t^s - \tilde{x}^s\|^2\big],$$

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

$$c_t = \begin{cases}\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho b} + \frac{2L}{b} + (1+\beta)c_{t+1}, & 1 \leq t \leq m,\\ 0, & t \geq m+1.\end{cases}$$

Further let $b = [n^{\frac{2}{3}}]$, $m = [n^{\frac{1}{3}}]$, $\eta = \frac{\alpha\sigma_{\min}(G)}{(9+\tau^2)L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\sigma_{\min}^A\alpha}$, we have

$$\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^{m-1}\big(\sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{1}{b}\|x_t^s - \tilde{x}^s\|_2^2 + \|x_{t+1}^s - x_t^s\|^2\big) \leq \frac{\Phi_0^1 - \Phi^*}{\gamma T}, \quad (9)$$

where Φ^* denotes a lower bound of Φ_t^s; $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi}_t, L)$ and $\tilde{\chi}_t \geq \frac{(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$.

Let $\theta_t^s = \|x_{t+1}^s - x_t^s\|^2 + \|x_t^s - x_{t-1}^s\|^2 + \sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 + \frac{1}{b}(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2) + \sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2$.

Theorem 2. *Suppose the sequence $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ is generated from Algorithm 2. Using the same conditions in Lemma 2, we have*

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

It implies that if the whole number of iteration $T = mS$ satisfies

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

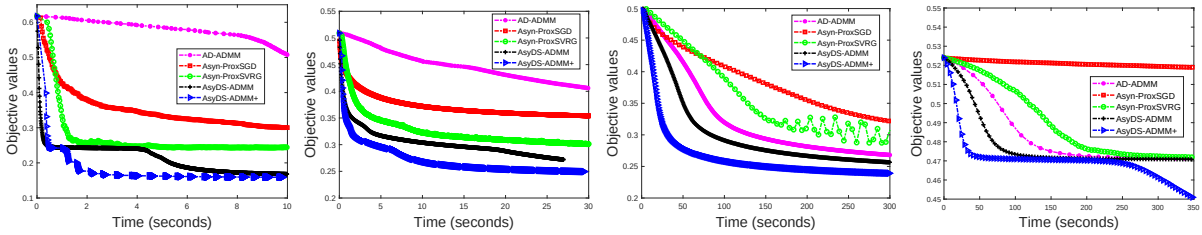
then $(x_{t^*}^{s^*}, y_{[k]}^{s^*,t^*}, \lambda_{t^*}^{s^*})$ is an ϵ-approximate solution of (3)*, where $(t^*, s^*) = \arg\min_{t,s}\theta_t^s$.*

Remark 2. *Theorem 2 states that given $\eta = \frac{\alpha\sigma_{\min}(G)}{(9+\tau^2)L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\sigma_{\min}^A\alpha}$, the AsyDS-ADMM+ has an $O(\frac{1}{T})$ convergence rate and the optimal IFO complexity $O(n + n^{\frac{2}{3}}\epsilon^{-1})$. In particular, set the delay number $\tau = 0$ in our theoretical analysis, we can obtain the optimal IFO complexity $O(n + n^{\frac{2}{3}}\epsilon^{-1})$ of the existing nonconvex SVRG-ADMM method [Zheng and Kwok, 2016b; Huang* et al.*, 2016].*

All related proofs are accessible in https://github.com/machine-learning-2019/ijcai-2019.

6 Experiments

In this section, we compare our algorithms (*i.e.*, AsyDS-ADMM and AsyDS-ADMM+) with some related asynchronous algorithms (*i.e.*, AD-ADMM [Chang *et al.*, 2016; Hong, 2018], Asyn-ProxSGD [Zhu *et al.*, 2018] and Asyn-ProxSVRG [Yu and Huang, 2018]) on the large-scale binary classification and multi-task learning. In the experiments, we run all distributed algorithms on 10 workers, and set the delay number $\tau = 5$.



(a) a9a (b) covtype.binary (c) SUSY (d) HIGGS

Figure 1: Objective values *versus* running time on binary classification.

Table 2: Real datasets for binary classification and multi-task.

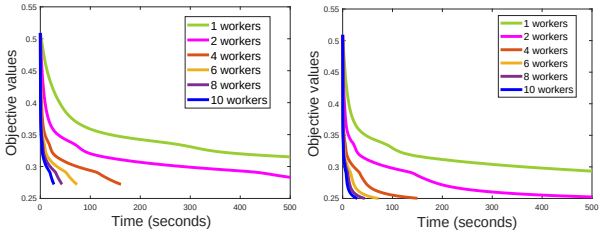
datasets	#samples	#features	#classes
a9a	32,561	123	2
covtype.binary	581,012	54	2
SUSY	5,000,000	18	2
HIGGS	11,000,000	28	2
covtype	581,012	54	7
mnist8m	8,100,000	784	10

6.1 Binary Classification

In this subsection, we apply our algorithms to solve the binary classification problem with graph-guided fused lasso regularization. Given a set of training samples $(a_i, b_i)_{i=1}^n$, where $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, 1\}$, then we solve the problem:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \beta_1\|\tilde{G}x\|_1 + \beta_2\|x\|_1, \qquad (10)$$

where $f_i(x) = \frac{1}{1+\exp(b_i a_i^T x)}$ is the nonconvex sigmoid loss function. Here $\tilde{G}$ decodes the sparsity pattern of graph, and can be obtained by sparse inverse covariance matrix estimation [Friedman *et al.*, 2008]. In the experiment, we set $\beta_1 = 10^{-4}$ and $\beta_2 = 10^{-5}$.



(a) AsyDS-ADMM (b) AsyDS-ADMM+

Figure 2: We run our algorithms on multiple workers from 1 to 10.

Figure 1 shows that objective values of our algorithms faster decrease than those of the other distributed algorithms, as the running time increases. Figure 2 shows that our algorithms have a near linear speedup, as number of workers increases.

6.2 Multi-task Learning

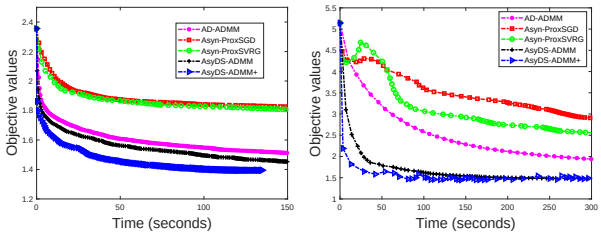
In this subsection, we apply our algorithms to multi-task learning with both sparse and low-rank penalties. Given a set of training samples $(a_i, b_i)_{i=1}^n$, where $a_i \in \mathbb{R}^d$ and $b_i \in \{1, 2, \cdots, c\}$, and let $W \in \mathbb{R}^{n\times c}$ with $W_{ij} = 1$ if $j = b_i$, and $W_{ij} = 0$ otherwise. Then this multi-task learning is equivalent to solving the following problem:

$$\min_{X\in\mathbb{R}^{c\times d}} \frac{1}{n}\sum_{i=1}^{n} f_i(X) + \beta_1 \sum_{ij} \kappa(|X_{ij}|) + \beta_2\|X\|_*,$$

where $f_i(X) = \log(\sum_{j=1}^c \exp(X_{j,.}a_i)) - \sum_{j=1}^c W_{ij}X_{j,.}a_i$ is a multinomial logistic loss function, $\kappa(x) = \beta\log(1 + \frac{x}{\alpha})$ is the nonconvex log-sum penalty function. Next, we change the above problem into the following form:

$$\min \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(X) + \beta_1\kappa_0\|X\|_1 + \beta_2\|X\|_*,$$

where $\kappa_0 = \kappa'(0)$ and $\hat{f}_i(X) = f_i(X) + \beta_1\big(\sum_{ij}\kappa(|X_{ij}|) - \kappa_0\|X\|_1\big)$, which is nonconvex and smooth. In the experiment, we set $\beta_1 = 10^{-4}$ and $\beta_2 = 10^{-5}$.



(a) covtype (b) mnist8m

Figure 3: Objective values *versus* running time on multi-task learning.

Figure 3 shows that objective values of our algorithms faster decrease than those of the other distributed algorithms, as the running time increases. These results demonstrate that our methods show better convergence behavior than the other methods in solving these complex problems.

7 Conclusions

In this paper, we proposed a class of fast asynchronous distributed stochastic ADMM methods (AsyDS-ADMM and AsyDS-ADMM+) for nonconvex nonsmooth optimization, and proved that both of them have convergence rate of $O(\frac{1}{T})$. Moreover, our theoretical analysis shows that the AsyDS-ADMM and AsyDS-ADMM+ also reach the optimal IFO complexities of $O(\epsilon^{-2})$ and $O(n + n^{2/3}\epsilon^{-1})$ for finding an ϵ-approximate solution, respectively. In particular, if setting the delay number $\tau = 0$ in our theoretical analysis, we can obtain the optimal IFO complexity of the existing nonconvex SVRG-ADMM method.

References

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

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

[Chang *et al.*, 2016] Tsung-Hui Chang, Mingyi Hong, Wei-Cheng Liao, and Xiangfeng Wang. Asynchronous distributed admm for large-scale optimization-part i: algorithm and convergence analysis. *IEEE Transactions on Signal Processing*, 64(12):3118–3130, 2016.

[Davis *et al.*, 2016] Damek Davis, Brent Edmunds, and Madeleine Udell. The sound of apalm clapping: faster nonsmooth nonconvex optimization with stochastic asynchronous palm. In *NIPS*, pages 226–234, 2016.

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

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

[Hong, 2018] Mingyi Hong. A distributed, asynchronous, and incremental algorithm for nonconvex optimization: An admm approach. *IEEE Transactions on Control of Network Systems*, 5(3):935–945, 2018.

[Huang *et al.*, 2016] Feihu Huang, Songcan Chen, and Zhaosong Lu. Stochastic alternating direction method of multipliers with variance reduction for nonconvex optimization. *arXiv preprint arXiv:1610.02758*, 2016.

[Huo and Huang, 2017] Zhouyuan Huo and Heng Huang. Asynchronous mini-batch gradient descent with variance reduction for non-convex optimization. In *AAAI*, 2017.

[Iutzeler *et al.*, 2013] Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *IEEE Annual Conference on Decision and Control*, pages 3671–3676, 2013.

[Jiang *et al.*, 2016] Bo Jiang, Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. Structured nonconvex and nonsmooth optimization: Algorithms and iteration complexity analysis. *arXiv preprint arXiv:1605.02408*, 2016.

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

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

[Lian *et al.*, 2015] Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *NIPS*, pages 2737–2745, 2015.

[Meng *et al.*, 2017] Qi Meng, Wei Chen, Jingcheng Yu, Taifeng Wang, Zhiming Ma, and Tie-Yan Liu. Asynchronous stochastic proximal optimization algorithms with variance reduction. In *AAAI*, pages 2329–2335, 2017.

[Mosci *et al.*, 2010] Sofia Mosci, Silvia Villa, Alessandro Verri, and Lorenzo Rosasco. A primal-dual algorithm for group sparse regularization with overlapping groups. In *NIPS*, pages 2604–2612, 2010.

[Ouyang *et al.*, 2013] Hua Ouyang, Niao He, Long Tran, and Alexander G Gray. Stochastic alternating direction method of multipliers. *ICML*, 28:80–88, 2013.

[Recht *et al.*, 2011] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, pages 693–701, 2011.

[Reddi *et al.*, 2015] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. In *NIPS*, pages 2647–2655, 2015.

[Reddi *et al.*, 2016] Sashank Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *NIPS*, pages 1145–1153, 2016.

[Wang *et al.*, 2015] Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block bregman admm for nonconvex composite problems. *arXiv preprint arXiv:1505.03063*, 2015.

[You *et al.*, 2016] Yang You, Xiangru Lian, Ji Liu, Hsiang-Fu Yu, Inderjit S Dhillon, James Demmel, and Cho-Jui Hsieh. Asynchronous parallel greedy coordinate descent. In *NIPS*, pages 4682–4690, 2016.

[Yu and Huang, 2018] Yue Yu and Longbo Huang. A new analysis of variance reduced stochastic proximal methods for composite optimization with serial and asynchronous realizations. *arXiv preprint arXiv:1805.10111*, 2018.

[Zhang and Kwok, 2014] Ruiliang Zhang and James Kwok. Asynchronous distributed admm for consensus optimization. In *ICML*, pages 1701–1709, 2014.

[Zhang *et al.*, 2016] Wei Zhang, Suyog Gupta, Xiangru Lian, and Ji Liu. Staleness-aware async-sgd for distributed deep learning. In *IJCAI*, pages 2350–2356, 2016.

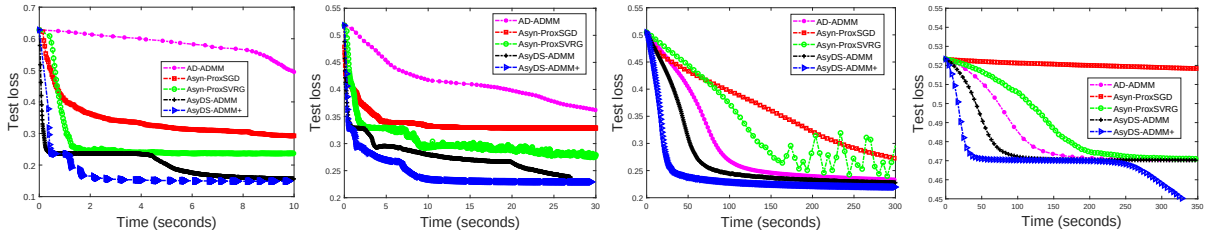
[Zheng and Kwok, 2016a] Shuai Zheng and James T Kwok. Fast and light stochastic admm. In *IJCAI*, 2016.

[Zheng and Kwok, 2016b] Shuai Zheng and James T Kwok. Stochastic variance-reduced admm. *arXiv preprint arXiv:1604.07070*, 2016.

[Zhu *et al.*, 2018] Rui Zhu, Di Niu, and Zongpeng Li. Asynchronous stochastic proximal methods for nonconvex nonsmooth optimization. *arXiv preprint arXiv:1802.08880*, 2018.

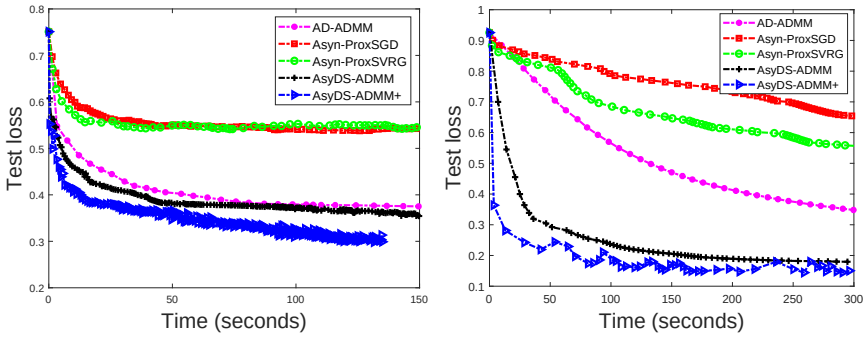
A Supplementary Materials

In this section, we first give some complementary results in the above experiments. Figures 4 and 5 show that test loss of our algorithms faster decrease than those of the other distributed algorithms, as the running time increases. From these result, we can find that our methods show better convergence behavior than the other methods in solving the complex nonconvex nonsmooth problems. In these experiments, we run all distributed algorithms on 10 workers, and set the delay number $\tau = 5$.



(a) a9a (b) covtype.binary (c) SUSY (d) HIGGS

Figure 4: Test loss *versus* running time on benchmark datasets in binary classification task.



(a) covtype (b) mnist8m

Figure 5: Test loss *versus* running time on multi-task learning.

Next, we at detail give the proof of the above lemmas and theorems.

Notations: To make the paper easier to follow, we give the following 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.
- $\|x\|_G = \sqrt{x^T G x}$, where G is a positive definite matrix.
- $\sigma^A_{\min}$ and $\sigma^A_{\max}$ denotes the minimum and maximum eigenvalues of $A^T A$, respectively; the conditional number $\kappa_A = \frac{\sigma^A_{\max}}{\sigma^A_{\min}}$.
- $\sigma^{B_j}_{\max}$ denotes the maximum eigenvalues of $B_j^T B_j$ for all $j \in [k]$, and $\sigma^B_{\max} = \max_{j=1}^k \sigma^{B_j}_{\max}$.
- $\sigma_{\min}(G)$ and $\sigma_{\max}(G)$ denote the minimum and maximum eigenvalues of matrix G, respectively; the conditional number $\kappa_G = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)}$.
- $\sigma_{\min}(Q_j)$ and $\sigma_{\max}(Q_j)$ denote the minimum and maximum eigenvalues of matrix Q_j for all $j \in [k]$, respectively; $\sigma_{\min}(Q) = \min_{j=1}^k \sigma_{\min}(Q_j)$ and $\sigma_{\max}(Q) = \max_{j=1}^k \sigma_{\max}(Q_j)$.
- η denotes the step size of updating variable x.
- L denotes the Lipschitz constant of $\nabla f(x)$.
- b denotes the mini-batch size of stochastic gradient.
- 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.

Lemma 3. *[Reddi* et al.*, 2016] For random variables $z_1, \cdots, z_r$ are independent and mean 0, we have*

$$\mathbb{E}[\|z_1 + \cdots + z_r\|^2] = \mathbb{E}[\|z_1\|^2 + \cdots + \|z_r\|^2]. \tag{11}$$

A.1 Theoretical Analysis of the AsyDS-ADMM

In this subsection, we in detail give the convergence analysis of the AsyDS-ADMM. We first give some useful lemmas as follows:

Lemma 4. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated by Algorithm 1, we have the following inequality:*

$$\begin{aligned}\mathbb{E}\|\lambda_{t+1}-\lambda_t\|^2 \leq & \frac{18L^2\tau}{\sigma_{\min}^A}\sum_{l=D(t)}^{t-1}\|x_{l+1}-x_l\|^2+\frac{18L^2\tau}{\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}-x_l\|^2 \\ & +(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2}+\frac{9L^2}{\sigma_{\min}^A})\|x_t-x_{t-1}\|^2+\frac{36\delta^2}{b\sigma_{\min}^A}. \end{aligned} \tag{12}$$

Proof. First, we define an unbiased stochastic gradient $u_t=\frac{1}{b}\sum_{i\in\mathcal{I}_t}\nabla f_i(x_t)$, then we give the upper bound of variance on the delayed stochastic $v_t=\frac{1}{b}\sum_{i\in\mathcal{I}_t}\nabla f_i(x_{D(t)})$ as follows:

$$\begin{aligned}\|v_t-\nabla f(x_t)\|^2 &= \|v_t-u_t+u_t-\nabla f(x_t)\|^2 \\ &\leq 2\|v_t-u_t\|^2+2\|u_t-\nabla f(x_t)\|^2 \\ &= 2\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}(\nabla f_i(x_{D(t)})-\nabla f_i(x_t))\|^2+2\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}\nabla f_i(x_t)-\nabla f(x_t)\|^2 \\ &\leq \frac{2}{b}\sum_{i\in\mathcal{I}_t}\|\nabla f_i(x_{D(t)})-\nabla f_i(x_t)\|^2+\frac{2}{b^2}\sum_{i\in\mathcal{I}_t}\|\nabla f_i(x_t)-\nabla f(x_t)\|^2 \\ &\leq 2L^2\|x_{D(t)}-x_t\|^2+\frac{2\delta^2}{b} \\ &= 2L^2\|\sum_{l=D(t)}^{t-1}(x_{l+1}-x_l)\|^2+\frac{2\delta^2}{b} \\ &\leq 2L^2\tau\sum_{l=D(t)}^{t-1}\|x_{l+1}-x_l\|^2+\frac{2\delta^2}{b}, \end{aligned} \tag{13}$$

where the second inequality holds by the inequality $\|\sum_{i=1}^r\alpha_i\|^2\leq r\sum_{i=1}^r\|\alpha_i\|$ and Lemma 3; the third inequality holds by Assumption 2.

Next, using the optimize condition of the the step 5.3 in Algorithm 1, we have

$$v_t+\frac{1}{\eta}G(x_{t+1}-x_t)-A^T\lambda_t+\rho A^T(Ax_{t+1}+\sum_{j=1}^k B_jy_j^{t+1}-c)=0. \tag{14}$$

Using the step 8 of Algorithm 1, then we have

$$A^T\lambda_{t+1}=v_t+\frac{G}{\eta}(x_{t+1}-x_t). \tag{15}$$

It follows that

$$A^T(\lambda_{t+1}-\lambda_t)=v_t-v_{t-1}+\frac{G}{\eta}(x_{t+1}-x_t)-\frac{1}{\eta}G(x_t-x_{t-1}). \tag{16}$$

By Assumption 4, we have

$$\|\lambda_{t+1}-\lambda_t\|^2\leq\frac{1}{\sigma_{\min}^A}\big[3\|v_t-v_{t-1}\|^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_{t+1}-x_t\|^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_t-x_{t-1}\|^2\big]. \tag{17}$$

Considering the upper bound of $\|v_t^s-v_{t-1}^s\|^2$, we have

$$\begin{aligned}\|v_t-v_{t-1}\|^2 &= \|v_t-\nabla f(x_t)+\nabla f(x_t)-\nabla f(x_{t-1})+\nabla f(x_{t-1})-v_{t-1}\|^2 \\ &\leq 3\|v_t-\nabla f(x_t)\|^2+3\|\nabla f(x_t)-\nabla f(x_{t-1})\|^2+3\|\nabla f(x_{t-1})-v_{t-1}\|^2 \\ &\leq 6L^2\tau\sum_{l=D(t)}^{t-1}\|x_{l+1}-x_l\|^2+6L^2\tau\sum_{l=D(t-1)}^{t-2}\|x_{l+1}-x_l\|^2+\frac{12\delta^2}{b}+3\|\nabla f(x_t)-\nabla f(x_{t-1})\|^2 \\ &\leq 6L^2\tau\big(\sum_{l=D(t)}^{t-1}\|x_{l+1}-x_l\|^2+\sum_{l=D(t-1)}^{t-2}\|x_{l+1}-x_l\|^2\big)+\frac{12\delta^2}{b}+3L^2\|x_t-x_{t-1}\|^2, \end{aligned} \tag{18}$$

where the second inequality holds by the above inequality (13), and the third inequality holds by Assumption 1. Finally, combining the inequalities (17) with (18), we can obtain the above result. □

Lemma 5. *Suppose the sequence* $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ *is generated from Algorithm 1, and define a* Lyapunov *function*

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2\big].$$

Let $\eta = \frac{\sigma_{\min}(G)\alpha}{(1+\tau^2)L}$ $(0 < \alpha \le 1)$ *and* $\rho = \frac{2(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\alpha\sigma_{\min}^A}$, *we have*

$$\frac{1}{T}\sum_{t=1}^T\big(\|x_t - x_{t+1}\|^2 + \sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2\big) \le \frac{\Omega_0 - \Omega^*}{\gamma T} + \frac{36\delta^2}{\gamma b\sigma_{\min}^A\rho} + \frac{\delta^2}{\gamma bL}, \tag{19}$$

where $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi})$, $\tilde{\chi} \ge \frac{(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$ *and* Ω^* *denotes a low bound of* Ω_t.

Proof. By the optimal condition of step 5.2 in Algorithm 1, we have, for $j \in [k]$

$$\begin{aligned}
0 &= (y_j^t - y_j^{t+1})^T\big(\partial h_j(y_j^{t+1}) - B^T\lambda_t + \rho B^T(Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c) + H_j(y_j^{t+1} - y_j^t)\big)\\
&\le h_j(y_j^t) - h_j(y_j^{t+1}) - (\lambda_t)^T(B_j y_j^t - B_j y_j^{t+1}) + \rho(By_j^t - By_j^{t+1})^T(Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c)\\
&\quad - \|y_j^{t+1} - y_j^t\|_{Q_j}^2\\
&= h_j(y_j^t) - h_j(y_j^{t+1}) - (\lambda_t)^T(Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^k B_i y_i^t - c) + (\lambda_t)^T(Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c)\\
&\quad + \frac{\rho}{2}\|Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^k B_i y_i^t - c\|^2 - \frac{\rho}{2}\|Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c\|^2 - \|y_j^{t+1} - y_j^t\|_{Q_j}^2\\
&\quad - \frac{\rho}{2}\|B_j y_j^t - B_j y_j^{t+1}\|^2\\
&\le \underbrace{f(x_t^s) + \sum_{l=1}^j h_j(y_j^{s,t}) + \sum_{l=j+1}^k h_j(y_j^{s,t+1}) - (\lambda_t^s)^T(Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^k B_i y_i^{s,t} - c) + \frac{\rho}{2}\|Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^k B_i y_i^{s,t} - c\|^2}_{\mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:k]}^{s,t}, \lambda_t^s)}\\
&\quad - \underbrace{\big(f(x_t^s) + \sum_{l=1}^{j-1} h_j(y_j^{s,t}) + \sum_{l=j}^k h_j(y_j^{s,t+1}) - (\lambda_t^s)^T(Ax_t^s + \sum_{i=1}^j B_i y_i^{s,t+1} + \sum_{i=j+1}^k B_i y_i^{s,t} - c) + \frac{\rho}{2}\|Ax_t^s + \sum_{i=1}^j B_i y_i^{s,t+1} + \sum_{i=j+1}^k B_i y_i^{s,t} - c\|^2\big)}_{\mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:k]}^{s,t}, \lambda_t^s)}\\
&\quad - \|y_j^{s,t+1} - y_j^{s,t}\|_{Q_j}^2 - \frac{\rho}{2}\|B_j y_j^{s,t} - B_j y_j^{s,t+1}\|^2\\
&\le \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:k]}^{s,t}, \lambda_t^s) - \mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:k]}^{s,t}, \lambda_t^s) - \sigma_{\min}(Q_j)\|y_j^{s,t} - y_j^{s,t+1}\|^2,
\end{aligned} \tag{20}$$

where the first inequality holds by the convexity of function $h_j(y)$, 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

$$\mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:k]}^t, \lambda_t) \le \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:k]}^t, \lambda_t) - \sigma_{\min}(Q_j)\|y_j^t - y_j^{t+1}\|^2, \text{ for } j \in [k]. \tag{21}$$

Telescoping inequality (21) over j from 1 to k, we obtain

$$\mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) \le \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) - \sigma_{\min}(Q)\sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2, \tag{22}$$

where $\sigma_{\min}(Q) = \min_{j\in[k]} \sigma_{\min}(Q_j)$.

By Assumption 1, we have

$$0 \le f(x_t) - f(x_{t+1}) + \nabla f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2. \tag{23}$$

Using the step 5.3 of Algorithm 1, we have

$$0 = (x_t - x_{t+1})^T\big(v_t - A^T\lambda_t + \rho A^T(Ax_{t+1} + \sum_{j=1}^{k} B_j y_j^{t+1} - c) + \frac{G}{\eta}(x_{t+1} - x_t)\big). \tag{24}$$

Combining (23) and (24), 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}{2}\|x_{t+1} - x_t\|^2 \\
& + (x_t - x_{t+1})^T\big(v_t - A^T\lambda_t + \rho A^T(Ax_{t+1} + \sum_{j=1}^{k} B_j y_j^{t+1} - c) + \frac{G}{\eta}(x_{t+1} - x_t)\big) \\
=\ & f(x_t) - f(x_{t+1}) + \frac{L}{2}\|x_t - x_{t+1}\|^2 - \frac{1}{\eta}\|x_t - x_{t+1}\|_G^2 + (x_t - x_{t+1})^T(v_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}{2}\|x_t - x_{t+1}\|^2 - \frac{1}{\eta}\|x_t - x_{t+1}\|_G^2 + (x_t - x_{t+1})^T(v_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} h_j(x_{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{\big(f(x_{t+1}) + \sum_{j=1}^{k} h_j(x_{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\big)}_{\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t)} \\
& + \frac{L}{2}\|x_t - x_{t+1}\|^2 + (x_t - x_{t+1})^T(v_t - \nabla f(x_t)) - \frac{1}{\eta}\|x_t - x_{t+1}\|_G^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}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L}{2})\|x_t - x_{t+1}\|^2 + (x_t - x_{t+1})^T(v_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}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 + \frac{1}{2L}\|v_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}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 + \tau L\sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 + \frac{\delta^2}{Lb},
\end{aligned} \tag{25}$$

where the equality (i) holds by applying the equality $(a-b)^Tb = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ 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 $a^Tb \le \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|a\|^2$, and the inequality (iii) holds by the inequality (13). Thus, we obtain

$$\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) \le \mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 + L\tau\sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 + \frac{\delta^2}{Lb}. \tag{26}$$

By the step 5.4 in Algorithm 1, we have

$$\begin{aligned}\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) - \mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) &= \frac{1}{\rho}\|\lambda_{t+1} - \lambda_t\|^2 \\ &\leq \frac{18L^2\tau}{\sigma_{\min}^A \rho} \sum_{l=D(t)}^{t-1} \|x_{l+1} - x_l\|^2 + \frac{18L^2\tau}{\sigma_{\min}^A \rho} \sum_{l=D(t-1)}^{t-2} \|x_{l+1} - x_l\|^2 \\ &\quad + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + \frac{36\delta^2}{b\sigma_{\min}^A \rho}. \end{aligned} \tag{27}$$

Combining (22), (26) and (27), we have

$$\begin{aligned}\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) &\leq \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 - \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2 \\ &\quad + \big(\frac{18L^2\tau}{\sigma_{\min}^A \rho} + L\tau\big) \sum_{l=D(t)}^{t-1} \|x_{l+1} - x_l\|^2 + \frac{18L^2\tau}{\sigma_{\min}^A \rho} \sum_{l=D(t-1)}^{t-2} \|x_{l+1} - x_l\|^2 \\ &\quad + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + \frac{36\delta^2}{\sigma_{\min}^A \rho b} + \frac{\delta^2}{Lb}. \end{aligned} \tag{28}$$

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

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A \rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A} \sum_{l=D(t-1)}^{t-2} \|x_{l+1}^s - x_l^s\|^2\big].$$

It follows that

$$\begin{aligned}\Omega_{t+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A \rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_{t+1} - x_t\|^2 + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A} \sum_{l=D(t)}^{t-1} \|x_{l+1}^s - x_l^s\|^2\big] \\ &\leq \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A \rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A} \sum_{l=D(t-1)}^{t-2} \|x_{l+1}^s - x_l^s\|^2 \\ &\quad - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t+1}\|^2 - \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2 \\ &\quad + \big(\frac{36L^2\tau}{\sigma_{\min}^A \rho} + L\tau\big) \sum_{l=D(t)}^{t-1} \|x_{l+1} - x_l\|^2 + \frac{36\delta^2}{b\sigma_{\min}^A \rho} + \frac{\delta^2}{Lb} \\ &= \Omega_t - \chi\|x_t - x_{t+1}\|^2 - \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2 + \big(\frac{36L^2\tau}{\sigma_{\min}^A \rho} + \tau L\big) \sum_{l=D(t)}^{t-1} \|x_{l+1} - x_l\|^2 + \frac{36\delta^2}{b\sigma_{\min}^A \rho} + \frac{\delta^2}{Lb}, \end{aligned} \tag{29}$$

where $\chi = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A \rho}$.

By (15), we have

$$\lambda_{t+1} = (A^T)^+(\hat{\nabla} f(x_t) + \frac{G}{\eta}(x_{t+1} - x_t)), \tag{30}$$

where $(A^T)^+$ is the pseudoinverse of A^T. Due to that A is full row rank, we have $(A^T)^+ = (AA^T)^{-1}A$. It follows that $\sigma_{\max}((A^T)^+)^T(A^T)^+) \leq \frac{\sigma_{\max}^A}{(\sigma_{\min}^A)^2} = \frac{\kappa_A}{\sigma_{\min}^A}$.

Then we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) = f(x_{t+1}) + \sum_{j=1}^k h_j(y_j^{t+1}) - \lambda_{t+1}^T(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&= f(x_{t+1}) + \sum_{j=1}^k h_j(y_j^{t+1}) - \langle (A^T)^+(v_t + \frac{G}{\eta}(x_{t+1} - x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&= f(x_{t+1}) + \sum_{j=1}^k h_j(y_j^{t+1}) - \langle (A^T)^+(v_t - \nabla f(x_t) + \nabla f(x_t) + \frac{G}{\eta}(x_{t+1} - x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle \\
&\quad + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&\geq f(x_{t+1}) + \sum_{j=1}^k h_j(y_j^{t+1}) - \frac{5\kappa_A}{2\sigma_{\min}^A \rho}\|v_t - \nabla f(x_t)\|^2 - \frac{5\kappa_A}{2\sigma_{\min}^A \rho}\|\nabla f(x_t)\|^2 - \frac{5\kappa_A \sigma_{\max}^2(G)}{2\sigma_{\min}^A \eta^2 \rho}\|x_{t+1} - x_t\|^2 \\
&\quad + \frac{\rho}{5}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&\geq f(x_{t+1}) + \sum_{j=1}^k h_j(y_j^{t+1}) - \frac{5L^2\tau\kappa_A}{\sigma_{\min}^A \rho}\sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 - \frac{5\kappa_A\delta^2}{b\sigma_{\min}^A\rho} - \frac{5\kappa_A\delta^2}{2\sigma_{\min}^A\rho} - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1} - x_t\|^2 \\
&= f(x_{t+1}) + \sum_{j=1}^k h_j(y_j^{t+1}) - \frac{5L^2\tau\kappa_A}{\sigma_{\min}^A \rho}\sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1} - x_t\|^2 - \frac{5(b+2)\kappa_A\delta^2}{2b\sigma_{\min}^A\rho},
\end{aligned} \tag{31}$$

where the first inequality is obtained by applying $\langle a, b\rangle \leq \frac{1}{2\beta}\|a\|^2 + \frac{\beta}{2}\|b\|^2$ to the terms $\langle (A^T)^+(\hat{\nabla} f(x_t) - \nabla f(x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$, $\langle (A^T)^+\nabla f(x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ and $\langle (A^T)^+\frac{G}{\eta}(x_{t+1} - x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ with $\beta = \frac{\rho}{5}$, respectively; the second inequality follows by Assumption 2. Using the definition of Ω_t and Assumption 4, we have

$$\Omega_{t+1} \geq f^* + \sum_{j=1}^k h_j^* - \frac{5(b+2)\kappa_A\delta^2}{2b\sigma_{\min}^A\rho},\ t = 0, 1, 2, \cdots. \tag{32}$$

It follows that the function Ω_t is bounded from below. Let Ω^* denotes a lower bound of Ω_t.

Telescoping inequality (29) over t from 0 to T, we have

$$\frac{1}{T}\sum_{t=1}^T \big(\chi\|x_t - x_{t+1}\|^2 + \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2\big) \leq \frac{\Omega_0 - \Omega^*}{T} + \big(\frac{36L^2\tau}{\sigma_{\min}^A\rho} + L\tau\big)\frac{1}{T}\sum_{t=1}^T\sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 + \frac{36\delta^2}{b\sigma_{\min}^A\rho} + \frac{\delta^2}{Lb},$$

where $\chi = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho}$. Since $\frac{1}{T}\sum_{t=1}^T\sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 \leq \tau\sum_{t=1}^T\|x_{t+1} - x_t\|^2$, we have

$$\frac{1}{T}\sum_{t=1}^T \big((\chi - \frac{36L^2\tau^2}{\sigma_{\min}^A\rho} - L\tau^2)\|x_t - x_{t+1}\|^2 + \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2\big) \leq \frac{\Omega_0 - \Omega^*}{T} + \frac{36\delta^2}{b\sigma_{\min}^A\rho} + \frac{\delta^2}{Lb}.$$

Next, considering the case of $\tilde{\chi} = \chi - \frac{36L^2\tau^2}{\sigma_{\min}^A\rho} - L\tau^2 > 0$, we have

$$\begin{aligned}
\tilde{\chi} &= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{36L^2\tau^2}{\sigma_{\min}^A\rho} - L\tau^2 \\
&= \underbrace{\frac{\sigma_{\min}(G)}{\eta} - L - L\tau^2}_{T_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{36L^2\tau^2}{\sigma_{\min}^A\rho}}_{T_2}
\end{aligned} \tag{33}$$

Let $0 < \eta \leq \frac{\sigma_{\min}(G)}{(1+\tau^2)L}$, we have $T_1 > 0$. Further let $\eta = \frac{\sigma_{\min}(G)\alpha}{(1+\tau^2)L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\alpha\sigma^A_{\min}}$, we have

$$\begin{aligned}
T_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\kappa_A\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{36L^2\tau^2}{\sigma^A_{\min}\rho} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{6(1+\tau^2)^2\kappa_A\kappa_G^2L^2}{\alpha^2\sigma^A_{\min}\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{36L^2\tau^2}{\sigma^A_{\min}\rho} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{6(1+\tau^2)^2\kappa_A\kappa_G^2L^2}{\alpha^2\sigma^A_{\min}\rho} - \frac{9\kappa_A\kappa_G^2L^2}{\alpha^2\sigma^A_{\min}\rho} - \frac{36\tau^2\kappa_A\kappa_G^2L^2}{\alpha^2\sigma^A_{\min}\rho} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{6(4+\tau^2)^2\kappa_A\kappa_G^2L^2}{\alpha^2\sigma^A_{\min}\rho} \\
&= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{6(4+\tau^2)^2\kappa_A\kappa_G^2L^2}{\alpha^2\sigma^A_{\min}\rho}}_{\geq 0} \\
&\geq \frac{(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha},
\end{aligned} \tag{34}$$

where $\kappa_G = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)} \geq 1$. Thus, we have $\tilde{\chi} \geq \frac{(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$. Finally, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\big(\|x_t - x_{t+1}\|^2 + \sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2\big) \leq \frac{\Omega_0 - \Omega^*}{\gamma T} + \frac{36\delta^2}{\gamma b\sigma^A_{\min}\rho} + \frac{\delta^2}{\gamma bL}, \tag{35}$$

where $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi})$ and $\tilde{\chi} \geq \frac{(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$.

□

Theorem 3. *Suppose the sequence $\{x_t, y^t_{[k]}, \lambda_t\}_{t=1}^T$ is generated from Algorithm 1. Let*

$$\begin{aligned}
&\nu_1 = k\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(Q)\big),\ \nu_2 = 6(1+\tau)L^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2} \\
&\nu_3 = \frac{18(1+\tau)L^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2},
\end{aligned}$$

and $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$. Further let $\eta = \frac{\sigma_{\min}(G)\alpha}{(1+\tau^2)L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\alpha\sigma^A_{\min}}$,

$$\min_{1\leq t\leq T}\mathbb{E}\big[dist(0, \partial L(x_t, y^t_{[k]}, \lambda_t))^2\big] \leq \frac{2(1+\tau)\nu_{\max}(\Omega_0 - \Omega^*)}{\gamma T} + \frac{36\delta^2}{\gamma b\sigma^A_{\min}\rho} + \frac{\delta^2}{\gamma bL} + \frac{6\delta^2}{b} + \frac{36\delta^2}{b\sigma^A_{\min}\rho^2}, \tag{36}$$

where $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi})$ and $\tilde{\chi} \geq \frac{(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$ and Ω^* is a lower bound of function Ω_t. It implies that the iteration number T and mini-batch size b satisfy

$$T = O(\frac{1}{\epsilon}),\ b = O(\frac{1}{\epsilon}),$$

then $(x_{t^*}, y^{t^*}_{[k]}, \lambda_{t^*})$ is an ϵ-approximate solution of (3)*, where $t^* = \arg\min_{1\leq t\leq T}\theta_t$.*

Proof. We begin with defining an useful sequence $\theta_t = \|x_{t+1} - x_t\|^2 + \|x_t - x_{t-1}\|^2 + \sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1} - x_l\|^2 + \sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2$. Summing the sequence θ_t over t from 1 to T, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\theta_t &= \frac{1}{T}\sum_{t=1}^{T}\big(\|x_{t+1} - x_t\|^2 + \|x_t - x_{t-1}\|^2 + \sum_{l=D(t)}^{t-1}\|x_{l+1} - x_l\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1} - x_l\|^2 + \sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2\big) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\big((1+\tau)\|x_{t+1}^s - x_t^s\|^2 + (1+\tau)\|x_t^s - x_{t-1}^s\|^2 + \sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2\big) \\
&\leq \frac{2(1+\tau)(\Omega_0 - \Omega^*)}{\gamma T},
\end{aligned} \tag{37}$$

where the first second inequality holds by the inequality $\frac{1}{T}\sum_{t=1}^{T}\sum_{l=D(t)}^{t-1}\|x_{l+1}-x_l\|^2 \le \tau\sum_{t=1}^{T}\|x_{t+1}-x_t\|^2$; the second inequality follows the above inequality (35).

By the optimal condition of the step 5.2 in Algorithm 1, we have, for all $i \in [k]$

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0,\partial_{y_j}L(x,y_{[k]},\lambda))^2\big]_{t+1} &= \mathbb{E}\big[\text{dist}(0,\partial\psi_j(y_j^{t+1})-B_j^T\lambda_{t+1})^2\big] \\
&= \|B_j^T\lambda_t-\rho B_j^T(Ax_t+\sum_{i=1}^{j}B_iy_i^{t+1}+\sum_{i=j+1}^{k}B_iy_i^t-c)-Q_j(y_j^{t+1}-y_j^t)-B_j^T\lambda_{t+1}\|^2 \\
&= \|\rho B_j^TA(x_{t+1}-x_t)+\rho B_j^T\sum_{i=j+1}^{k}B_i(y_i^{t+1}-y_i^t)-Q_j(y_j^{t+1}-y_j^t)\|^2 \\
&\le k\rho^2\sigma_{\max}^{B_j}\sigma_{\max}^{A}\|x_{t+1}-x_t\|^2+k\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^{k}\sigma_{\max}^{B_i}\|y_i^{t+1}-y_i^t\|^2 \\
&\quad +k\sigma_{\max}^2(Q_j)\|y_j^{t+1}-y_j^t\|^2 \\
&\le k\big(\rho^2\sigma_{\max}^{B}\sigma_{\max}^{A}+\rho^2(\sigma_{\max}^{B})^2+\sigma_{\max}^2(Q)\big)\theta_t, \qquad (38)
\end{aligned}$$

where the first inequality follows by the inequality $\|\sum_{i=1}^{r}\alpha_i\|^2 \le r\sum_{i=1}^{r}\|\alpha_i\|^2$.

By the step 5.3 in Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}[\text{dist}(0,\nabla_xL(x,y_{[k]},\lambda))]_{t+1} &= \mathbb{E}\|A^T\lambda_{t+1}-\nabla f(x_{t+1})\|^2 \\
&= \mathbb{E}\|v_t-\nabla f(x_{t+1})-\frac{G}{\eta}(x_t-x_{t+1})\|^2 \\
&= \mathbb{E}\|v_t-\nabla f(x_t)+\nabla f(x_t)-\nabla f(x_{t+1})-\frac{G}{\eta}(x_t-x_{t+1})\|^2 \\
&\le 3\mathbb{E}\|v_t-\nabla f(x_t)\|^2+3\|\nabla f(x_t)-\nabla f(x_{t+1})\|^2+3\|\frac{G}{\eta}(x_t-x_{t+1})\|^2 \\
&\le 6L^2\tau\sum_{l=D(t)}^{t-1}\|x_{l+1}-x_l\|^2+\frac{6\delta^2}{b}+3(L^2+\frac{\sigma_{\max}^2(G)}{\eta^2})\|x_t-x_{t+1}\|^2 \\
&\le \big(6(1+\tau)L^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\big)\theta_t+\frac{6\delta^2}{b}, \qquad (39)
\end{aligned}$$

where the second inequality follows the above inequality (13).

By the step 5.4 of Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}[\text{dist}(0,\nabla_\lambda L(x,y_{[k]},\lambda))]_{t+1} &= \mathbb{E}\|Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c\|^2 \\
&= \frac{1}{\rho^2}\mathbb{E}\|\lambda_{t+1}-\lambda_t\|^2 \\
&\le \frac{18L^2\tau}{\sigma_{\min}^A\rho^2}\sum_{l=D(t)}^{t-1}\|x_{l+1}-x_l\|^2+\frac{18L^2\tau}{\sigma_{\min}^A\rho^2}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}-x_l\|^2 \\
&\quad +(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}+\frac{9L^2}{\sigma_{\min}^A\rho^2})\|x_t-x_{t-1}\|^2+\frac{36\delta^2}{b\sigma_{\min}^A\rho^2} \\
&\le \big(\frac{18(1+\tau)L^2}{\sigma_{\min}^A\rho^2}+\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\big)\theta_t+\frac{36\delta^2}{b\sigma_{\min}^A\rho^2},
\end{aligned} \qquad (40)$$

where the first inequality holds by Lemma 3.

Let

$$\nu_1 = k\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(Q)\big),\ \nu_2 = 6(1+\tau)L^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}$$
$$\nu_3 = \frac{18(1+\tau)L^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2},$$

and $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$. Then combining the inequalities (38), (39) with (40), we have

$$\begin{aligned}\min_{1\le t\le T}\mathbb{E}\big[\text{dist}(0, \partial L(x_t, y^t_{[k]}, \lambda_t))^2\big] &\le \frac{\nu_{\max}}{T}\sum_{t=1}^{T}\theta_t + \frac{6\delta^2}{b} + \frac{36\delta^2}{b\sigma^A_{\min}\rho^2} \\ &\le \frac{2(1+\tau)\nu_{\max}(\Omega_0 - \Omega^*)}{\gamma T} + \frac{36\delta^2}{\gamma b\sigma^A_{\min}\rho} + \frac{\delta^2}{\gamma bL} + \frac{6\delta^2}{b} + \frac{36\delta^2}{b\sigma^A_{\min}\rho^2}, \end{aligned} \tag{41}$$

where $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi})$ and $\tilde{\chi} \ge \frac{(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$

Given $\eta = \frac{\sigma_{\min}(G)\alpha}{(1+\tau^2)L}$ $(0 < \alpha \le 1)$ and $\rho = \frac{2(4+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\alpha\sigma^A_{\min}}$, since k is relatively small, it is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(1)$, which are independent on n. Thus, we obtain

$$\min_{1\le t\le T}\mathbb{E}\big[\text{dist}(0, \partial L(x_t, y^t_{[k]}, \lambda_t))^2\big] \le O(\frac{1}{T}) + O(\frac{1}{b}). \tag{42}$$

□

A.2 Theoretical Analysis of the AsyDS-ADMM+

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

Lemma 6. *Suppose the sequence $\big\{(x^s_t, y^{s,t}_{[k]}, \lambda^s_t)^m_{t=1}\big\}^S_{s=1}$ is generated by Algorithm 2, we have the following inequality:*

$$\begin{aligned}\mathbb{E}\|\lambda^s_{t+1} - \lambda^s_t\|^2 \le{}& \frac{18L^2}{\sigma^A_{\min}b}\big(\|x^s_t - \tilde{x}^s\|^2 + \|x^s_{t-1} - \tilde{x}^s\|^2\big) + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2}\|x^s_{t+1} - x^s_t\|^2 \\ &+ (\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2} + \frac{9L^2}{\sigma^A_{\min}})\|x^s_t - x^s_{t-1}\|^2 + \frac{18L^2\tau}{\sigma^A_{\min}}(\sum_{l=D(t)}^{t-1}\|x^s_{l+1} - x^s_l\|^2 + \sum_{l=D(t-1)}^{t-2}\|x^s_{l+1} - x^s_l\|^2). \end{aligned} \tag{43}$$

Proof. First, we define an undelayed variance reduced stochastic gradient $u^s_t = \frac{1}{b}\sum_{i\in\mathcal{I}_t}(\nabla f_i(x^s_t) - \nabla f_i(\tilde{x}^s)) + \nabla f(\tilde{x}^s)$. Next, considering the upper bound on the variance of delayed gradient v^s_t, we have

$$\begin{aligned}\|v^s_t - \nabla f(x^s_t)\|^2 &= \|v^s_t - u^s_t + u^s_t - \nabla f(x^s_t)\|^2 \\ &\le 2\|v^s_t - u^s_t\|^2 + 2\|u^s_t - \nabla f(x^s_t)\|^2 \\ &= 2\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}(\nabla f_i(x^s_{D(t)}) - \nabla f_i(x^s_t))\|^2 + 2\|\frac{1}{b}\sum_{i\in\mathcal{I}_t}(\nabla f_i(x^s_t) - \nabla f_i(\tilde{x}^s)) + \nabla f(\tilde{x}^s) - \nabla f(x^s_t)\|^2 \\ &\le \frac{2}{b}\sum_{i\in\mathcal{I}_t}\|\nabla f_i(x^s_{D(t)}) - \nabla f_i(x^s_t)\|^2 + \frac{2}{b^2}\sum_{i\in\mathcal{I}_t}\|\nabla f_i(x^s_t) - \nabla f_i(\tilde{x}^s) + \nabla f(\tilde{x}^s) - \nabla f(x^s_t)\|^2 \\ &\le \frac{2}{b}\sum_{i\in\mathcal{I}_t}\|\nabla f_i(x^s_{D(t)}) - \nabla f_i(x^s_t)\|^2 + \frac{2}{b^2}\sum_{i\in\mathcal{I}_t}\|\nabla f_i(x^s_t) - \nabla f_i(\tilde{x}^s)\|^2 \\ &\le 2L^2\|x^s_{D(t)} - x^s_t\|^2 + \frac{2L^2}{b}\|x^s_t - \tilde{x}^s\|^2 \\ &= 2L^2\|\sum_{l=D(t)}^{t-1}(x^s_{l+1} - x^s_l)\|^2 + \frac{2L^2}{b}\|x^s_t - \tilde{x}^s\|^2 \\ &\le 2L^2\tau\sum_{l=D(t)}^{t-1}\|x^s_{l+1} - x^s_l\|^2 + \frac{2L^2}{b}\|x^s_t - \tilde{x}^s\|^2, \end{aligned} \tag{44}$$

where the second inequality holds by the inequality $\|\sum_{i=1}^n \alpha_i\|^2 \leq n\sum_{i=1}^n \|\alpha_i\|^2$ and Lemma 3; the third inequality follows the equality $\mathbb{E}\|\xi - \mathbb{E}[\xi]\|^2 = \mathbb{E}\|\xi\|^2 - \|\mathbb{E}[\xi]\|^2$; the forth inequality holds by Assumption 1.

Using the optimal condition for the step 8.3 of Algorithm 2, we have

$$v_t^s + \frac{1}{\eta}G(x_{t+1}^s - x_t^s) - A^T\lambda_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) = 0, \tag{45}$$

By the step 8.4 of Algorithm 2, we have

$$A^T\lambda_{t+1}^s = v_t^s + \frac{1}{\eta}G(x_{t+1}^s - x_t^s). \tag{46}$$

Since

$$A^T(\lambda_{t+1}^s - \lambda_t^s) = v_t^s - v_{t-1}^s + \frac{G}{\eta}(x_{t+1}^s - x_t^s) - \frac{G}{\eta}(x_t^s - x_{t-1}^s), \tag{47}$$

then we have

$$\|\lambda_{t+1}^s - \lambda_t^s\|^2 \leq \frac{1}{\sigma_{\min}^A}\big[3\|v_t^s - v_{t-1}^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_{t+1}^s - x_t^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_t^s - x_{t-1}^s\|^2\big]. \tag{48}$$

Next, considering the upper bound of $\|v_t^s - v_{t-1}^s\|^2$, we have

$$\begin{aligned}
\|v_t^s - v_{t-1}^s\|^2 &= \|v_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) - \nabla f(x_{t-1}^s) + \nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 \\
&\leq 3\|v_t^s - \nabla f(x_t^s)\|^2 + 3\|\nabla f(x_t^s) - \nabla f(x_{t-1}^s)\|^2 + 3\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 \\
&\leq 6L^2\tau\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{6L^2}{b}\|x_t^s - \tilde{x}^s\|^2 + 6L^2\tau\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 \\
&\quad + \frac{6L^2}{b}\|x_{t-1}^s - \tilde{x}^s\|^2 + 3L^2\|x_t^s - x_{t-1}^s\|^2,
\end{aligned} \tag{49}$$

where the second inequality holds by Assumption 1. Finally, combining (48) and (49), we obtain the above result. □

Lemma 7. *Suppose the sequence* $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ *is generated from Algorithm 2, and define a* Lyapunov *function:*

$$\begin{aligned}
\Phi_t^s =& \mathbb{E}\big[\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + \frac{18\kappa_A L^2}{\sigma_{\min}^A\rho b}\|x_{t-1}^s - \tilde{x}^s\|^2 \\
&+ \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 + c_t\|x_t^s - \tilde{x}^s\|^2\big],
\end{aligned} \tag{50}$$

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

$$c_t = \begin{cases} \dfrac{36\kappa_A L^2}{\sigma_{\min}^A\rho b} + \dfrac{2L}{b} + (1+\beta)c_{t+1},\ 1 \leq t \leq m, \\ 0,\ t \geq m+1. \end{cases}$$

Further let $b = [n^{\frac{2}{3}}]$, $m = [n^{\frac{1}{3}}]$, $\eta = \frac{\alpha\sigma_{\min}(G)}{(9+\tau^2)L}$ $(0 < \alpha \leq 1)$ *and* $\rho = \frac{2(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\sigma_{\min}^A\alpha}$, *we have*

$$\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^{m-1}\big(\sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{1}{b}\|x_t^s - \tilde{x}^s\|_2^2 + \|x_{t+1}^s - x_t^s\|^2\big) \leq \frac{\Phi_0^1 - \Phi^*}{\gamma T}, \tag{51}$$

where Φ^* *denotes a lower bound of* Φ_t^s; $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi}_t, L)$ *and* $\tilde{\chi}_t \geq \frac{(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$.

Proof. By the optimal condition of step 8.2 in Algorithm 2, we have, for $j \in [k]$

$$
\begin{aligned}
0 &= (y_j^{s,t} - y_j^{s,t+1})^T \big(\partial h_j(y_j^{s,t+1}) - B^T\lambda_t^s + \rho B^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) + Q_j(y_j^{s,t+1} - y_j^{s,t})\big) \\
&\le h_j(y_j^{s,t}) - h_j(y_j^{s,t+1}) - (\lambda_t^s)^T(B_j y_j^{s,t} - B_j y_j^{s,t+1}) + \rho(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) \\
&\quad - \|y_j^{s,t+1} - y_j^{s,t}\|_{Q_j}^2 \\
&= h_j(y_j^{s,t}) - h_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) \\
&\quad + \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}\|_{Q_j}^2 \\
&\quad - \frac{\rho}{2}\|B_j y_j^{s,t} - B_j y_j^{s,t+1}\|^2 \\
&\le \underbrace{f(x_t^s) + \sum_{l=1}^{j} h_j(y_j^{s,t}) + \sum_{l=j+1}^{k} h_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) + \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)} \\
&\quad - \underbrace{\big(f(x_t^s) + \sum_{l=1}^{j-1} h_j(y_j^{s,t}) + \sum_{l=j}^{k} h_j(y_j^{s,t+1}) - (\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\big)}_{\mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:k]}^{s,t}, \lambda_t^s)} \\
&\quad - \|y_j^{s,t+1} - y_j^{s,t}\|_{Q_j}^2 - \frac{\rho}{2}\|B_j y_j^{s,t} - B_j y_j^{s,t+1}\|^2 \\
&\le \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}(Q_j)\|y_j^{s,t} - y_j^{s,t+1}\|^2, \qquad (52)
\end{aligned}
$$

where the first inequality holds by the convexity of function $h_j(y)$, 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 $j \in [k]$

$$
\mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:k]}^{s,t}, \lambda_t^s) \le \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:k]}^{s,t}, \lambda_t^s) - \sigma_{\min}(Q_j)\|y_j^{s,t} - y_j^{s,t+1}\|^2. \qquad (53)
$$

Telescoping inequality (53) over j from 1 to k, we obtain

$$
\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) \le \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) - \sigma_{\min}(Q)\sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2, \qquad (54)
$$

where $\sigma_{\min}(Q) = \min_{j\in[k]} \sigma_{\min}(Q_j)$.

By Assumption 1, we have

$$
0 \le f(x_t^s) - f(x_{t+1}^s) + \nabla f(x_t^s)^T(x_{t+1}^s - x_t^s) + \frac{L}{2}\|x_{t+1}^s - x_t^s\|^2. \qquad (55)
$$

Using optimal condition of the step 8.3 in Algorithm 2, we have

$$
0 = (x_t^s - x_{t+1}^s)^T\big(v_t^s - A^T\lambda_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^{k} B_j y_j^{s,t+1} - c) + \frac{G}{\eta}(x_{t+1}^s - x_t^s)\big). \qquad (56)
$$

Combining (55) and (56), 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}{2}\|x_{t+1}^s - x_t^s\|^2 \\
& + (x_t^s - x_{t+1}^s)^T\big(v_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{G}{\eta}(x_{t+1}^s - x_t^s)\big) \\
=\ & f(x_t^s) - f(x_{t+1}^s) + \frac{L}{2}\|x_t^s - x_{t+1}^s\|^2 - \frac{1}{\eta}\|x_t^s - x_{t+1}^s\|_G^2 + (x_t^s - x_{t+1}^s)^T\big(v_t^s - \nabla f(x_t^s)\big) \\
& - (\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}{2}\|x_t^s - x_{t+1}^s\|^2 - \frac{1}{\eta}\|x_t^s - x_{t+1}^s\|_G^2 + (x_t^s - x_{t+1}^s)^T\big(v_t^s - \nabla f(x_t^s)\big) - (\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 h_j(x_t^s) - (\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{\big(f(x_{t+1}^s) + \sum_{j=1}^k h_j(x_{t+1}^s) - (\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\big)}_{\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s)} \\
& + \frac{L}{2}\|x_t^s - x_{t+1}^s\|^2 + (x_t^s - x_{t+1}^s)^T(v_t^s - \nabla f(x_t^s)) - \frac{1}{\eta}\|x_t^s - x_{t+1}^s\|_G^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}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L}{2})\|x_t^s - x_{t+1}^s\|^2 + (x_t^s - x_{t+1}^s)^T(v_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}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 + \frac{1}{2L}\|v_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}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 \\
& + \tau L \sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{L}{b}\|x_t^s - \tilde{x}^s\|^2,
\end{aligned} \tag{57}
$$

where the equality (i) holds by applying the equality $(a-b)^Tb = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ 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 $a^Tb \leq \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|b\|^2$, and the inequality (iii) follows the above inequality (44). Thus, we obtain

$$
\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 + \tau L \sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{L}{b}\|x_t^s - \tilde{x}^s\|^2. \tag{58}
$$

By the step 8.4 in Algorithm 2, we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) = \frac{1}{\rho}\|\lambda_{t+1}^s - \lambda_t^s\|^2 \\
&\leq \frac{18L^2}{\rho\sigma_{\min}^A b}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A \eta^2}\|x_{t+1}^s - x_t^s\|^2 \\
&\quad + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A \eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A})\|x_t^s - x_{t-1}^s\|^2 + \frac{18L^2\tau}{\rho\sigma_{\min}^A}(\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2). \qquad (59)
\end{aligned}$$

Combining (54), (58) and (59), we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) \leq \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) - \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^{s,t} - y_j^{s,t+1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 \\
&+ \big(\frac{18L^2}{\rho\sigma_{\min}^A b} + \frac{L}{b}\big)\|x_t^s - \tilde{x}^s\|^2 + \frac{18L^2}{\rho\sigma_{\min}^A b}\|x_{t-1}^s - \tilde{x}^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A \eta^2}\|x_{t+1}^s - x_t^s\|^2 + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A \eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A})\|x_t^s - x_{t-1}^s\|^2 \\
&+ \big(\frac{18\tau L^2}{\rho\sigma_{\min}^A} + \tau L\big)\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{18L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2. \qquad (60)
\end{aligned}$$

Next, we define a *Lyapunov* function Φ_t^s as follows:

$$\begin{aligned}
\Phi_t^s =& \mathbb{E}\big[\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + \frac{18\kappa_A L^2}{\sigma_{\min}^A\rho b}\|x_{t-1}^s - \tilde{x}^s\|^2 \\
&+ \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 + c_t\|x_t^s - \tilde{x}^s\|^2\big]. \qquad (61)
\end{aligned}$$

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

$$\begin{aligned}
\|x_{t+1}^s - x_t^s + x_t^s - \tilde{x}^s\|^2 &= \|x_{t+1}^s - x_t^s\|^2 + 2(x_{t+1}^s - x_t^s)^T(x_t^s - \tilde{x}^s) + \|x_t^s - \tilde{x}^s\|^2 \\
&\leq \|x_{t+1}^s - x_t^s\|^2 + 2\big(\frac{1}{2\beta}\|x_{t+1}^s - x_t^s\|^2 + \frac{\beta}{2}\|x_t^s - \tilde{x}^s\|^2\big) + \|x_t^s - \tilde{x}^s\|^2 \\
&= (1 + 1/\beta)\|x_{t+1}^s - x_t^s\|^2 + (1+\beta)\|x_t^s - \tilde{x}^s\|^2, \qquad (62)
\end{aligned}$$

where the above inequality holds by the Cauchy-Schwarz inequality with $\beta > 0$. Combining (61) with (62), we have

$$\begin{aligned}
\Phi_{t+1}^s &= \mathbb{E}\big[\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A \eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A})\|x_{t+1}^s - x_t^s\|^2 + \frac{18\kappa_A L^2}{\rho\sigma_{\min}^A b}\|x_t^s - \tilde{x}^s\|^2 \\
&\quad + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + c_{t+1}\|x_{t+1}^s - \tilde{x}^s\|^2\big] \\
&\leq \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A \eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A})\|x_t^s - x_{t-1}^s\|^2 + \frac{18\kappa_A L^2}{\sigma_{\min}^A b\rho}\|x_{t-1}^s - \tilde{x}^s\|^2 + \frac{18\kappa_A L^2\tau}{\rho\sigma_{\min}^A}\sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 \\
&\quad + \big(\frac{36\kappa_A L^2}{\sigma_{\min}^A b\rho} + \frac{2L}{b} + (1+\beta)c_{t+1}\big)\|x_t^s - \tilde{x}^s\|^2 - \sigma_{\min}(Q)\sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2 \\
&\quad - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+1/\beta)c_{t+1}\big)\|x_t^s - x_{t+1}^s\|^2 \\
&\quad - \frac{L}{b}\|x_t^s - \tilde{x}^s\|^2 + \big(\frac{36\kappa_A L^2\tau}{\rho\sigma_{\min}^A} + \tau L\big)\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 \\
&= \Phi_t^s - \sigma_{\min}(Q)\sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2 - \frac{L}{b}\|x_t^s - \tilde{x}^s\|^2 - \chi_t\|x_t^s - x_{t+1}^s\|^2 + \big(\frac{36\kappa_A L^2\tau}{\rho\sigma_{\min}^A} + \tau L\big)\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2, \\
&\qquad (63)
\end{aligned}$$

where $\kappa_A \geq 1$, $c_t = \frac{36\kappa_A L^2 d}{\sigma_{\min}^A b\rho} + \frac{2L}{b} + (1+\beta)c_{t+1}$ and $\chi_t = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+1/\beta)c_{t+1}$.

Next, we will give the relationship between Φ_1^{s+1} and Φ_m^s. Due to $x_0^{s+1} = x_m^s = \tilde{x}^{s+1}$, we have

$$v_0^{s+1} = \nabla f_{\mathcal{I}}(x_0^{s+1}) - \nabla f_{\mathcal{I}}(x_0^{s+1}) + \nabla f(x_0^{s+1}) = \nabla f(x_0^{s+1}) = \nabla f(x_m^s). \tag{64}$$

It follows that

$$\mathbb{E}\|v_0^{s+1} - v_m^s\|^2 = \mathbb{E}\|\nabla f(x_m^s) - v_m^s\|^2 \leq 2L^2\tau \sum_{l=D(m)}^{t-1} \|x_{l+1}^s - x_l^s\|^2 + \frac{2L^2}{b}\|x_m^s - \tilde{x}^s\|^2, \tag{65}$$

where the above inequality follows the inequality (44).

Following Lemma 6, we have

$$\begin{aligned}
\|\lambda_1^{s+1} - \lambda_m^s\|^2 &\leq \frac{1}{\sigma_{\min}^A}\|v_0^{s+1} - v_m^s + \frac{G}{\eta}(x_1^{s+1} - x_0^{s+1}) + \frac{G}{\eta}(x_m^s - x_{m-1}^s)\|^2 \\
&= \frac{1}{\sigma_{\min}^A}\|\nabla f(x_m^s) - v_m^s + \frac{G}{\eta}(x_1^{s+1} - x_m^s) + \frac{G}{\eta}(x_m^s - x_{m-1}^s)\|^2 \\
&\leq \frac{1}{\sigma_{\min}^A}\big(3\|\nabla f(x_m^s) - v_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_m^s - x_{m-1}^s\|^2\big) \\
&\leq \frac{6L^2\tau}{\sigma_{\min}^A}\sum_{l=D(m)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{6L^2}{\sigma_{\min}^A b}\|x_m^s - \tilde{x}^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2}\|x_m^s - x_{m-1}^s\|^2.
\end{aligned} \tag{66}$$

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 (54), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) &\leq \mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,0}, \lambda_0^{s+1}) - \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^{s+1,0} - y_j^{s+1,1}\|^2 \\
&= \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) - \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^{s,m} - y_j^{s+1,1}\|^2.
\end{aligned} \tag{67}$$

Since there is a synchronization at the beginning of each epoch, by (58), we have

$$\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) \leq \mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_0^{s+1} - x_1^{s+1}\|^2. \tag{68}$$

By (59), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) &\leq \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) + \frac{1}{\rho}\|\lambda_1^{s+1} - \lambda_0^{s+1}\|^2 \\
&\leq \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) + \frac{6L^2\tau}{\sigma_{\min}^A\rho}\sum_{l=D(m)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{6L^2}{\sigma_{\min}^A b\rho}\|x_m^s - \tilde{x}^s\|^2 \\
&\quad + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_m^s - x_{m-1}^s\|^2.
\end{aligned} \tag{69}$$

where the second inequality holds by (66).

Combining (67), (68) with (69), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) \leq \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) &- \sigma_{\min}(Q)\sum_{j=1}^k \|y_j^{s,m} - y_j^{s+1,1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_0^{s+1} - x_1^{s+1}\|^2 \\
&+ \frac{6L^2\tau}{\sigma_{\min}^A\rho}\sum_{l=D(m)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{6L^2}{\sigma_{\min}^A b\rho}\|x_m^s - \tilde{x}^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_1^{s+1} - x_m^s\|^2 \\
&+ \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_m^s - x_{m-1}^s\|^2.
\end{aligned} \tag{70}$$

It follows that

$$\begin{aligned}
\Phi_1^{s+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_1^{s+1} - x_0^{s+1}\|^2 + \frac{18\kappa_A L^2}{\rho\sigma_{\min}^A b}\|x_0^{s+1} - \tilde{x}^{s+1}\|^2 + c_1\|x_1^{s+1} - \tilde{x}^{s+1}\|^2\big] \\
&= \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) + \big(\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho} + c_1\big)\|x_1^{s+1} - x_m^s\|^2 \\
&\le \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_m^s - x_{m-1}^s\|^2 + \frac{18\kappa_A L^2}{\sigma_{\min}^A\rho b}\|x_{m-1}^s - \tilde{x}^s\|_2^2 + \frac{18\kappa_A L^2\tau}{\sigma_{\min}^A\rho}\sum_{l=D(m)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 \\
&\quad + (\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho b} + \frac{2L}{b})\|x_m^s - \tilde{x}^s\|_2^2 - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - c_1\big)\|x_1^{s+1} - x_m^s\|_2^2 \\
&\quad - \sigma_{\min}(Q)\sum_{j=1}^k\|y_j^{s,m} - y_j^{s+1,1}\|^2 - \frac{L}{b}\|x_m^s - \tilde{x}^s\|_2^2 \\
&= \Phi_m^s - \sigma_{\min}(Q)\sum_{j=1}^k\|y_j^{s,m} - y_j^{s+1,1}\|^2 - \frac{L}{b}\|x_m^s - \tilde{x}^s\|_2^2 - \chi_m\|x_1^{s+1} - x_m^s\|^2,
\end{aligned} \tag{71}$$

where $\kappa_A \ge 1$, $c_m = \frac{36\kappa_A L^2}{\sigma_{\min}^A\rho b} + \frac{2L}{b}$ and $\chi_m = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - c_1$.

By (46), we have

$$\lambda_{t+1} = (A^T)^+(\hat{\nabla}f(x_t) + \frac{G}{\eta}(x_{t+1} - x_t)), \tag{72}$$

where $(A^T)^+$ is the pseudoinverse of A^T. Due to that A is full row rank, we have $(A^T)^+ = (AA^T)^{-1}A$. It follows that $\sigma_{\max}((A^T)^+)^T(A^T)^+) \le \frac{\sigma_{\max}^A}{(\sigma_{\min}^A)^2} = \frac{\kappa_A}{\sigma_{\min}^A}$.

Then we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) = f(x_{t+1}^s) + \sum_{j=1}^k h_j(y_j^{s,t+1}) - \lambda_{t+1}^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&= f(x_{t+1}^s) + \sum_{j=1}^k h_j(y_j^{s,t+1}) - \langle(A^T)^+(v_t^s + \frac{G}{\eta}(x_{t+1}^s - x_t^s)), Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\rangle + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&= f(x_{t+1}^s) + \sum_{j=1}^k h_j(y_j^{s,t+1}) - \langle(A^T)^+(v_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) + \frac{G}{\eta}(x_{t+1}^s - x_t^s)), Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\rangle \\
&\quad + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&\ge f(x_{t+1}^s) + \sum_{j=1}^k h_j(y_j^{s,t+1}) - \frac{5\kappa_A}{2\sigma_{\min}^A\rho}\|v_t^s - \nabla f(x_t^s)\|^2 - \frac{5\kappa_A}{2\sigma_{\min}^A\rho}\|\nabla f(x_t^s)\|^2 - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2 \\
&\quad + \frac{\rho}{5}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&\ge f(x_{t+1}^s) + \sum_{j=1}^k h_j(y_j^{s,t+1}) - \frac{5\kappa_A L^2\tau}{\sigma_{\min}^A\rho}\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 - \frac{5\kappa_A L^2}{\sigma_{\min}^A\rho b}\|x_t^s - \tilde{x}^s\|^2 - \frac{5\kappa_A\delta^2}{2\sigma_{\min}^A\rho} - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2,
\end{aligned} \tag{73}$$

where the first inequality is obtained by applying $\langle a, b\rangle \le \frac{1}{2\beta}\|a\|^2 + \frac{\beta}{2}\|b\|^2$ to the terms $\langle(A^T)^+(\hat{\nabla}f(x_t) - \nabla f(x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$, $\langle(A^T)^+\nabla f(x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ and $\langle(A^T)^+\frac{G}{\eta}(x_{t+1} - x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ with

$\beta = \frac{\rho}{5}$, respectively; the second inequality follows by Assumption 2. Using the definition of Φ_t^s and Assumption 4, we have

$$\Phi_{t+1}^s \geq f^* + \sum_{j=1}^{k} h_j^* - \frac{5\kappa_A \delta^2}{2\sigma_{\min}^A \rho},\ t = 0, 1, 2, \cdots. \tag{74}$$

It follows that the function Φ_t^s is bounded from below. Let Φ^* denotes a lower bound of Φ_t^s.

Telescoping (63) and (71) over t from 0 to $m-1$ and over s from 1 to S, we have

$$\begin{aligned}\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\big(\sigma_{\min}(Q)\sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|_2^2 + \chi_t\|x_{t+1}^s - x_t^s\|^2\big) \\ \leq \frac{\Phi_0^1 - \Phi^*}{T} + \big(\frac{36\kappa_A L^2\tau}{\rho\sigma_{\min}^A} + \tau L\big)\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2,\end{aligned} \tag{75}$$

where $T = mS$. Since $\sum_{t=0}^{m-1}\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 \leq \tau\sum_{t=0}^{m-1}\|x_{t+1}^s - x_t^s\|^2$, we have

$$\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\big(\sigma_{\min}(Q)\sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{L}{b}\|x_t^s - \tilde{x}^s\|_2^2 + (\chi_t - \frac{36\kappa_A L^2\tau^2}{\rho\sigma_{\min}^A} - \tau^2 L)\|x_{t+1}^s - x_t^s\|^2\big) \leq \frac{\Phi_0^1 - \Phi^*}{T}. \tag{76}$$

Next, we consider the case of $\tilde{\chi}_t = \chi_t - \frac{36\kappa_A L^2\tau^2}{\rho\sigma_{\min}^A} - \tau^2 L = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+1/\beta)c_{t+1} - \frac{36\kappa_A L^2\tau^2}{\rho\sigma_{\min}^A} - \tau^2 L \geq 0$. Let $c_{m+1} = 0$ and $\beta = \frac{1}{m}$, recursing on t, we have

$$\begin{aligned}c_{t+1} = (\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho b} + \frac{2L}{b})\frac{(1+\beta)^{m-t} - 1}{\beta} &= \frac{m}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L)\big((1+\frac{1}{m})^{m-t} - 1\big) \\ &\leq \frac{m}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L)(e-1) \leq \frac{2m}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L).\end{aligned} \tag{77}$$

where the first 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 $t = 1, 2, \cdots, m$

$$\begin{aligned}\tilde{\chi}_t &\geq \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \tau^2 L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{36\kappa_A L^2\tau^2}{\rho\sigma_{\min}^A} - (1+1/\beta)\frac{2m}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L) \\ &= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \tau^2 L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{36\kappa_A L^2\tau^2}{\rho\sigma_{\min}^A} - (1+m)\frac{2m}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L) \\ &\geq \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \tau^2 L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{36\kappa_A L^2\tau^2}{\rho\sigma_{\min}^A} - \frac{4m^2}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L) \\ &= \underbrace{\frac{\sigma_{\min}(G)}{\eta} - L - \tau^2 L - \frac{8m^2 L}{b}}_{T_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{36\kappa_A L^2\tau^2}{\rho\sigma_{\min}^A} - \frac{144m^2\kappa_A L^2}{b\sigma_{\min}^A\rho}}_{T_2}.\end{aligned} \tag{78}$$

Let $m = [n^{\frac{1}{3}}]$, $b = [n^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{(9+\tau^2)L}$, we have $T_1 \geq 0$. Further, let $\eta = \frac{\alpha\sigma_{\min}(G)}{(9+\tau^2)L}$ $(0 < \alpha \leq 1)$ and

$\rho = \frac{2(9+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\sigma^A_{\min}\alpha}$, we have

$$\begin{aligned}
T_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{36\kappa_A L^2\tau^2}{\rho\sigma^A_{\min}} - \frac{144m^2\kappa_A L^2}{b\sigma^A_{\min}\rho} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{6(9+\tau^2)^2\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{36\kappa_A L^2\tau^2}{\rho\sigma^A_{\min}} - \frac{144\kappa_A L^2}{\sigma^A_{\min}\rho} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{6(9+\tau^2)^2\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{(144+36\tau^2)\kappa_A L^2}{\rho\sigma^A_{\min}\alpha^2} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{6(12+\tau^2)^2\kappa_A\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} \\
&= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{6(12+\tau^2)^2\kappa_A\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\
&\geq \frac{(9+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha},
\end{aligned} \tag{79}$$

where $\kappa_G \geq 1$. Thus, we have $\tilde{\chi}_t \geq \frac{(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$ for all $t \in \{1, 2, \cdots, m\}$.

Thus, by (76), we have

$$\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\big(\sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{1}{b}\|x_t^s - \tilde{x}^s\|_2^2 + \|x_{t+1}^s - x_t^s\|^2\big) \leq \frac{\Phi_0^1 - \Phi^*}{\gamma T}, \tag{80}$$

where $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi}_t, L)$ and $\tilde{\chi}_t \geq \frac{(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$.

□

Next, based on the above lemmas, we give the convergence analysis of AsyDS-ADMM+ in the following:

Theorem 4. *Suppose the sequence $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ is generated from Algorithm 2. Given*

$$\begin{aligned}
&\nu_1 = k\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(Q)\big),\ \nu_2 = 6(1+\tau)L^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2} \\
&\nu_3 = \frac{18(1+\tau)L^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2},
\end{aligned}$$

and let $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$. Further, let $m = [n^{\frac{1}{3}}]$, $b = [n^{\frac{2}{3}}]$, $\eta = \frac{\alpha\sigma_{\min}(G)}{(9+\tau^2)L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\sigma^A_{\min}\alpha}$, then we have

$$\min_{1\leq s\leq S,\ 0\leq t\leq m-1}\mathbb{E}\big[dist(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \leq \frac{\nu_{\max}}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\theta_t^s \leq \frac{2(1+\tau)\nu_{\max}(\Phi_0^1 - \Phi^*)}{\gamma T} = O(\frac{1}{T}). \tag{81}$$

It implies that if the whole number of iteration $T = mS$ satisfy

$$T = \frac{4\nu_{\max}(\Phi_0^1 - \Phi^*)}{\epsilon\gamma} \tag{82}$$

where Φ^* denotes a lower bound of Φ_t^s; $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi}_t, L)$ and $\tilde{\chi}_t \geq \frac{(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$, then $(x_{t^*}^{s^*}, y_{[k]}^{s^*,t^*}, \lambda_{t^*}^{s^*})$ is an ϵ-approximate solution of (3)*, where $(t^*, s^*) = \arg\min_{t,s}\theta_t^s$.*

Proof. First, we define an useful sequence $\theta_t^s = \|x_{t+1}^s - x_t^s\|^2 + \|x_t^s - x_{t-1}^s\|^2 + \sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 + \frac{1}{b}(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2) + \sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2$. Summing the sequence θ_t^s over t from 0 to $m-1$ and over

s from 1 to S, we have

$$\begin{aligned}\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\theta_t^s &= \frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\big(\|x_{t+1}^s - x_t^s\|^2 + \|x_t^s - x_{t-1}^s\|^2 + \sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2 + \frac{1}{b}(\|x_t^s - \tilde{x}^s\|^2 \\ &\quad + \|x_{t-1}^s - \tilde{x}^s\|^2) + \sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2\big) \\ &\le \frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\big((1+\tau)\|x_{t+1}^s - x_t^s\|^2 + (1+\tau)\|x_t^s - x_{t-1}^s\|^2 + \frac{1}{b}(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2) + \sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2\big) \\ &\le \frac{2(1+\tau)(\Phi_0^1 - \Phi^*)}{\gamma T}, \end{aligned} \tag{83}$$

where the first inequality holds by the inequality $\sum_{t=0}^{m-1}\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 \le \tau\sum_{t=0}^{m-1}\|x_{t+1}^s - x_t^s\|^2$; the second inequality holds by the inequality (80); $\gamma = \min(\sigma_{\min}(Q), \tilde{\chi}_t, L)$ and $\tilde{\chi}_t \ge \frac{(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{2\alpha} > 0$.

By the step 8.2 of Algorithm 2, we have, for all $i \in [k]$

$$\begin{aligned}\mathbb{E}\big[\text{dist}(0, \partial_{y_j} L(x, y_{[k]}, \lambda))^2\big]_{s,t+1} &= \mathbb{E}\big[\text{dist}(0, \partial h_j(y_j^{s,t+1}) - B_j^T\lambda_{t+1}^s)^2\big] \\ &= \|B_j^T\lambda_t^s - \rho B_j^T(Ax_t^s + \sum_{i=1}^{j}B_i y_i^{s,t+1} + \sum_{i=j+1}^{k}B_i y_i^{s,t} - c) - Q_j(y_j^{s,t+1} - y_j^{s,t}) - B_j^T\lambda_{t+1}^s\|^2 \\ &= \|\rho B_j^T A(x_{t+1}^s - x_t^s) + \rho B_j^T\sum_{i=j+1}^{k}B_i(y_i^{s,t+1} - y_i^{s,t}) - Q_j(y_j^{s,t+1} - y_j^{s,t})\|^2 \\ &\le k\rho^2\sigma_{\max}^{B_j}\sigma_{\max}^{A}\|x_{t+1}^s - x_t^s\|^2 + k\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^{k}\sigma_{\max}^{B_i}\|y_i^{s,t+1} - y_i^{s,t}\|^2 \\ &\quad + k\sigma_{\max}^2(Q_j)\|y_j^{s,t+1} - y_j^{s,t}\|^2 \\ &\le k\big(\rho^2\sigma_{\max}^{B}\sigma_{\max}^{A} + \rho^2(\sigma_{\max}^{B})^2 + \sigma_{\max}^2(Q)\big)\theta_t^s, \end{aligned} \tag{84}$$

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

By the step 8.3 of Algorithm 2, we have

$$\begin{aligned}\mathbb{E}[\text{dist}(0, \nabla_x L(x, y_{[k]}, \lambda))]_{s,t+1} &= \mathbb{E}\|A^T\lambda_{t+1}^s - \nabla f(x_{t+1}^s)\|^2 \\ &= \mathbb{E}\|v_t^s - \nabla f(x_{t+1}^s) - \frac{G}{\eta}(x_t^s - x_{t+1}^s)\|^2 \\ &= \mathbb{E}\|v_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) - \nabla f(x_{t+1}^s) - \frac{G}{\eta}(x_t^s - x_{t+1}^s)\|^2 \\ &\le 6L^2\tau\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \frac{6L^2}{b}\|x_t^s - \tilde{x}^s\|^2 + 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\|x_t^s - x_{t+1}^s\|^2 \\ &\le \big(6(1+\tau)L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\big)\theta_t^s, \end{aligned} \tag{85}$$

where the first inequality holds by the above inequality (44).

By the step 8.4 of Algorithm 2, we have

$$\begin{aligned}\mathbb{E}[\text{dist}(0,\nabla_\lambda L(x,y_{[k]},\lambda))]_{s,t+1} &= \mathbb{E}\|Ax_{t+1}^s + By_{t+1}^s - c\|^2 \\ &= \frac{1}{\rho^2}\mathbb{E}\|\lambda_{t+1}^s - \lambda_t^s\|^2 \\ &\le \frac{18L^2}{\sigma_{\min}^A b\rho^2}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\|x_{t+1}^s - x_t^s\|^2 \\ &\quad + \big(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2} + \frac{9L^2}{\sigma_{\min}^A\rho^2}\big)\|x_t^s - x_{t-1}^s\|^2 + \frac{18L^2\tau}{\sigma_{\min}^A\rho^2}\big(\sum_{l=D(t)}^{t-1}\|x_{l+1}^s - x_l^s\|^2 + \sum_{l=D(t-1)}^{t-2}\|x_{l+1}^s - x_l^s\|^2\big) \\ &\le \big(\frac{18(1+\tau)L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\big)\theta_t^s,\end{aligned} \tag{86}$$

where the first inequality follows Lemma 6.

Let

$$\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(Q)\big),\ \nu_2 = 6(1+\tau)L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}$$
$$\nu_3 = \frac{18(1+\tau)L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2},$$

and let $\nu_{\max} = \max(\nu_1,\nu_2,\nu_3)$. Further, given $\eta = \frac{\alpha\sigma_{\min}(G)}{(9+\tau^2)L}$ $(0<\alpha\le 1)$ and $\rho = \frac{2(12+\tau^2)\sqrt{6\kappa_A}\kappa_G L}{\sigma_{\min}^A\alpha}$, it is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(1)$. Thus, we obtain

$$\min_{1\le s\le S,\ 0\le t\le m-1}\mathbb{E}\big[\text{dist}(0,\partial L(x_t^s,y_{[k]}^{s,t},\lambda_t^s))^2\big] \le \frac{\nu_{\max}}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\theta_t^s \le \frac{2(1+\tau)\nu_{\max}(\Phi_0^1 - \Phi^*)}{\gamma T} = O(\frac{1}{T}). \tag{87}$$

□