Supplementary Materials for " Faster Stochastic Alternating Direction Method of Multipliers for Nonconvex Optimization"

Feihu Huang [1] **Songcan Chen** [2 3] **Heng Huang** [1 4]

1. Preliminaries

In this section, we give some preliminaries for the following theoretical analysis.

In this paper, we focus on the following nonconvex nonsmooth problem:

$$\min_{x,\{y_j\}_{j=1}^m} f(x) := \begin{cases} \frac{1}{n}\sum_{i=1}^n f_i(x) \text{ (finite-sum)} \\ \mathbb{E}_\zeta[f(x,\zeta)] \text{ (online)} \end{cases} + \sum_{j=1}^m g_j(y_j)$$
$$\text{s.t. } Ax + \sum_{j=1}^m B_j y_j = c, \quad (1)$$

Next, we give some standard assumptions regarding problem (1) as follows:

Assumption 1. *Each loss function $f_i(x)$ is L-smooth such that*

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

which is equivalent to

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

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

Assumption 3. *$f(x)$ and $g_j(y_j)$ for all $j \in [m]$ are all lower bounded, and let $f^* = \inf_x f(x) > -\infty$ and $g_j^* = \inf_{y_j} g_j(y_j) > -\infty$.*

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

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

$$\mathbb{E}\big[dist(0, \partial L(x^*, y^*_{[m]}, z^*))^2\big] \le \epsilon, \quad (2)$$

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

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

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

Notations:

- $\|\cdot\|$ 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 $\sigma^A_{\max}$ denote the minimum and maximum eigenvalues of $A^T A$, respectively.
- $\sigma^{B_j}_{\max}$ denotes the maximum eigenvalues of $B_j^T B_j$ for all $j \in [k]$, and $\sigma^B_{\max} = \max_{j=1}^k \sigma^{B_j}_{\max}$.
- $\sigma_{\min}(H_j)$ and $\sigma_{\max}(H_j)$ denote the minimum and maximum eigenvalues of matrix H_j for all $j \in [m]$, respectively; $\sigma_{\min}(H) = \min_{j\in[m]} \sigma_{\min}(H_j)$ and $\sigma_{\max}(H) = \max_{j\in[m]} \sigma_{\max}(H_j)$.
- $\sigma_{\min}(G)$ and $\sigma_{\max}(G)$ denotes the minimum and maximum eigenvalues of matrix G, respectively; the conditional number $\kappa_G = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)}$.
- 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.
- In both SPIDER-ADMM and online SPIDER-ADMM, K denotes the total number of iteration. In both SVRG-ADMM and SAGA-ADMM, 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.
- In the SVRG-ADMM algorithm, $y_j^{s,t}$ denotes output of the variable y_j in t-th inner loop and s-th outer loop.

[1]Department of Electrical & Computer Engineering, University of Pittsburgh, PA 15261, USA [2]College of Computer Science & Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China [3]MIIT Key Laboratory of Pattern Analysis & Machine Intelligence [4]JD Finance America Corporation. Correspondence to: Heng Huang <heng.huang@pitt.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. Theoretical Analysis

In this section, we at detail provide the theoretical analysis of the SPIDER-ADMM, online SPIDER-ADMM, nonconvex SVRG-ADMM and SAGA-ADMM. First, we introduce an useful lemma from Fang et al. (2018). Throughout the paper, let $n_k = \lceil k/q \rceil$ such that $(n_k - 1)q \le k \le n_k q - 1$.

Lemma 1. *(Fang et al., 2018) Under Assumption 1, the SPIDER generates stochastic gradient v_k satisfies for all $(n_k - 1)q + 1 \le k \le n_k q - 1$,*

$$\mathbb{E}\|v_k - \nabla f(x_k)\|^2 \le \frac{L^2}{|S_2|}\mathbb{E}\|x_k - x_{k-1}\|^2 + \mathbb{E}\|v_{k-1} - \nabla f(x_{k-1})\|^2. \tag{3}$$

From the above Lemma, telescoping (3) over i from $(n_k - 1)q + 1$ to k, we have

$$\mathbb{E}\|v_k - \nabla f(x_k)\|^2 \le \sum_{i=(n_k-1)q}^{k-1} \frac{L^2}{|S_2|}\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|v_{(n_k-1)q} - \nabla f(x_{(n_k-1)q})\|^2. \tag{4}$$

In Algorithm 1, due to $v_{(n_k-1)q} = \nabla f(x_{(n_k-1)q})$ and $|S_2| = b$, we have

$$\mathbb{E}\|v_k - \nabla f(x_k)\|^2 \le \sum_{i=(n_k-1)q}^{k-1} \frac{L^2}{b}\mathbb{E}\|x_{i+1} - x_i\|^2. \tag{5}$$

In Algorithm 2, using Assumption 2 and $|S_2| = b_2$, we obtain

$$\mathbb{E}\|v_k - \nabla f(x_k)\|^2 \le \sum_{i=(n_k-1)q}^{k-1} \frac{L^2}{b_2}\mathbb{E}\|x_{i+1} - x_i\|^2 + \frac{4\delta^2}{b_1}. \tag{6}$$

2.1. Convergence Analysis of the SPIDER-ADMM Algorithm

In this subsection, we conduct convergence analysis of the SPIDER-ADMM. We begin with giving some useful lemmas.

Algorithm 1 SPIDER-ADMM Algorithm

Input: b, q, K, $\eta > 0$ and $\rho > 0$;
Initialize: $x_0 \in \mathbb{R}^d$, $y_j^0 \in \mathbb{R}^p,\ j \in [m]$ and $z_0 \in \mathbb{R}^l$;
for $k = 0, 1, \cdots, K-1$ **do**
  if $\mathrm{mod}(k, q) = 0$ **then**
    Compute $v_k = \nabla f(x_k)$;
  else
    Uniformly randomly pick a mini-batch $\mathcal{I}_k$ (with replacement) from $\{1, 2, \cdots, n\}$ with $|\mathcal{I}_k| = b$, and compute

$$v_k = \nabla f_{\mathcal{I}_k}(x_k) - \nabla f_{\mathcal{I}_k}(x_{k-1}) + v_{k-1};$$

  end if
  $y_j^{k+1} = \arg\min_{y_j} \big\{\mathcal{L}_\rho(x_k, y_{[j-1]}^{k+1}, y_j, y_{[j+1:m]}^k, z_k) + \frac{1}{2}\|y_j - y_j^k\|^2_{H_j}\big\}$ for all $j \in [m]$;
  $x_{k+1} = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[m]}^{k+1}, z_k, v_k\big)$;
  $z_{k+1} = z_k - \rho(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c)$;
end for
Output: $\{x, y_{[m]}, z\}$ chosen uniformly random from $\{x_k, y_{[m]}^k, z_k\}_{k=1}^K$.

Lemma 2. *Under Assumption 1 and given the sequence* $\{x_k, y^k_{[m]}, z_k\}_{k=1}^K$ *from Algorithm 1, it holds that*

$$\begin{aligned}\mathbb{E}\|z_{k+1}-z_k\|^2 \leq & \frac{18L^2}{\sigma^A_{\min} b}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2 + (\frac{9L^2}{\sigma^A_{\min}} + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2})\|x_k - x_{k-1}\|^2 \\ & + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2}\|x_{k+1}-x_k\|^2. \end{aligned} \tag{7}$$

Proof. Using the optimal condition of the step 10 in Algorithm 1, we have

$$v_k + \frac{G}{\eta}(x_{k+1}-x_k) - A^T z_k + \rho A^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) = 0. \tag{8}$$

Using the step 11 of Algorithm 1, we have

$$A^T z_{k+1} = v_k + \frac{G}{\eta}(x_{k+1}-x_k). \tag{9}$$

It follows that

$$z_{k+1} = (A^T)^+\big(v_k + \frac{G}{\eta}(x_{k+1}-x_k)\big), \tag{10}$$

where $(A^T)^+$ is the pseudoinverse of A^T. By Assumption 4, *i.e.,* A is a full column matrix, we have $(A^T)^+ = A(A^TA)^{-1}$. By (10), we have

$$\begin{aligned}\mathbb{E}\|z_{k+1}-z_k\|^2 &= \mathbb{E}\|(A^T)^+\big(v_k + \frac{G}{\eta}(x_{k+1}-x_k) - v_{k-1} - \frac{G}{\eta}(x_k - x_{k-1})\big)\|^2 \\ &\leq \frac{1}{\sigma^A_{\min}}\mathbb{E}\|v_k + \frac{G}{\eta}(x_{k+1}-x_k) - v_{k-1} - \frac{G}{\eta}(x_k-x_{k-1})\|^2 \\ &\leq \frac{1}{\sigma^A_{\min}}\big[3\mathbb{E}\|v_k - v_{k-1}\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\mathbb{E}\|x_{k+1}-x_k\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\mathbb{E}\|x_k - x_{k-1}\|^2\big], \end{aligned} \tag{11}$$

where the first inequality follows by $((A^T)^+)^T(A^T)^+ = (A(A^TA)^{-1})^TA(A^TA)^{-1} = (A^TA)^{-1}$; the second inequality holds by the inequality $\|\sum_{i=1}^r \alpha_i\|^2 \leq r\sum_{i=1}^r\|\alpha_i\|^2$.

Next, considering the upper bound of $\|v_k - v_{k-1}\|^2$, we have

$$\begin{aligned}\mathbb{E}\|v_k - v_{k-1}\|^2 &= \mathbb{E}\|v_k - \nabla f(x_k) + \nabla f(x_k) - \nabla f(x_{k-1}) + \nabla f(x_{k-1}) - v_{k-1}\|^2 \\ &\leq 3\mathbb{E}\|v_k - \nabla f(x_k)\|^2 + 3\mathbb{E}\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 + 3\mathbb{E}\|\nabla f(x_{k-1}) - v_{k-1}\|^2 \\ &\leq \frac{3L^2}{b}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2 + 3L^2\mathbb{E}\|x_{k-1}-x_k\|^2 + \frac{3L^2}{b}\sum_{i=(n_k-1)q}^{k-2}\mathbb{E}\|x_{i+1}-x_i\|^2 \\ &\leq \frac{6L^2}{b}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2 + 3L^2\mathbb{E}\|x_{k-1}-x_k\|^2, \end{aligned} \tag{12}$$

where the second inequality holds by Assumption 1 and the inequality (5).

Finally, combining the inequalities (11) and (12), we obtain the above result.

□

Lemma 3. *Suppose the sequence* $\{x_k, y^k_{[m]}, z_k\}_{k=1}^K$ *is generated from Algorithm 1, and define a* Lyapunov *function* R_k *as follows:*

$$R_k = \mathcal{L}_\rho(x_k, y^k_{[m]}, z_k) + (\frac{9L^2}{\sigma^A_{\min}\rho} + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho})\|x_k - x_{k-1}\|^2 + \frac{2L^2}{\sigma^A_{\min}\rho b}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2. \tag{13}$$

Let $b=q$, $\eta=\frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0<\alpha\leq 1)$ *and* $\rho=\frac{\sqrt{170}\kappa_G L}{\sigma^A_{\min}\alpha}$*, then we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}(\|x_{k+1}-x_k\|^2+\sum_{j=1}^m\|y_j^k-y_j^{k+1}\|^2)\leq\frac{\mathbb{E}[R_0]-R^*}{K\gamma}, \tag{14}$$

where $\gamma=\min(\chi,\sigma^H_{\min})$ *with* $\chi\geq\frac{\sqrt{170}\kappa_G L}{4\alpha}$ *and* R^* *is a lower bound of the function* R_k*.*

Proof. By the optimal condition of step 9 in Algorithm 1, we have, for $j\in[m]$

$$\begin{aligned}
0 &= (y_j^k-y_j^{k+1})^T\big(\partial g_j(y_j^{k+1})-B_j^Tz_k+\rho B_j^T(Ax_k+\sum_{i=1}^j B_iy_i^{k+1}+\sum_{i=j+1}^m B_iy_i^k-c)+H_j(y_j^{k+1}-y_j^k)\big)\\
&\leq g_j(y_j^k)-g_j(y_j^{k+1})-(z_k)^T(B_jy_j^k-B_jy_j^{k+1})+\rho(B_jy_j^k-B_jy_j^{k+1})^T(Ax_k+\sum_{i=1}^j B_iy_i^{k+1}+\sum_{i=j+1}^m B_iy_i^k-c)-\|y_j^{k+1}-y_j^k\|^2_{H_j}\\
&= g_j(y_j^k)-g_j(y_j^{k+1})-(z_k)^T(Ax_k+\sum_{i=1}^{j-1}B_iy_i^{k+1}+\sum_{i=j}^m B_iy_i^k-c)+(z_k)^T(Ax_k+\sum_{i=1}^j B_iy_i^{k+1}+\sum_{i=j+1}^m B_iy_i^k-c)\\
&\quad+\frac{\rho}{2}\|Ax_k+\sum_{i=1}^{j-1}B_iy_i^{k+1}+\sum_{i=j}^m B_iy_i^k-c\|^2-\frac{\rho}{2}\|Ax_k+\sum_{i=1}^j B_iy_i^{k+1}+\sum_{i=j+1}^m B_iy_i^k-c\|^2-\frac{\rho}{2}\|B_jy_j^k-B_jy_j^{k+1}\|^2-\|y_j^{k+1}-y_j^k\|^2_{H_j}\\
&= \underbrace{f(x_k)+\sum_{i=1}^{j-1}g_i(y_i^{k+1})+\sum_{i=j}^m g_i(y_i^k)-(z_k)^T(Ax_k+\sum_{i=1}^{j-1}B_iy_i^{k+1}+\sum_{i=j}^m B_iy_i^k-c)+\frac{\rho}{2}\|Ax_k+\sum_{i=1}^{j-1}B_iy_i^{k+1}+\sum_{i=j}^m B_iy_i^k-c\|^2}_{\mathcal{L}_\rho(x_k,y^{k+1}_{[j-1]},y^k_{[j:m]},z_k)}\\
&\quad-\underbrace{\big(f(x_k)+\sum_{i=1}^{j}g_i(y_i^{k+1})+\sum_{i=j+1}^m g_i(y_i^k)-(z_k)^T(Ax_k+\sum_{i=1}^{j}B_iy_i^{k+1}+\sum_{i=j+1}^m B_iy_i^k-c)+\frac{\rho}{2}\|Ax_k+\sum_{i=1}^{j}B_iy_i^{k+1}+\sum_{i=j+1}^m B_iy_i^k-c\|^2\big)}_{\mathcal{L}_\rho(x_k,y^{k+1}_{[j]},y^k_{[j+1:m]},z_k)}\\
&\quad-\frac{\rho}{2}\|B_jy_j^k-B_jy_j^{k+1}\|^2-\|y_j^{k+1}-y_j^k\|^2_{H_j}\\
&\leq \mathcal{L}_\rho(x_k,y^{k+1}_{[j-1]},y^k_{[j:m]},z_k)-\mathcal{L}_\rho(x_k,y^{k+1}_{[j]},y^k_{[j+1:m]},z_k)-\sigma_{\min}(H_j)\|y_j^k-y_j^{k+1}\|^2,
\end{aligned} \tag{15}$$

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

$$\mathcal{L}_\rho(x_k,y^{k+1}_{[j]},y^k_{[j+1:m]},z_k)\leq\mathcal{L}_\rho(x_k,y^{k+1}_{[j-1]},y^k_{[j:m]},z_k)-\sigma_{\min}(H_j)\|y_j^k-y_j^{k+1}\|^2. \tag{16}$$

Telescoping inequality (16) over j from 1 to m, we obtain

$$\mathcal{L}_\rho(x_k,y^{k+1}_{[m]},z_k)\leq\mathcal{L}_\rho(x_k,y^k_{[m]},z_k)-\sigma^H_{\min}\sum_{j=1}^m\|y_j^k-y_j^{k+1}\|^2, \tag{17}$$

where $\sigma^H_{\min}=\min_{j\in[m]}\sigma_{\min}(H_j)$.

By Assumption 1, we have

$$0\leq f(x_k)-f(x_{k+1})+\nabla f(x_k)^T(x_{k+1}-x_k)+\frac{L}{2}\|x_{k+1}-x_k\|^2. \tag{18}$$

Using the optimal condition of step 10 in Algorithm 1, we have

$$0=(x_k-x_{k+1})^T\big(v_k-A^Tz_k+\rho A^T(Ax_{k+1}+\sum_{j=1}^m B_jy_j^{k+1}-c)+\frac{G}{\eta}(x_{k+1}-x_k)\big). \tag{19}$$

Combining (18) and (19), we have

$$
\begin{aligned}
0 &\le f(x_k) - f(x_{k+1}) + \nabla f(x_k)^T(x_{k+1}-x_k) + \frac{L}{2}\|x_{k+1}-x_k\|^2 \\
&\quad + (x_k - x_{k+1})^T\big(v_k - A^T z_k + \rho A^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{G}{\eta}(x_{k+1}-x_k)\big) \\
&= f(x_k) - f(x_{k+1}) + \frac{L}{2}\|x_k - x_{k+1}\|^2 - \frac{1}{\eta}\|x_k - x_{k+1}\|^2_G + (x_k - x_{k+1})^T(v_k - \nabla f(x_k)) \\
&\quad - (z_k)^T(Ax_k - Ax_{k+1}) + \rho(Ax_k - Ax_{k+1})^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) \\
&= f(x_k) - f(x_{k+1}) + \frac{L}{2}\|x_k - x_{k+1}\|^2 - \frac{1}{\eta}\|x_k - x_{k+1}\|^2_G + (x_k - x_{k+1})^T\big(v_k - \nabla f(x_k)\big) - (z_k)^T(Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c) \\
&\quad + (z_k)^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{\rho}{2}\big(\|Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 - \|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 - \|Ax_k - Ax_{k+1}\|^2\big) \\
&= \underbrace{f(x_k) + \sum_{j=1}^m g_j(y_j^{k+1}) - z_k^T(Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{\rho}{2}\|Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2}_{L_\rho(x_k, y_{[m]}^{k+1}, z_k)} \\
&\quad - \underbrace{\big(f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - z_k^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2\big)}_{L_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k)} \\
&\quad + \frac{L}{2}\|x_k - x_{k+1}\|^2 + (x_k - x_{k+1})^T\big(v_k - \nabla f(x_k)\big) - \frac{1}{\eta}\|x_k - x_{k+1}\|^2_G - \frac{\rho}{2}\|Ax_k - Ax_{k+1}\|^2 \\
&\le \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L}{2})\|x_{k+1}-x_k\|^2 + (x_k - x_{k+1})^T(v_k - \nabla f(x_k)) \\
&\le \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_{k+1}-x_k\|^2 + \frac{1}{2L}\|v_k - \nabla f(x_k)\|^2 \\
&\le \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_{k+1}-x_k\|^2 + \frac{L}{2b}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2,
\end{aligned}
$$

where the second equality follows by applying the equality $(a-b)^Tb = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ over the term $(Ax_k - Ax_{k+1})^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c)$; the third inequality follows by the inequality $a^Tb \le \frac{1}{2L}\|a\|^2 + \frac{L}{2}\|b\|^2$, and the forth inequality holds by the inequality (5). It follows that

$$
\mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) \le \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_{k+1}-x_k\|^2 + \frac{L}{2b}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2. \tag{20}
$$

Using the step 10 in Algorithm 1, we have

$$
\begin{aligned}
\mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_{k+1}) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) &= \frac{1}{\rho}\|z_{k+1} - z_k\|^2 \\
&\le \frac{18L^2}{\sigma_{\min}^A b\rho}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2 + (\frac{9L^2}{\sigma_{\min}^A\rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho})\|x_k - x_{k-1}\|^2 \\
&\quad + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_{k+1}-x_k\|^2,
\end{aligned} \tag{21}
$$

where the above inequality holds by Lemma 2.

Combining (17), (20) and (21), we have

$$\mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_{k+1}) \leq \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho})\|x_{k+1} - x_k\|^2$$
$$+ (\frac{L}{2b} + \frac{18L^2}{\sigma_{\min}^A b\rho}) \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + (\frac{9L^2}{\sigma_{\min}^A\rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho})\|x_k - x_{k-1}\|^2. \tag{22}$$

Next, we define a *Lyapunov* function R_k:

$$R_k = \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) + (\frac{9L^2}{\sigma_{\min}^A\rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho})\|x_k - x_{k-1}\|^2 + \frac{2L^2}{\sigma_{\min}^A\rho b}\sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2. \tag{23}$$

It follows that

$$\begin{aligned}
R_{k+1} &= \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_{k+1}) + (\frac{9L^2}{\sigma_{\min}^A\rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho})\|x_{k+1} - x_k\|^2 + \frac{2L^2}{\sigma_{\min}^A\rho b}\sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1} - x_i\|^2 \\
&\leq \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) + (\frac{9L^2}{\sigma_{\min}^A\rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho})\|x_k - x_{k-1}\|^2 + \frac{2L^2}{\sigma_{\min}^A\rho b}\sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 \\
&\quad - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{2L^2}{\sigma_{\min}^A\rho b})\|x_{k+1} - x_k\|^2 - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|^2 \\
&\quad + (\frac{L}{2b} + \frac{18L^2}{\sigma_{\min}^A b\rho}) \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 \\
&\leq R_k - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{2L^2}{\sigma_{\min}^A\rho b})\|x_{k+1} - x_k\|^2 - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|^2 \\
&\quad + (\frac{L}{2b} + \frac{18L^2}{\sigma_{\min}^A b\rho}) \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2,
\end{aligned} \tag{24}$$

where the first inequality holds by the inequality (22) and the equality

$$\sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1} - x_i\|^2 = \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|x_{k+1} - x_k\|^2.$$

Since $(n_k-1)q \leq k \leq n_k q - 1$, and let $(n_k-1)q \leq l \leq n_k q - 1$, telescoping inequality (24) over k from $(n_k-1)q$ to k, we have

$$\begin{aligned}
\mathbb{E}[R_{k+1}] &\leq \mathbb{E}[R_{(n_k-1)q}] - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b}) \sum_{l=(n_k-1)q}^{k} \|x_{l+1} - x_l\|^2 \\
&\quad - \sigma^H_{\min} \sum_{l=(n_k-1)q}^{k} \sum_{j=1}^{m} \|y^l_j - y^{l+1}_j\|^2 + (\frac{L}{2b} + \frac{18L^2}{\sigma^A_{\min} b\rho}) \sum_{l=(n_k-1)q}^{k} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 \\
&\leq \mathbb{E}[R_{(n_k-1)q}] - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b}) \sum_{i=(n_k-1)q}^{k} \|x_{i+1} - x_i\|^2 \\
&\quad - \sigma^H_{\min} \sum_{i=(n_k-1)q}^{k-1} \sum_{j=1}^{m} \|y^i_j - y^{i+1}_j\|^2 + (\frac{Lq}{2b} + \frac{18L^2 q}{\sigma^A_{\min} b\rho}) \sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1} - x_i\|^2 \\
&= \mathbb{E}[R_{(n_k-1)q}] - \underbrace{(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b} - \frac{Lq}{2b} - \frac{18L^2 q}{\sigma^A_{\min} b\rho})}_{\chi} \sum_{i=(n_k-1)q}^{k} \|x_{i+1} - x_i\|^2 \\
&\quad - \sigma^H_{\min} \sum_{i=(n_k-1)q}^{k-1} \sum_{j=1}^{m} \|y^i_j - y^{i+1}_j\|^2, \qquad (25)
\end{aligned}$$

where the second inequality holds by the fact that

$$\sum_{l=(n_k-1)q}^{k} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 \leq \sum_{l=(n_k-1)q}^{k} \sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1} - x_i\|^2 \leq q \sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1} - x_i\|^2.$$

Since $b = q$, we have

$$\begin{aligned}
\chi &= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b} - \frac{Lq}{2b} - \frac{18L^2 q}{\sigma^A_{\min} b\rho} \\
&= \underbrace{\frac{\sigma_{\min}(G)}{\eta} - \frac{3L}{2}}_{L_1} + \underbrace{\frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{27L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b}}_{L_2}. \qquad (26)
\end{aligned}$$

Given $0 < \eta \leq \frac{2\sigma_{\min}(G)}{3L}$, we have $L_1 \geq 0$. Further, let $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma^A_{\min}\alpha}$, we have

$$\begin{aligned}
L_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{27L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{27L^2\kappa_G^2}{2\sigma^A_{\min}\rho\alpha^2} - \frac{27L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{27L^2\kappa_G^2}{2\sigma^A_{\min}\rho\alpha^2} - \frac{27L^2\kappa_G^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{2\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} \\
&= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{85L^2\kappa_G^2}{2\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\
&\geq \frac{\sqrt{170}\kappa_G L}{4\alpha}, \qquad (27)
\end{aligned}$$

where the first inequality holds by $\kappa_G \geq 1$ and $b \geq 1 \geq \alpha^2$ and the second inequality holds by $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma^A_{\min}\alpha}$. Thus, we obtain $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$.

Since A is a full column rank matrix, we have $(A^T)^+ = A(A^TA)^{-1}$. It follows that $\sigma_{\max}((A^T)^+)^T(A^T)^+) = \sigma_{\max}((A^TA)^{-1}) = \frac{1}{\sigma^A_{\min}}$. Using (10), then we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{k+1}, y^{k+1}_{[m]}, z_{k+1}) = f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - z_{k+1}^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 \\
&= f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \langle (A^T)^+(v_k + \frac{G}{\eta}(x_{k+1} - x_k)), Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\rangle + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 \\
&= f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \langle (A^T)^+(v_k - \nabla f(x_k) + \nabla f(x_k) + \frac{G}{\eta}(x_{k+1} - x_k)), Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\rangle \\
&\quad + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 \\
&\geq f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \frac{2}{\sigma^A_{\min}\rho}\|v_k - \nabla f(x_k)\|^2 - \frac{2}{\sigma^A_{\min}\rho}\|\nabla f(x_k)\|^2 - \frac{2\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{\rho}{8}\|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 \\
&\geq f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \frac{2L^2}{\sigma^A_{\min}b\rho}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1} - x_i\|^2 - \frac{2\delta^2}{\sigma^A_{\min}\rho} - \frac{2\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho}\|x_{k+1} - x_k\|^2
\end{aligned} \tag{28}$$

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 $\langle (A^T)^+(v_k - \nabla f(x_k)), Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\rangle$, $\langle (A^T)^+ v_k, Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\rangle$ and $\langle (A^T)^+\frac{G}{\eta}(x_{k+1} - x_k), Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\rangle$ with $\beta = \frac{\rho}{4}$, respectively; and the second inequality follows by the inequality (5) and Assumption 3. Using the definition of R_k, we have

$$R_{k+1} \geq f^* + \sum_{j=1}^m g_j^* - \frac{2\delta^2}{\sigma^A_{\min}\rho}, \ \forall\, k = 0, 1, 2, \cdots. \tag{29}$$

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

Telescoping inequality (25) over k from 0 to K, we have

$$\begin{aligned}
\mathbb{E}[R_K] - \mathbb{E}[R_0] &= (\mathbb{E}[R_q] - \mathbb{E}[R_0]) + (\mathbb{E}[R_{2q}] - \mathbb{E}[R_q]) + \cdots + (\mathbb{E}[R_K] - \mathbb{E}[R_{(n_K-1)q}]) \\
&\leq -\sum_{i=0}^{q-1}(\chi\|x_{i+1} - x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2) - \sum_{i=q}^{2q-1}(\chi\|x_{i+1} - x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2) \\
&\quad - \cdots - \sum_{i=(n_k-1)q}^{K-1}(\chi\|x_{i+1} - x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2) \\
&= -\sum_{i=0}^{K-1}(\chi\|x_{i+1} - x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2).
\end{aligned} \tag{30}$$

Finally, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}(\|x_{k+1} - x_k\|^2 + \sum_{j=1}^m\|y_j^k - y_j^{k+1}\|^2) \leq \frac{\mathbb{E}[R_0] - R^*}{K\gamma}, \tag{31}$$

where $\gamma = \min(\chi, \sigma^H_{\min})$ with $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$.

□

Theorem 1. *Suppose the sequence* $\{x_k, y^k_{[m]}, z_k)^K_{k=1}$ *is generated from Algorithm 1, and let* $b = q$, $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0 < \alpha \leq 1)$, $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma^A_{\min}\alpha}$, *and*

$$\nu_1 = m\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(H)\big),\ \nu_2 = 3(L^2 + \frac{\sigma^2_{\max}(G)}{\eta^2}),\ \nu_3 = \frac{18L^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2}, \tag{32}$$

then we have

$$\min_{1\leq k\leq K} \mathbb{E}\big[dist(0, \partial L(x_k, y^k_{[m]}, z_k))^2\big] \leq \frac{\nu_{\max}}{K}\sum_{k=1}^{K-1}\theta_k \leq \frac{3\nu_{\max}(R_0 - R^*)}{K\gamma}, \tag{33}$$

where $\gamma = \min(\chi, \sigma^H_{\min})$ *with* $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$, $\nu_{\max} = \max\{\nu_1, \nu_2, \nu_3\}$ *and* R^* *is a lower bound of the function* R_k. *It implies that the number of iteration* K *satisfies*

$$K = \frac{3\nu_{\max}(R_0 - R^*)}{\epsilon\gamma}$$

then $(x_{k^*}, y^{k^*}_{[m]}, z_{k^*})$ *is an* ϵ*-approximate stationary point of* (1), *where* $k^* = \arg\min_k \theta_k$.

Proof. First, we define an useful variable $\theta_k = \mathbb{E}\big[\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2 + \frac{1}{q}\sum_{i=(n_k-1)q}^{k}\|x_{i+1} - x_i\|^2 + \sum_{j=1}^m\|y^k_j - y^{k+1}_j\|^2\big]$. By the optimal condition of the step 9 in Algorithm 1, we have, for all $j \in [m]$

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \partial_{y_j} L(x, y_{[m]}, z))^2\big]_{k+1} &= \mathbb{E}\big[\text{dist}(0, \partial g_j(y^{k+1}_j) - B^T_j z_{k+1})^2\big] \\
&= \|B^T_j z_k - \rho B^T_j(Ax_k + \sum_{i=1}^{j} B_i y^{k+1}_i + \sum_{i=j+1}^{m} B_i y^k_i - c) - H_j(y^{k+1}_j - y^k_j) - B^T_j z_{k+1}\|^2 \\
&= \|\rho B^T_j A(x_{k+1} - x_k) + \rho B^T_j \sum_{i=j+1}^{m} B_i(y^{k+1}_i - y^k_i) - H_j(y^{k+1}_j - y^k_j)\|^2 \\
&\leq m\rho^2\sigma^{B_j}_{\max}\sigma^A_{\max}\|x_{k+1} - x_k\|^2 + m\rho^2\sigma^{B_j}_{\max}\sum_{i=j+1}^{m}\sigma^{B_i}_{\max}\|y^{k+1}_i - y^k_i\|^2 + m\sigma^2_{\max}(H_j)\|y^{k+1}_j - y^k_j\|^2 \\
&\leq m\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(H)\big)\theta_k,
\end{aligned} \tag{34}$$

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

By the step 10 of Algorithm 1, we have

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \nabla_x L(x, y_{[m]}, z))^2\big]_{k+1} &= \mathbb{E}\|A^T z_{k+1} - \nabla f(x_{k+1})\|^2 \\
&= \mathbb{E}\|v_k - \nabla f(x_{k+1}) - \frac{G}{\eta}(x_k - x_{k+1})\|^2 \\
&= \mathbb{E}\|v_k - \nabla f(x_k) + \nabla f(x_k) - \nabla f(x_{k+1}) - \frac{G}{\eta}(x_k - x_{k+1})\|^2 \\
&\leq \sum_{i=(n_k-1)q}^{k-1}\frac{3L^2}{b}\mathbb{E}\|x_{i+1} - x_i\|^2 + 3(L^2 + \frac{\sigma^2_{\max}(G)}{\eta^2})\|x_k - x_{k+1}\|^2 \\
&\leq 3(L^2 + \frac{\sigma^2_{\max}(G)}{\eta^2})\theta_k,
\end{aligned} \tag{35}$$

where the second inequality holds by $b = q$.

By the step 11 of Algorithm 1, we have

$$\begin{aligned}\mathbb{E}\big[\text{dist}(0,\nabla_z L(x,y_{[m]},z))^2\big]_{k+1} &= \mathbb{E}\|Ax_{k+1}+\sum_{j=1}^m B_j y_j^{k+1}-c\|^2 \\ &= \frac{1}{\rho^2}\mathbb{E}\|z_{k+1}-z_k\|^2 \\ &\leq \frac{18L^2}{\sigma_{\min}^A b\rho^2}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2+(\frac{9L^2}{\sigma_{\min}^A\rho^2}+\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2})\|x_k-x_{k-1}\|^2 \\ &\quad+\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\|x_{k+1}-x_k\|^2 \\ &\leq (\frac{18L^2}{\sigma_{\min}^A\rho^2}+\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2})\theta_k, \end{aligned} \tag{36}$$

where the second inequality holds by $b=q$.

Let

$$\nu_1 = m\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A+\rho^2(\sigma_{\max}^B)^2+\sigma_{\max}^2(H)\big),\ \nu_2 = 3(L^2+\frac{\sigma_{\max}^2(G)}{\eta^2}),\ \nu_3 = \frac{18L^2}{\sigma_{\min}^A\rho^2}+\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}. \tag{37}$$

By (31), we have

$$\frac{1}{K}\sum_{k=0}^{K-1}(\|x_{k+1}-x_k\|^2+\sum_{j=1}^m\|y_j^k-y_j^{k+1}\|^2) \leq \frac{\mathbb{E}[R_0]-R^*}{K\gamma}, \tag{38}$$

where $\gamma=\min(\chi,\sigma_{\min}^H)$ with $\chi\geq\frac{\sqrt{170}\kappa_G L}{4\alpha}$. Since

$$\sum_{k=0}^{K-1}\sum_{i=(n_k-1)q}^{k}\|x_{i+1}-x_i\|^2 \leq q\sum_{k=0}^{K-1}\|x_{k+1}-x_k\|^2, \tag{39}$$

by (34), (35) and (36), we have

$$\min_{1\leq k\leq K}\mathbb{E}\big[\text{dist}(0,\partial L(x_k,y_{[m]}^k,z_k))^2\big] \leq \frac{\nu_{\max}}{K}\sum_{k=1}^{K-1}\theta_k \leq \frac{3\nu_{\max}(R_0-R^*)}{K\gamma}, \tag{40}$$

where $\gamma=\min(\chi,\sigma_{\min}^H)$ with $\chi\geq\frac{\sqrt{170}\kappa_G L}{4\alpha}$ and $\nu_{\max}=\max\{\nu_1,\nu_2,\nu_3\}$.

Given $\eta=\frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0<\alpha\leq 1)$ and $\rho=\frac{\sqrt{170}\kappa_G L}{\sigma_{\min}^A\alpha}$, since m is relatively small, it easy verifies that $\nu_{\max}=O(1)$ and $\gamma=O(1)$, which are independent on n and K. Thus, we obtain

$$\min_{1\leq k\leq K}\mathbb{E}\big[\text{dist}(0,\partial L(x_k,y_{[m]}^k,z_k))^2\big] \leq O(\frac{1}{K}). \tag{41}$$

□

2.2. Convergence Analysis of the Online SPIDER-ADMM Algorithm

In this subsection, we conduct convergence analysis of the online SPIDER-ADMM. First, we give some useful lemmas.

Algorithm 2 Online SPIDER-ADMM Algorithm

Input: $b_1, b_2, q, K, \eta > 0$ and $\rho > 0$;
Initialize: $x_0 \in \mathbb{R}^d$, $y_j^0 \in \mathbb{R}^p,\ j \in [m]$ and $z_0 \in \mathbb{R}^l$;
for $k = 0, 1, \cdots, K-1$ **do**
  if $\mathrm{mod}(k, q) = 0$ **then**
    Draw S_1 samples with $|S_1| = b_1$, and compute $v_k = \frac{1}{b_1}\sum_{i\in S_1}\nabla f_i(x_k)$;
  else
    Draw S_2 samples with $|S_2| = b_2 = \sqrt{b_1}$, and compute

$$v_k = \frac{1}{b_2}\sum_{i\in S_2}\big(\nabla f_i(x_k) - f_i(x_{k-1})\big) + v_{k-1};$$

  end if
  $y_j^{k+1} = \arg\min_{y_j}\big\{\mathcal{L}_\rho(x_k, y_{[j-1]}^{k+1}, y_j, y_{[j+1:m]}^k, z_k) + \frac{1}{2}\|y_j - y_j^k\|^2_{H_j}\big\}$ for all $j \in [m]$;
  $x_{k+1} = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[m]}^{k+1}, z_k, v_k\big)$;
  $z_{k+1} = z_k - \rho(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c)$;
end for
Output: $\{x, y_{[m]}, z\}$ chosen uniformly random from $\{x_k, y_{[m]}^k, z_k\}_{k=1}^K$.

Lemma 4. *Under Assumption 1 and given the sequence $\{x_k, y_{[m]}^k, z_k\}_{k=1}^K$ from Algorithm 2, it holds that*

$$\begin{aligned}\mathbb{E}\|z_{k+1} - z_k\|^2 \leq & \frac{18L^2}{\sigma_{\min}^A b_2}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1} - x_i\|^2 + \frac{72\delta^2}{\sigma_{\min}^A b_1} + (\frac{9L^2}{\sigma_{\min}^A} + \frac{3\sigma^2_{\max}(G)}{\sigma_{\min}^A\eta^2})\|x_k - x_{k-1}\|^2 \\ & + \frac{3\sigma^2_{\max}(G)}{\sigma_{\min}^A\eta^2}\|x_{k+1} - x_k\|^2. \end{aligned} \tag{42}$$

Proof. The proof of this lemma is the same to the proof of Lemma 2. □

Lemma 5. *Suppose the sequence $\{x_k, y_{[m]}^k, z_k\}_{k=1}^K$ is generated from Algorithm 2, and define a* Lyapunov *function Φ_k as follows:*

$$\Phi_k = \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) + (\frac{9L^2}{\sigma_{\min}^A\rho} + \frac{3\sigma^2_{\max}(G)}{\sigma_{\min}^A\eta^2\rho})\|x_k - x_{k-1}\|^2 + \frac{2L^2}{\sigma_{\min}^A\rho b}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1} - x_i\|^2. \tag{43}$$

Let $b_2 = q$, $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma_{\min}^A\alpha}$, then we have

$$\frac{1}{K}\sum_{k=0}^{K-1}(\|x_{k+1} - x_k\|^2 + \sum_{j=1}^m\|y_j^k - y_j^{k+1}\|^2) \leq \frac{\mathbb{E}[\Phi_0] - \Phi^*}{K\gamma} + \frac{2\delta^2}{b_1 L} + \frac{72\delta^2}{\sigma_{\min}^A b_1\rho}, \tag{44}$$

where $\gamma = \min(\chi, \sigma_{\min}^H)$, $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$ and Φ^* is a lower bound of the function Φ_k.

Proof. This proof is the same as the proof of Lemma 3.

By the optimal condition of step 9 in Algorithm 2, we have, for $j \in [m]$

$$
\begin{aligned}
0 &= (y_j^k - y_j^{k+1})^T \big(\partial g_j(y_j^{k+1}) - B_j^T z_k + \rho B_j^T(Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) + H_j(y_j^{k+1} - y_j^k)\big) \\
&\leq g_j(y_j^k) - g_j(y_j^{k+1}) - (z_k)^T(B_j y_j^k - B_j y_j^{k+1}) + \rho(B_j y_j^k - B_j y_j^{k+1})^T(Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) - \|y_j^{k+1} - y_j^k\|^2_{H_j} \\
&= g_j(y_j^k) - g_j(y_j^{k+1}) - (z_k)^T(Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c) + (z_k)^T(Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) \\
&\quad + \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c\|^2 - \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\|^2 - \frac{\rho}{2}\|B_j y_j^k - B_j y_j^{k+1}\|^2 - \|y_j^{k+1} - y_j^k\|^2_{H_j} \\
&= \underbrace{f(x_k) + \sum_{i=1}^{j-1} g_i(y_i^{k+1}) + \sum_{i=j}^{m} g_i(y_i^k) - (z_k)^T(Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c) + \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c\|^2}_{\mathcal{L}_\rho(x_k, y_{[j-1]}^{k+1}, y_{[j:m]}^k, z_k)} \\
&\quad - \underbrace{\big(f(x_k) + \sum_{i=1}^{j} g_i(y_i^{k+1}) + \sum_{i=j+1}^{m} g_i(y_i^k) - (z_k)^T(Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) + \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\|^2\big)}_{\mathcal{L}_\rho(x_k, y_{[j]}^{k+1}, y_{[j+1:m]}^k, z_k)} \\
&\quad - \frac{\rho}{2}\|B_j y_j^k - B_j y_j^{k+1}\|^2 - \|y_j^{k+1} - y_j^k\|^2_{H_j} \\
&\leq \mathcal{L}_\rho(x_k, y_{[j-1]}^{k+1}, y_{[j:m]}^k, z_k) - \mathcal{L}_\rho(x_k, y_{[j]}^{k+1}, y_{[j+1:m]}^k, z_k) - \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|^2,
\end{aligned} \tag{45}
$$

where the first inequality holds by the convexity of function $g_j(y)$, 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 $(By_j^k - By_j^{k+1})^T(Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c)$. Thus, we have, for all $j \in [m]$

$$
\mathcal{L}_\rho(x_k, y_{[j]}^{k+1}, y_{[j+1:m]}^k, z_k) \leq \mathcal{L}_\rho(x_k, y_{[j-1]}^{k+1}, y_{[j:m]}^k, z_k) - \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|^2. \tag{46}
$$

Telescoping inequality (46) over j from 1 to m, we obtain

$$
\mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) \leq \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) - \sigma_{\min}^H \sum_{j=1}^{m} \|y_j^k - y_j^{k+1}\|^2. \tag{47}
$$

Using Assumption 1, we have

$$
0 \leq f(x_k) - f(x_{k+1}) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{48}
$$

Using the optimal condition of step 10 in Algorithm 2, we have

$$
0 = (x_k - x_{k+1})^T \big(v_k - A^T z_k + \rho A^T(Ax_{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c) + \frac{G}{\eta}(x_{k+1} - x_k)\big). \tag{49}
$$

Combining (48) and (49), we have

$$
\begin{aligned}
0 \leq\ & f(x_k) - f(x_{k+1}) + \nabla f(x_k)^T(x_{k+1}-x_k) + \frac{L}{2}\|x_{k+1}-x_k\|^2 \\
& + (x_k - x_{k+1})^T\big(v_k - A^T z_k + \rho A^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{G}{\eta}(x_{k+1}-x_k)\big) \\
=\ & f(x_k) - f(x_{k+1}) + \frac{L}{2}\|x_k - x_{k+1}\|^2 - \frac{1}{\eta}\|x_k - x_{k+1}\|_G^2 + (x_k - x_{k+1})^T(v_k - \nabla f(x_k)) \\
& - (z_k)^T(Ax_k - Ax_{k+1}) + \rho(Ax_k - Ax_{k+1})^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) \\
=\ & f(x_k) - f(x_{k+1}) + \frac{L}{2}\|x_k - x_{k+1}\|^2 - \frac{1}{\eta}\|x_k - x_{k+1}\|_G^2 + (x_k - x_{k+1})^T\big(v_k - \nabla f(x_k)\big) - (z_k)^T(Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c) \\
& + (z_k)^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{\rho}{2}\big(\|Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 - \|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 - \|Ax_k - Ax_{k+1}\|^2\big) \\
=\ & \underbrace{f(x_k) + \sum_{j=1}^m g_j(y_j^{k+1}) - z_k^T(Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{\rho}{2}\|Ax_k + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2}_{L_\rho(x_k, y_{[m]}^{k+1}, z_k)} \\
& - \underbrace{\big(f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - z_k^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2\big)}_{L_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k)} \\
& + \frac{L}{2}\|x_k - x_{k+1}\|^2 + (x_k - x_{k+1})^T\big(v_k - \nabla f(x_k)\big) - \frac{1}{\eta}\|x_k - x_{k+1}\|_G^2 - \frac{\rho}{2}\|Ax_k - Ax_{k+1}\|^2 \\
\leq\ & \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L}{2})\|x_{k+1}-x_k\|^2 + (x_k - x_{k+1})^T(v_k - \nabla f(x_k)) \\
\leq\ & \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_{k+1}-x_k\|^2 + \frac{1}{2L}\|v_k - \nabla f(x_k)\|^2 \\
\leq\ & \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_{k+1}-x_k\|^2 + \frac{L}{2b_2}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2 + \frac{2\delta^2}{b_1 L},
\end{aligned}
$$

where the second equality follows by applying the equality $(a-b)^T b = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ over the term $(Ax_k - Ax_{k+1})^T(Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c)$; the third inequality follows by the inequality $a^T b \leq \frac{1}{2L}\|a\|^2 + \frac{L}{2}\|b\|^2$, and the forth inequality holds by the inequality (6). It follows that

$$
\begin{aligned}
\mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) \leq & \mathcal{L}_\rho(x_k, y_{[m]}^{k+1}, z_k) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_{k+1}-x_k\|^2 \\
& + \frac{L}{2b_2}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2 + \frac{2\delta^2}{b_1 L}.
\end{aligned} \tag{50}
$$

Using the step 11 in Algorithm 2, we have

$$\begin{aligned}
\mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_{k+1}) - \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_k) &= \frac{1}{\rho}\|z_{k+1} - z_k\|^2 \\
&\leq \frac{18L^2}{\sigma_{\min}^A b_2 \rho} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + (\frac{9L^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho})\|x_k - x_{k-1}\|^2 \\
&\quad + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x_{k+1} - x_k\|^2 + \frac{72\delta^2}{\sigma_{\min}^A b_1 \rho},
\end{aligned} \tag{51}$$

where the above inequality holds by Lemma 4.

Combining (47), (50) and (51), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_{k+1}) &\leq \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho})\|x_{k+1} - x_k\|^2 \\
&\quad + (\frac{L}{2b_2} + \frac{18L^2}{\sigma_{\min}^A b_2 \rho}) \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + (\frac{9L^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho})\|x_k - x_{k-1}\|^2 \\
&\quad + \frac{2\delta^2}{b_1 L} + \frac{72\delta^2}{\sigma_{\min}^A b_1 \rho}.
\end{aligned} \tag{52}$$

Next, we define an useful *Lyapunov* function Φ_k:

$$\Phi_k = \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) + (\frac{9L^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho})\|x_k - x_{k-1}\|^2 + \frac{2L^2}{\sigma_{\min}^A \rho b_2} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2. \tag{53}$$

It follows that

$$\begin{aligned}
\Phi_{k+1} &= \mathcal{L}_\rho(x_{k+1}, y_{[m]}^{k+1}, z_{k+1}) + (\frac{9L^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho})\|x_{k+1} - x_k\|^2 + \frac{2L^2}{\sigma_{\min}^A \rho b_2} \sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1} - x_i\|^2 \\
&\leq \mathcal{L}_\rho(x_k, y_{[m]}^k, z_k) + (\frac{9L^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho})\|x_k - x_{k-1}\|^2 + \frac{2L^2}{\sigma_{\min}^A \rho b_2} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 \\
&\quad - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L^2}{\sigma_{\min}^A \rho} - \frac{2L^2}{\sigma_{\min}^A \rho b_2})\|x_{k+1} - x_k\|^2 \\
&\quad - \frac{3(-1)\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}(\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2) + (\frac{L}{2b_2} + \frac{18L^2}{\sigma_{\min}^A b_2 \rho}) \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + \frac{2\delta^2}{b_1 L} + \frac{72\delta^2}{\sigma_{\min}^A b_1 \rho} \\
&\leq \Phi_k - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L^2}{\sigma_{\min}^A \rho} - \frac{2L^2}{\sigma_{\min}^A \rho b_2})\|x_{k+1} - x_k\|^2 \\
&\quad + (\frac{L}{2b_2} + \frac{18L^2}{\sigma_{\min}^A b_2 \rho}) \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + \frac{2\delta^2}{b_1 L} + \frac{72\delta^2}{\sigma_{\min}^A b_1 \rho},
\end{aligned} \tag{54}$$

where the first inequality follows by the inequality (52) and the equality

$$\sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1} - x_i\|^2 = \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + \|x_{k+1} - x_k\|^2.$$

Since $(n_k-1)q \leq k \leq n_k q - 1$ and let $(n_k-1)q \leq l \leq n_k q - 1$, then telescoping equality (54) over k from $(n_k-1)q$ to k, we have

$$\begin{aligned}
\mathbb{E}[\Phi_{k+1}] &\leq \mathbb{E}[\Phi_{(n_k-1)q}] - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b_2}\Big) \sum_{l=(n_k-1)q}^{k} \|x_{l+1}-x_l\|^2 \\
&\quad - \sigma^H_{\min} \sum_{l=(n_k-1)q}^{k} \sum_{j=1}^{m} \|y_j^l - y_j^{l+1}\|^2 + \Big(\frac{L}{2b_2} + \frac{18L^2}{\sigma^A_{\min} b_2 \rho}\Big) \sum_{l=(n_k-1)q}^{k} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1}-x_i\|^2 + \frac{2\delta^2}{b_1 L} + \frac{72\delta^2}{\sigma^A_{\min} b_1 \rho} \\
&\leq \mathbb{E}[\Phi_{(n_k-1)q}] - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b_2}\Big) \sum_{i=(n_k-1)q}^{k} \|x_{i+1}-x_i\|^2 \\
&\quad - \sigma^H_{\min} \sum_{i=(n_k-1)q}^{k} \sum_{j=1}^{m} \|y_j^i - y_j^{i+1}\|^2 + \Big(\frac{Lq}{2b_2} + \frac{18L^2 q}{\sigma^A_{\min} b_2 \rho}\Big) \sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1}-x_i\|^2 + \frac{2\delta^2}{b_1 L} + \frac{72\delta^2}{\sigma^A_{\min} b_1 \rho} \\
&= \mathbb{E}[\Phi_{(n_k-1)q}] - \underbrace{\Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b_2} - \frac{Lq}{2b_2} - \frac{18L^2 q}{\sigma^A_{\min} b_2 \rho}\Big)}_{\chi} \sum_{i=(n_k-1)q}^{k} \|x_{i+1}-x_i\|^2 \\
&\quad - \sigma^H_{\min} \sum_{i=(n_k-1)q}^{k} \sum_{j=1}^{m} \|y_j^i - y_j^{i+1}\|^2 + \frac{2\delta^2}{b_1 L} + \frac{72\delta^2}{\sigma^A_{\min} b_1 \rho}, \qquad (55)
\end{aligned}$$

where the second inequality holds by the fact that

$$\sum_{l=(n_k-1)q}^{k} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1}-x_i\|^2 \leq \sum_{l=(n_k-1)q}^{k} \sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1}-x_i\|^2 \leq q \sum_{i=(n_k-1)q}^{k} \mathbb{E}\|x_{i+1}-x_i\|^2.$$

Since $b_2 = q$, we have

$$\begin{aligned}
\chi &= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho b_2} - \frac{Lq}{2b_2} - \frac{18L^2 q}{\sigma^A_{\min}\rho b_2} \\
&= \underbrace{\frac{\sigma_{\min}(G)}{\eta} - \frac{3L}{2}}_{L_1} + \underbrace{\frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{27L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho}}_{L_2}. \qquad (56)
\end{aligned}$$

Given $0 < \eta \leq \frac{2\sigma_{\min}(G)}{3L}$, we have $L_1 \geq 0$. Further, let $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma^A_{\min}\alpha}$, we have

$$\begin{aligned}
L_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{27L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{27\kappa_G^2 L^2}{2\sigma^A_{\min}\rho\alpha^2} - \frac{27L^2}{\sigma^A_{\min}\rho} - \frac{2L^2}{\sigma^A_{\min}\rho} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{27\kappa_G^2 L^2}{2\sigma^A_{\min}\rho\alpha^2} - \frac{27\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{2\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} \\
&= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{85\kappa_G^2 L^2}{2\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\
&\geq \frac{\sqrt{170}\kappa_G L}{4\alpha}, \qquad (57)
\end{aligned}$$

where the first inequality follows by $\kappa_G \geq 1$, ≥ 1 and $0 < \alpha \leq 1$; and the second inequality holds by $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma^A_{\min}\alpha}$. It follows that $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$.

Since $\sigma_{\max}((A^T)^+)^T(A^T)^+) = \frac{1}{\sigma^A_{\min}}$, using (10), we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{k+1}, y^{k+1}_{[m]}, z_{k+1}) = f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - z_{k+1}^T(Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c) + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\|^2 \\
&= f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \langle (A^T)^+(v_k + \frac{G}{\eta}(x_{k+1}-x_k)), Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\rangle + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\|^2 \\
&= f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \langle (A^T)^+(v_k - \nabla f(x_k) + \nabla f(x_k) + \frac{G}{\eta}(x_{k+1}-x_k)), Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\rangle \\
&\quad + \frac{\rho}{2}\|Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\|^2 \\
&\geq f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \frac{2}{\sigma^A_{\min}\rho}\|v_k - \nabla f(x_k)\|^2 - \frac{2}{\sigma^A_{\min}\rho}\|\nabla f(x_k)\|^2 - \frac{2\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho}\|x_{k+1}-x_k\|^2 \\
&\quad + \frac{\rho}{8}\|Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\|^2 \\
&\geq f(x_{k+1}) + \sum_{j=1}^m g_j(y_j^{k+1}) - \frac{2L^2}{\sigma^A_{\min}b_2\rho}\sum_{i=(n_k-1)q}^{k-1}\mathbb{E}\|x_{i+1}-x_i\|^2 - \frac{8\delta^2}{\sigma^A_{\min}b_1\rho} - \frac{2\delta^2}{\sigma^A_{\min}\rho} - \frac{2\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho}\|x_{k+1}-x_k\|^2
\end{aligned} \tag{58}$$

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 $\langle (A^T)^+(v_k - \nabla f(x_k)), Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\rangle$, $\langle (A^T)^+v_k, Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\rangle$ and $\langle (A^T)^+\frac{G}{\eta}(x_{k+1}-x_k), Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\rangle$ with $\beta = \frac{\rho}{4}$, respectively; and the second inequality follows by the inequality (6) and Assumption 3. Using the definition of Φ_k, we have

$$\Phi_{k+1} \geq f^* + \sum_{j=1}^m g_j^* - \frac{2(4+b_1)\delta^2}{\sigma^A_{\min}\rho b_1}, \ \forall\, k = 0, 1, 2, \cdots. \tag{59}$$

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

Further, telescoping equality (55) over k from 0 to K, we have

$$\begin{aligned}
\mathbb{E}[\Phi_K] - \mathbb{E}[\Phi_0] &= (\mathbb{E}[\Phi_q] - \mathbb{E}[\Phi_0]) + (\mathbb{E}[\Phi_{2q}] - \mathbb{E}[\Phi_q]) + \cdots + (\mathbb{E}[\Phi_K] - \mathbb{E}[\Phi_{(n_K-1)q}]) \\
&\leq -\sum_{i=0}^{q-1}(\chi\|x_{i+1}-x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2) - \sum_{i=q}^{2q-1}(\chi\|x_{i+1}-x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2) \\
&\quad - \cdots - \sum_{i=(n_k-1)q}^{K-1}(\chi\|x_{i+1}-x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2) + \frac{2K\delta^2}{b_1L} + \frac{72K\delta^2}{\sigma^A_{\min}b_1\rho} \\
&= -\sum_{i=0}^{K-1}(\chi\|x_{i+1}-x_i\|^2 + \sigma^H_{\min}\sum_{j=1}^m\|y_j^i - y_j^{i+1}\|^2) + \frac{2K\delta^2}{b_1L} + \frac{72K\delta^2}{\sigma^A_{\min}b_1\rho}.
\end{aligned} \tag{60}$$

Thus, the above inequality implies that

$$\frac{1}{K}\sum_{k=0}^{K-1}(\|x_{k+1}-x_k\|^2 + \sum_{j=1}^m\|y_j^k - y_j^{k+1}\|^2) \leq \frac{\mathbb{E}[\Phi_0] - \Phi^*}{K\gamma} + \frac{2\delta^2}{b_1L\gamma} + \frac{72\delta^2}{\sigma^A_{\min}b_1\rho\gamma}, \tag{61}$$

where $\gamma = \min(\chi, \sigma^H_{\min})$ and $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$.

□

Theorem 2. *Suppose the sequence* $\{x_k, y_{[m]}^k, z_k\}_{k=1}^K$ *is generated from Algorithm 2, and let* $b_2 = q = \sqrt{b_1}$, $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0 < \alpha \leq 1)$, $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma_{\min}^A \alpha}$, *and*

$$\nu_1 = m\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big),\ \nu_2 = 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2}),\ \nu_3 = \frac{18L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}, \tag{62}$$

then we have

$$\begin{aligned}\min_{1\leq k\leq K} \mathbb{E}\big[\mathit{dist}(0, \partial L(x_k, y_{[m]}^k, z_k))^2\big] &\leq \frac{\nu_{\max}}{K}\sum_{k=1}^{K-1}\theta_k + \frac{w}{b_1} \\ &\leq \frac{3\nu_{\max}(\Phi_0 - \Phi^*)}{K\gamma} + \frac{6\nu_{\max}\delta^2}{b_1\gamma}(\frac{1}{L} + \frac{36}{\sigma_{\min}^A\rho}) + \frac{w}{b_1},\end{aligned} \tag{63}$$

where $w = 12\delta^2\max\{1, \frac{6}{\sigma_{\min}^A\rho^2}\}$, $\gamma = \min(\chi, \sigma_{\min}^H)$ *with* $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$, $\nu_{\max} = \max\{\nu_1, \nu_2, \nu_3\}$ *and* Φ^* *is a lower bound of the function* Φ_k. *It implies that* K *and* b_1 *satisfy*

$$K = \frac{6\nu_{\max}(\Phi_0 - \Phi^*)}{\epsilon\gamma}, \quad b_1 = \frac{12\nu_{\max}\delta^2}{\epsilon\gamma}(\frac{1}{L} + \frac{36}{\sigma_{\min}^A\rho}) + \frac{2w}{\epsilon}$$

then $(x_{k^*}, y_{[m]}^{k^*}, z_{k^*})$ *is an* ϵ*-approximate stationary point of (1), where* $k^* = \arg\min_k \theta_k$.

Proof. First, we define an useful variable $\theta_k = \mathbb{E}\big[\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2 + \frac{1}{q}\sum_{i=(n_k-1)q}^{k}\|x_{i+1} - x_i\|^2 + \sum_{j=1}^m\|y_j^k - y_j^{k+1}\|^2\big]$. By the optimal condition of the step 9 in Algorithm 2, we have, for all $j \in [m]$

$$\begin{aligned}\mathbb{E}\big[\text{dist}(0, \partial_{y_j}L(x, y_{[m]}, z))^2\big]_{k+1} &= \mathbb{E}\big[\text{dist}(0, \partial g_j(y_j^{k+1}) - B_j^T z_{k+1})^2\big] \\ &= \|B_j^T z_k - \rho B_j^T(Ax_k + \sum_{i=1}^j B_i y_i^{k+1} + \sum_{i=j+1}^m B_i y_i^k - c) - H_j(y_j^{k+1} - y_j^k) - B_j^T z_{k+1}\|^2 \\ &= \|\rho B_j^T A(x_{k+1} - x_k) + \rho B_j^T\sum_{i=j+1}^m B_i(y_i^{k+1} - y_i^k) - H_j(y_j^{k+1} - y_j^k)\|^2 \\ &\leq m\rho^2\sigma_{\max}^{B_j}\sigma_{\max}^A\|x_{k+1} - x_k\|^2 + m\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^m\sigma_{\max}^{B_i}\|y_i^{k+1} - y_i^k\|^2 + m\sigma_{\max}^2(H_j)\|y_j^{k+1} - y_j^k\|^2 \\ &\leq m\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big)\theta_k,\end{aligned} \tag{64}$$

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

By the step 10 of Algorithm 2, we have

$$\begin{aligned}\mathbb{E}\big[\text{dist}(0, \nabla_x L(x, y_{[m]}, z))^2\big]_{k+1} &= \mathbb{E}\|A^T z_{k+1} - \nabla f(x_{k+1})\|^2 \\ &= \mathbb{E}\|v_k - \nabla f(x_{k+1}) - \frac{G}{\eta}(x_k - x_{k+1})\|^2 \\ &= \mathbb{E}\|v_k - \nabla f(x_k) + \nabla f(x_k) - \nabla f(x_{k+1}) - \frac{G}{\eta}(x_k - x_{k+1})\|^2 \\ &\leq \sum_{i=(n_k-1)q}^{k-1}\frac{3L^2}{b_2}\mathbb{E}\|x_{i+1} - x_i\|^2 + \frac{12\delta^2}{b_1} + 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\|x_k - x_{k+1}\|^2 \\ &\leq 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\theta_k + \frac{12\delta^2}{b_1},\end{aligned} \tag{65}$$

where the second inequality holds by $b_2 = q$.

By the step 11 of Algorithm 2, we have

$$
\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \nabla_z L(x, y_{[m]}, z))^2\big]_{k+1} &= \mathbb{E}\|Ax_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|^2 \\
&= \frac{1}{\rho^2}\mathbb{E}\|z_{k+1} - z_k\|^2 \\
&\leq \frac{18L^2}{\sigma_{\min}^A b_2 \rho^2} \sum_{i=(n_k-1)q}^{k-1} \mathbb{E}\|x_{i+1} - x_i\|^2 + \big(\frac{9L^2}{\sigma_{\min}^A \rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho^2}\big)\|x_k - x_{k-1}\|^2 \\
&\quad + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho^2}\|x_{k+1} - x_k\|^2 + \frac{72\delta^2}{b_1 \sigma_{\min}^A \rho^2} \\
&\leq \big(\frac{18L^2}{\sigma_{\min}^A \rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho^2}\big)\theta_k + \frac{72\delta^2}{b_1 \sigma_{\min}^A \rho^2},
\end{aligned} \tag{66}
$$

where the second inequality holds by $b_2 = q$.

Let

$$
\nu_1 = m\big(\rho^2 \sigma_{\max}^B \sigma_{\max}^A + \rho^2 (\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big), \ \nu_2 = 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2}), \ \nu_3 = \frac{18L^2}{\sigma_{\min}^A \rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho^2}. \tag{67}
$$

By (61), we have

$$
\frac{1}{K}\sum_{k=0}^{K-1}(\|x_{k+1} - x_k\|^2 + \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|^2) \leq \frac{\mathbb{E}[\Phi_0] - \Phi^*}{K\gamma} + \frac{2\delta^2}{b_1 L \gamma} + \frac{72\delta^2}{\sigma_{\min}^A b_1 \rho \gamma}, \tag{68}
$$

where $\gamma = \min(\chi, \sigma_{\min}^H)$ and $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$. Since

$$
\sum_{k=0}^{K-1} \sum_{i=(n_k-1)q}^{k} \|x_{i+1} - x_i\|^2 \leq q \sum_{k=0}^{K-1} \|x_{k+1} - x_k\|^2, \tag{69}
$$

by (64), (65) and (66), we have

$$
\begin{aligned}
\min_{1\leq k \leq K} \mathbb{E}\big[\text{dist}(0, \partial L(x_k, y_{[m]}^k, z_k))^2\big] &\leq \frac{\nu_{\max}}{K}\sum_{k=1}^{K-1}\theta_k + \max\{\frac{12\delta^2}{b_1}, \frac{72\delta^2}{b_1 \sigma_{\min}^A \rho^2}\} \\
&\leq \frac{3\nu_{\max}(\Phi_0 - \Phi^*)}{K\gamma} + \frac{6\nu_{\max}\delta^2}{b_1\gamma}(\frac{1}{L} + \frac{36}{\sigma_{\min}^A \rho}) + \max\{\frac{12\delta^2}{b_1}, \frac{72\delta^2}{b_1 \sigma_{\min}^A \rho^2}\},
\end{aligned} \tag{70}
$$

where $\gamma = \min(\chi, \sigma_{\min}^H)$ with $\chi \geq \frac{\sqrt{170}\kappa_G L}{4\alpha}$ and $\nu_{\max} = \max\{\nu_1, \nu_2, \nu_3\}$.

Given $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{\sqrt{170}\kappa_G L}{\sigma_{\min}^A \alpha}$, since m is relatively small, it easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(1)$, which are independent on b_1 and K. Thus, we obtain

$$
\min_{1\leq k \leq K} \mathbb{E}\big[\text{dist}(0, \partial L(x_k, y_{[m]}^k, z_k))^2\big] \leq O(\frac{1}{K}) + O(\frac{1}{b_1}). \tag{71}
$$

□

2.3. Theoretical Analysis of Non-convex SVRG-ADMM Algorithm

In this subsection, we first extend the existing nonconvex SVRG-ADMM (Huang et al., 2016; Zheng & Kwok, 2016) to the multi-blocks setting for solving the problem (1), which is summarized in Algorithm 3. Then we study the convergence analysis of this non-convex SVRG-ADMM.

Algorithm 3 Nonconvex SVRG-ADMM Algorithm

1: **Input:** $b, T, M, S = [T/M]$, $\eta > 0$ and $\rho > 0$;
2: **Initialize:** $x_0^1 = \tilde{x}^1$, z_0^1, and $y_j^{0,1}$ for $j \in [m]$;
3: **for** $s = 1, 2, \cdots, S$ **do**
4: $\quad \nabla f(\tilde{x}^s) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{x}^s)$;
5: $\quad$ **for** $t = 0, 1, \cdots, M-1$ **do**
6: $\quad\quad$ Uniformly random pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n\}$ with $|\mathcal{I}_t| = b$, and compute

$$v_t^s = \nabla f_{\mathcal{I}_t}(x_t^s) - \nabla f_{\mathcal{I}_t}(\tilde{x}^s) + \nabla f(\tilde{x}^s);$$

7: $\quad\quad y_j^{s,t+1} = \arg\min_{y_j} \big\{\mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_j, y_{[j+1:m]}^{s,t}, z_t) + \frac{1}{2}\|y_j - y_j^{s,t}\|_{H_j}^2\big\}$ with $H_j \succ 0$ for all $j \in [m]$;
8: $\quad\quad x_{t+1}^s = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[m]}^{s,t+1}, z_t^s, v_t^s\big)$;
9: $\quad\quad z_{t+1}^s = z_t^s - \rho(Ax_{t+1}^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c)$;
10: $\quad$ **end for**
11: $\quad \tilde{x}^{s+1} = x_0^{s+1} = x_M^s$, $y_j^{s+1,0} = y_j^{s,M}$ for all $j \in [m]$, $z_0^{s+1} = z_M^s$;
12: **end for**
13: **Output:** $\{x, y_{[m]}, z\}$ chosen uniformly random from $\{(x_t^s, y_{[m]}^{s,t}, z_t^s)_{t=1}^M\}_{s=1}^S$.

In Algorithm 3, we give

$$\begin{aligned}\hat{\mathcal{L}}_\rho(x, y_{[m]}^{s,t+1}, z_t^s, v_t^s) =& f(x_t) + (v_t^s)^T(x - x_t^s) + \frac{1}{2\eta}\|x - x_t^s\|_G^2 + \sum_{j=1}^m g_j(y_j^{s,t+1}) \\ & - (z_t^s)^T(Ax + \sum_{j=1}^m B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax + \sum_{j=1}^m B_j y_j^{s,t+1} - c\|^2, \end{aligned} \tag{72}$$

where $\eta > 0$ and $G \succ 0$.

Lemma 6. *Suppose the sequence $\big\{(x_t^s, y_{[m]}^{s,t}, z_t^s)_{t=1}^M\big\}_{s=1}^S$ is generated by Algorithm 3. The following inequality holds*

$$\begin{aligned}\mathbb{E}\|z_{t+1}^s - z_t^s\|^2 \leq & \frac{9L^2}{\sigma_{\min}^A b}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2}\mathbb{E}\|x_{t+1}^s - x_t^s\|^2 \\ & + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2} + \frac{9L^2}{\sigma_{\min}^A})\mathbb{E}\|x_t^s - x_{t-1}^s\|^2. \end{aligned} \tag{73}$$

Proof. Using the optimal condition for the step 8 of Algorithm 3, we have

$$v_t^s + \frac{1}{\eta}G(x_{t+1}^s - x_t^s) - A^T z_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c) = 0, \tag{74}$$

By the step 10 of Algorithm 3, we have

$$A^T z_{t+1}^s = v_t^s + \frac{1}{\eta}G(x_{t+1}^s - x_t^s). \tag{75}$$

It follows that

$$z_{t+1}^s = (A^T)^+\big(v_t^s + \frac{G}{\eta}(x_{t+1}^s - x_t^s)\big), \tag{76}$$

where $(A^T)^+$ is the pseudoinverse of A^T. By Assumption 4, *i.e.,* A is a full column matrix, we have $(A^T)^+ = A(A^TA)^{-1}$. Using (76), then we have

$$\begin{aligned}\mathbb{E}\|z^s_{t+1} - z^s_t\|^2 &= \mathbb{E}\|(A^T)^+\big(v^s_t + \frac{G}{\eta}(x^s_{t+1} - x^s_t) - v^s_{t-1} - \frac{G}{\eta}(x^s_t - x^s_{t-1})\big)\|^2 \\ &\leq \frac{1}{\sigma^A_{\min}}\big[3\mathbb{E}\|v^s_t - v^s_{t-1}\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\mathbb{E}\|x^s_{t+1} - x^s_t\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\mathbb{E}\|x^s_t - x^s_{t-1}\|^2\big]. \qquad (77)\end{aligned}$$

Next, considering the upper bound of $\|v^s_t - v^s_{t-1}\|^2$, we have

$$\begin{aligned}\mathbb{E}\|v^s_t - v^s_{t-1}\|^2 &= \mathbb{E}\|v^s_t - \nabla f(x^s_t) + \nabla f(x^s_t) - \nabla f(x^s_{t-1}) + \nabla f(x^s_{t-1}) - v^s_{t-1}\|^2 \\ &\leq 3\mathbb{E}\|v^s_t - \nabla f(x^s_t)\|^2 + 3\mathbb{E}\|\nabla f(x^s_t) - \nabla f(x^s_{t-1})\|^2 + 3\mathbb{E}\|\nabla f(x^s_{t-1}) - v^s_{t-1}\|^2 \\ &\leq \frac{3L^2}{b}\|x^s_t - \tilde{x}^s\|^2 + \frac{3L^2}{b}\|x^s_{t-1} - \tilde{x}^s\|^2 + 3L^2\mathbb{E}\|x^s_t - x^s_{t-1}\|^2, \qquad (78)\end{aligned}$$

where the second inequality holds by Lemma 3 of (Reddi et al., 2016) and Assumption 1. Finally, combining (77) with (78), we obtain the above result. □

Lemma 7. *Suppose the sequence* $\{(x^s_t, y^{s,t}_{[m]}, z^s_t)^M_{t=1}\}^S_{s=1}$ *is generated from Algorithm 3, and define a* Lyapunov *function:*

$$\Gamma^s_t = \mathbb{E}\big[\mathcal{L}_\rho(x^s_t, y^{s,t}_{[m]}, z^s_t) + (\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x^s_t - x^s_{t-1}\|^2 + \frac{9L^2}{\sigma^A_{\min}\rho b}\|x^s_{t-1} - \tilde{x}^s\|^2 + c_t\|x^s_t - \tilde{x}^s\|^2\big], \qquad (79)$$

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

$$c_t = \begin{cases}\frac{18L^2}{\sigma^A_{\min}\rho b} + \frac{L}{b} + (1+\beta)c_{t+1}, \ 1 \leq t \leq M, \\ 0, \ t \geq M+1.\end{cases}$$

Let $M = [n^{\frac{1}{3}}]$, $b = [n^{\frac{2}{3}}]$, $\eta = \frac{\alpha\sigma_{\min}(G)}{5L}$ $(0 < \alpha \leq 1)$ *and* $\rho = \frac{2\sqrt{231}\kappa_G L}{\sigma^A_{\min}\alpha}$, *we have*

$$\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^{M-1}(\sigma^H_{\min}\sum_{j=1}^m\|y^{s,t}_j - y^{s,t+1}_j\|^2 + \frac{L}{2b}\|x^s_t - \tilde{x}^s\|^2_2 + \chi_t\|x^s_{t+1} - x^s_t\|^2) \leq \frac{\Gamma^1_0 - \Gamma^*}{T}. \qquad (80)$$

where Γ^* *denotes a lower bound of* Γ^s_t *and and* $\chi_t \geq \frac{\sqrt{231}\kappa_G L}{2\alpha} > 0$.

Proof. By the optimal condition of step 7 in Algorithm 3, we have, for $j \in [m]$

$$
\begin{aligned}
0 &= (y_j^{s,t} - y_j^{s,t+1})^T \big(\partial g_j(y_j^{s,t+1}) - B_j^T z_t^s + \rho B_j^T (Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{m} B_i y_i^{s,t} - c) + H_j (y_j^{s,t+1} - y_j^{s,t})\big) \\
&\leq g_j(y_j^{s,t}) - g_j(y_j^{s,t+1}) - (z_t^s)^T (B_j y_j^{s,t} - B_j y_j^{s,t+1}) + \rho (B_j y_j^{s,t} - B_j y_j^{s,t+1})^T (Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{m} B_i y_i^{s,t} - c) \\
&\quad - \|y_j^{s,t+1} - y_j^{s,t}\|_{H_j}^2 \\
&= g_j(y_j^{s,t}) - g_j(y_j^{s,t+1}) - (z_t^s)^T (Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^{m} B_i y_i^{s,t} - c) + (z_t^s)^T (Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{m} 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}^{m} 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}^{m} B_i y_i^{s,t} - c\|^2 - \|y_j^{s,t+1} - y_j^{s,t}\|_{H_j}^2 \\
&\quad - \frac{\rho}{2} \|B_j y_j^{s,t} - B_j y_j^{s,t+1}\|^2 \\
&= \underbrace{f(x_t^s) + \sum_{i=1}^{j-1} g_i(y_i^{s,t+1}) + \sum_{i=j}^{m} g_i(y_i^{s,t}) - (z_t^s)^T (Ax_t^s + \sum_{i=1}^{j-1} B_i y_i^{s,t+1} + \sum_{i=j}^{m} 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}^{m} B_i y_i^{s,t} - c\|^2}_{\mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:m]}^{s,t}, z_t^s)} \\
&\quad \underbrace{- \big(f(x_t^s) + \sum_{i=1}^{j} g_i(y_i^{s,t+1}) + \sum_{i=j+1}^{m} g_i(y_i^{s,t}) - (z_t^s)^T (Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{m} 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}^{m} B_i y_i^{s,t} - c\|^2\big)}_{\mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:m]}^{s,t}, z_t^s)} \\
&\quad - \|y_j^{s,t+1} - y_j^{s,t}\|_{H_j}^2 - \frac{\rho}{2} \|B_j y_j^{s,t} - B_j y_j^{s,t+1}\|^2 \\
&\leq \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:m]}^{s,t}, z_t^s) - \mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:m]}^{s,t}, z_t^s) - \sigma_{\min}(H_j) \|y_j^{s,t} - y_j^{s,t+1}\|^2,
\end{aligned} \tag{81}
$$

where the first inequality holds by the convexity of function $g_j(y)$, 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 $(By_j^{s,t} - By_j^{s,t+1})^T (Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{m} B_i y_i^{s,t} - c)$. Thus, we have, for all $j \in [m]$

$$
\mathcal{L}_\rho(x_t^s, y_{[j]}^{s,t+1}, y_{[j+1:m]}^{s,t}, z_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_{[j:m]}^{s,t}, z_t^s) - \sigma_{\min}(H_j) \|y_j^{s,t} - y_j^{s,t+1}\|^2. \tag{82}
$$

Telescoping inequality (82) over j from 1 to m, we obtain

$$
\mathcal{L}_\rho(x_t^s, y_{[m]}^{s,t+1}, z_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[m]}^{s,t}, z_t^s) - \sigma_{\min}^H \sum_{j=1}^{m} \|y_j^{s,t} - y_j^{s,t+1}\|^2, \tag{83}
$$

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

By Assumption 1, we have

$$
0 \leq f(x_t^s) - f(x_{t+1}^s) + \nabla f(x_t^s)^T (x_{t+1}^s - x_t^s) + \frac{L}{2} \|x_{t+1}^s - x_t^s\|^2. \tag{84}
$$

Using optimal condition of the step 8 in Algorithm 3, we have

$$
0 = (x_t^s - x_{t+1}^s)^T \big(v_t^s - A^T z_t^s + \rho A^T (Ax_{t+1}^s + \sum_{j=1}^{m} B_j y_j^{s,t+1} - c) + \frac{G}{\eta}(x_{t+1}^s - x_t^s)\big). \tag{85}
$$

Combining (84) and (85), 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 z_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^m 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(v_t^s - \nabla f(x_t^s)) \\
& - (z_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}^m 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(v_t^s - \nabla f(x_t^s)) - (z_t^s)^T(Ax_t^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c) \\
& + (z_t^s)^T(Ax_{t+1}^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\big(\|Ax_t^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c\|^2 - \|Ax_{t+1}^s + \sum_{j=1}^m 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}^m g_j(y_j^{s,t+1}) - (z_t^s)^T(Ax_t^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_t^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c\|^2}_{\mathcal{L}_\rho(x_t^s, y_{[m]}^{s,t+1}, z_t^s)} \\
& - \underbrace{\big(f(x_{t+1}^s) + \sum_{j=1}^m g_j(y_j^{s,t+1}) - (z_t^s)^T(Ax_{t+1}^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c\|^2\big)}_{\mathcal{L}_\rho(x_{t+1}^s, y_{[m]}^{s,t+1}, z_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_{[m]}^{s,t+1}, z_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[m]}^{s,t+1}, z_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_{[m]}^{s,t+1}, z_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[m]}^{s,t+1}, z_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_{[m]}^{s,t+1}, z_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[m]}^{s,t+1}, z_t^s) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 + \frac{L}{2b}\|x_t^s - \tilde{x}^s\|^2, \qquad (86)
\end{aligned}
$$

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 $(Ax_t^s - Ax_{t+1}^s)^T(Ax_{t+1}^s + \sum_{j=1}^m B_j y_j^{s,t+1} - c)$, the inequality (ii) holds by the inequality $a^Tb \leq \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|b\|^2$, and the inequality (iii) holds by Lemma 3 of (Reddi et al., 2016). Thus, we obtain

$$
\mathcal{L}_\rho(x_{t+1}^s, y_{[m]}^{s,t+1}, z_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[m]}^{s,t+1}, z_t^s) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 + \frac{L}{2b}\|x_t^s - \tilde{x}^s\|^2. \qquad (87)
$$

By the step 9 in Algorithm 3, we have

$$
\begin{aligned}
\mathcal{L}_\rho(x_{t+1}^s, y_{[m]}^{s,t+1}, z_{t+1}^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[m]}^{s,t+1}, z_t^s) &= \frac{1}{\rho}\|z_{t+1}^s - z_t^s\|^2 \\
&\leq \frac{9L^2}{\sigma_{\min}^A b\rho}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2 \\
&\quad + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2, \qquad (88)
\end{aligned}
$$

where the first inequality follows by Lemma 6.

Combining (83), (87) and (88), we have

$$\mathcal{L}_\rho(x^s_{t+1}, y^{s,t+1}_{[m]}, z^s_{t+1}) \leq \mathcal{L}_\rho(x^s_t, y^{s,t}_{[m]}, z^s_t) - \sigma^H_{\min}\sum_{j=1}^m \|y^{s,t}_j - y^{s,t+1}_j\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L)\|x^s_t - x^s_{t+1}\|^2$$
$$+ \frac{L}{2b}\|x^s_t - \tilde{x}^s\|^2 + \frac{9L^2}{\sigma^A_{\min} b\rho}\big(\|x^s_t - \tilde{x}^s\|^2 + \|x^s_{t-1} - \tilde{x}^s\|^2\big) + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho}\|x^s_{t+1} - x^s_t\|^2$$
$$+ (\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x^s_t - x^s_{t-1}\|^2. \tag{89}$$

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

$$\Gamma^s_t = \mathbb{E}\big[\mathcal{L}_\rho(x^s_t, y^{s,t}_{[m]}, z^s_t) + (\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x^s_t - x^s_{t-1}\|^2 + \frac{9L^2}{\sigma^A_{\min}\rho b}\|x^s_{t-1} - \tilde{x}^s\|^2 + c_t\|x^s_t - \tilde{x}^s\|^2\big]. \tag{90}$$

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

$$\begin{aligned}\|x^s_{t+1} - x^s_t + x^s_t - \tilde{x}^s\|^2 &= \|x^s_{t+1} - x^s_t\|^2 + 2(x^s_{t+1} - x^s_t)^T(x^s_t - \tilde{x}^s) + \|x^s_t - \tilde{x}^s\|^2 \\ &\leq \|x^s_{t+1} - x^s_t\|^2 + 2\big(\frac{1}{2\beta}\|x^s_{t+1} - x^s_t\|^2 + \frac{\beta}{2}\|x^s_t - \tilde{x}^s\|^2\big) + \|x^s_t - \tilde{x}^s\|^2 \\ &= (1 + 1/\beta)\|x^s_{t+1} - x^s_t\|^2 + (1+\beta)\|x^s_t - \tilde{x}^s\|^2,\end{aligned} \tag{91}$$

where the above inequality holds by the Cauchy-Schwarz inequality with $\beta > 0$. Using (89), we have

$$\begin{aligned}\Gamma^s_{t+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x^s_{t+1}, y^{s,t+1}_{[m]}, z^s_{t+1}) + (\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x^s_{t+1} - x^s_t\|^2 + \frac{9L^2}{\sigma^A_{\min} b\rho}\|x^s_t - \tilde{x}^s\|^2 + c_{t+1}\|x^s_{t+1} - \tilde{x}^s\|^2\big] \\ &\leq \mathcal{L}_\rho(x^s_t, y^{s,t}_{[m]}, z^s_t) + (\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x^s_t - x^s_{t-1}\|^2 + \frac{9L^2}{\sigma^A_{\min}\rho b}\|x^s_{t-1} - \tilde{x}^s\|^2 + \big(\frac{18L^2}{\sigma^A_{\min}\rho b} + \frac{L}{b} + (1+\beta)c_{t+1}\big)\|x^s_t - \tilde{x}^s\|^2 \\ &\quad - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - (1+1/\beta)c_{t+1}\big)\|x^s_t - x^s_{t+1}\|^2 \\ &\quad - \sigma^H_{\min}\sum_{j=1}^m\|y^{s,t}_j - y^{s,t+1}_j\|^2 - \frac{L}{2b}\|x^s_t - \tilde{x}^s\|^2 \\ &\leq \Gamma^s_t - \chi_t\|x^s_t - x^s_{t+1}\|^2 - \sigma^H_{\min}\sum_{j=1}^m\|y^{s,t}_j - y^{s,t+1}_j\|^2 - \frac{L}{2b}\|x^s_t - \tilde{x}^s\|^2,\end{aligned} \tag{92}$$

where $c_t = \frac{18L^2}{\sigma^A_{\min}\rho b} + \frac{L}{b} + (1+\beta)c_{t+1}$ and $\chi_t = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - (1+1/\beta)c_{t+1}$.

Next, we will prove the relationship between Γ^{s+1}_1 and Γ^s_M. Since $x^{s+1}_0 = x^s_M = \tilde{x}^{s+1}$, we have

$$v^{s+1}_0 = \nabla f_{\mathcal{I}}(x^{s+1}_0) - \nabla f_{\mathcal{I}}(x^{s+1}_0) + \nabla f(x^{s+1}_0) = \nabla f(x^{s+1}_0) = \nabla f(x^s_M). \tag{93}$$

Thus, we obtain

$$\begin{aligned}\mathbb{E}\|v^{s+1}_0 - v^s_M\|^2 &= \mathbb{E}\|\nabla f(x^s_M) - \nabla f_{\mathcal{I}}(x^s_M) + \nabla f_{\mathcal{I}}(\tilde{x}^s) - \nabla f(\tilde{x}^s)\|^2 \\ &= \|\nabla f_{\mathcal{I}}(x^s_M) - \nabla f_{\mathcal{I}}(\tilde{x}^s) - \mathbb{E}_{\mathcal{I}}[\nabla f_{\mathcal{I}}(x^s_M) - \nabla f_{\mathcal{I}}(\tilde{x}^s)]\|^2 \\ &\leq \frac{1}{bn}\sum_{i=1}^n \mathbb{E}\|\nabla f_i(x^s_M) - \nabla f_i(\tilde{x}^s)\|^2 \\ &\leq \frac{L^2}{b}\|x^s_M - \tilde{x}^s\|^2.\end{aligned} \tag{94}$$

By the step 9 of Algorithm 3, we have

$$\begin{aligned}\|z_1^{s+1}-z_M^s\|^2 &\leq \frac{1}{\sigma_{\min}^A}\|v_0^{s+1}-v_M^s+\frac{G}{\eta}(x_1^{s+1}-x_0^{s+1})+\frac{G}{\eta}(x_M^s-x_{M-1}^s)\|^2 \\ &= \frac{1}{\sigma_{\min}^A}\|\nabla f(x_M^s)-v_M^s+\frac{G}{\eta}(x_1^{s+1}-x_M^s)+\frac{G}{\eta}(x_M^s-x_{M-1}^s)\|^2 \\ &\leq \frac{1}{\sigma_{\min}^A}\big(3\|\nabla f(x_M^s)-v_M^s\|^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1}-x_M^s\|^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_M^s-x_{M-1}^s\|^2\big) \\ &\leq \frac{1}{\sigma_{\min}^A}\big(\frac{3L^2}{b}\|x_M^s-\tilde{x}^s\|_2^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1}-x_M^s\|^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_M^s-x_{M-1}^s\|^2\big). \end{aligned} \tag{95}$$

Since $x_M^s=x_0^{s+1}$, $y_j^{s,M}=y_j^{s+1,0}$ for all $j\in[m]$ and $z_M^s=z_0^{s+1}$, using (83), we have

$$\mathcal{L}_\rho(x_0^{s+1},y_{[m]}^{s+1,1},z_0^{s+1}) \leq \mathcal{L}_\rho(x_M^s,y_{[m]}^{s,M},z_M^s)-\sigma_{\min}^H\sum_{j=1}^m\|y_j^{s,M}-y_j^{s+1,1}\|^2. \tag{96}$$

By (87), we have

$$\mathcal{L}_\rho(x_1^{s+1},y_{[m]}^{s+1,1},z_0^{s+1}) \leq \mathcal{L}_\rho(x_0^{s+1},y_{[m]}^{s+1,1},z_0^{s+1})-(\frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L)\|x_0^{s+1}-x_1^{s+1}\|^2. \tag{97}$$

By (88), we have

$$\begin{aligned}\mathcal{L}_\rho(x_1^{s+1},y_{[m]}^{s+1,1},z_1^{s+1}) &\leq \mathcal{L}_\rho(x_1^{s+1},y_{[m]}^{s+1,1},z_0^{s+1})+\frac{1}{\rho}\|z_1^{s+1}-z_0^{s+1}\|^2 \\ &\leq \mathcal{L}_\rho(x_1^{s+1},y_{[m]}^{s+1,1},z_0^{s+1})+\frac{1}{\sigma_{\min}^A\rho}\big(\frac{3L^2}{b}\|x_M^s-\tilde{x}^s\|_2^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1}-x_M^s\|^2 \\ &\quad +\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_M^s-x_{M-1}^s\|^2\big). \end{aligned} \tag{98}$$

where the second inequality holds by the inequality (95).

Combining (96), (97) with (98), we have

$$\begin{aligned}\mathcal{L}_\rho(x_1^{s+1},y_{[m]}^{s+1,1},z_1^{s+1}) &\leq \mathcal{L}_\rho(x_M^s,y_{[m]}^{s,M},z_M^s)-\sigma_{\min}^H\sum_{j=1}^m\|y_j^{s,M}-y_j^{s+1,1}\|^2-(\frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L)\|x_0^{s+1}-x_1^{s+1}\|^2+ \\ &\quad \frac{1}{\sigma_{\min}^A\rho}\big(\frac{3L^2d}{b}\|x_M^s-\tilde{x}^s\|_2^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1}-x_M^s\|^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_M^s-x_{M-1}^s\|^2\big). \end{aligned} \tag{99}$$

Therefore, we have

$$
\begin{aligned}
\Gamma_1^{s+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_1^{s+1}, y_{[m]}^{s+1,1}, z_1^{s+1}) + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_1^{s+1} - x_0^{s+1}\|^2 + \frac{9L^2}{\sigma_{\min}^A b\rho}\|x_0^{s+1} - \tilde{x}^{s+1}\|^2 + c_1\|x_1^{s+1} - \tilde{x}^{s+1}\|^2\big] \\
&= \mathcal{L}_\rho(x_1^{s+1}, y_{[m]}^{s+1,1}, z_1^{s+1}) + \big(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho} + c_1\big)\|x_1^{s+1} - x_0^{s+1}\|^2 \\
&\leq \mathcal{L}_\rho(x_M^s, y_{[m]}^{s,M}, z_M^s) + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_M^s - x_{M-1}^s\|^2 + \frac{9L^2}{\sigma_{\min}^A\rho b}\|x_{M-1}^s - \tilde{x}^s\|_2^2 + (\frac{18L^2}{\sigma_{\min}^A\rho b} + \frac{L}{b})\|x_M^s - \tilde{x}^s\|_2^2 \\
&\quad - \sigma_{\min}^H\sum_{j=1}^m\|y_j^{s,M} - y_j^{s+1,1}\|^2 - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - c_1\big)\|x_1^{s+1} - x_M^s\|_2^2 \\
&\quad - \frac{9L^2}{\sigma_{\min}^A\rho}\|x_M^s - x_{M-1}^s\|_2^2 - \frac{9L^2}{\sigma_{\min}^A\rho b}\|x_{M-1}^s - \tilde{x}^s\|_2^2 - (\frac{15L^2}{\sigma_{\min}^A\rho b} + \frac{L}{b})\|x_M^s - \tilde{x}^s\|_2^2 \\
&\leq \Gamma_M^s - \sigma_{\min}^H\sum_{j=1}^m\|y_j^{s,M} - y_j^{s+1,1}\|^2 - \frac{L}{2b}\|x_M^s - \tilde{x}^s\|_2^2 - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - c_1\big)\|x_1^{s+1} - x_M^s\|^2 \\
&= \Gamma_M^s - \sigma_{\min}^H\sum_{j=1}^m\|y_j^{s,M} - y_j^{s+1,1}\|^2 - \frac{L}{2b}\|x_M^s - \tilde{x}^s\|_2^2 - \chi_M\|x_1^{s+1} - x_M^s\|^2, \qquad (100)
\end{aligned}
$$

where $c_M = \frac{18L^2}{\sigma_{\min}^A\rho b} + \frac{L}{b}$, and $\chi_M = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - c_1$.

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

$$
\begin{aligned}
c_{t+1} = (\frac{18L^2}{\sigma_{\min}^A\rho b} + \frac{L}{b})\frac{(1+\beta)^{M-t} - 1}{\beta} &= \frac{M}{b}(\frac{18L^2}{\sigma_{\min}^A\rho} + L)\big((1 + \frac{1}{M})^{M-t} - 1\big) \\
&\leq \frac{M}{b}(\frac{18L^2}{\sigma_{\min}^A\rho} + L)(e-1) \leq \frac{2M}{b}(\frac{18L^2}{\sigma_{\min}^A\rho} + L). \qquad (101)
\end{aligned}
$$

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 $t = 1, 2, \cdots, M$

$$
\begin{aligned}
\chi_t &\geq \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1 + 1/\beta)\frac{2M}{b}(\frac{18L^2}{\sigma_{\min}^A\rho} + L) \\
&= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1 + M)\frac{2M}{b}(\frac{18L^2}{\sigma_{\min}^A\rho} + L) \\
&\geq \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{4M^2}{b}(\frac{18L^2}{\sigma_{\min}^A\rho} + L) \\
&= \underbrace{\frac{\sigma_{\min}(G)}{\eta} - L - \frac{4M^2L}{b}}_{Q_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{72M^2L^2}{b\sigma_{\min}^A\rho}}_{Q_2}. \qquad (102)
\end{aligned}
$$

Let $M = [n^{\frac{1}{3}}]$, $b = [n^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{5L}$, we have $Q_1 \geq 0$. Further, set $\eta = \frac{\alpha\sigma_{\min}(G)}{5L}$ $(0 < \alpha \leq 1)$ and

$\rho = \frac{2\sqrt{231}\kappa_G L}{\sigma^A_{\min}\alpha}$, we have

$$
\begin{aligned}
Q_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{72M^2L^2}{b\sigma^A_{\min}\rho} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{150\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{72L^2}{\sigma^A_{\min}\rho} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{150\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9\kappa_G^2L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{72\kappa_G^2L^2}{\sigma^A_{\min}\rho\alpha^2} \\
&= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{231\kappa_G^2L^2}{\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\
&\geq \frac{\sqrt{231}\kappa_G L}{2\alpha} > 0
\end{aligned}
$$

where $\kappa_G = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)} \geq 1$. Thus, we have $\chi_t \geq \frac{\sqrt{231}\kappa_G L}{2\alpha} > 0$ for all t.

Since $\frac{L}{2b} > 0$ and $\chi_t > 0$, by (92) and (100), the function Γ^s_t is monotone decreasing. By the definition of function Γ^s_t, we have

$$
\begin{aligned}
\Gamma^s_t &\geq \mathbb{E}\big[\mathcal{L}_\rho(x^s_t, y^{s,t}_{[m]}, z^s_t)\big] \\
&= f(x^s_t) + \sum_{j=1}^m g(y^{s,t}_j) - (z^s_t)^T(Ax^s_t + \sum_{j=1}^m B_j y^{s,t}_j - c) + \frac{\rho}{2}\|Ax^s_t + \sum_{j=1}^m B_j y^{s,t}_j - c\| \\
&= f(x^s_t) + \sum_{j=1}^m g(y^{s,t}_j) - \frac{1}{\rho}(z^s_t)^T(z^s_{t-1} - z^s_t) + \frac{1}{2\rho}\|z^s_t - z^s_{t-1}\|^2 \\
&= f(x^s_t) + \sum_{j=1}^m g(y^{s,t}_j) - \frac{1}{2\rho}\|z^s_{t-1}\|^2 + \frac{1}{2\rho}\|z^s_t\|^2 + \frac{1}{\rho}\|z^s_t - z^s_{t-1}\|^2 \\
&\geq f^* + \sum_{j=1}^m g^*_j - \frac{1}{2\rho}\|z^s_{t-1}\|^2 + \frac{1}{2\rho}\|z^s_t\|^2.
\end{aligned} \tag{103}
$$

Summing the inequality (105) over $t = 0, 1\cdots, M$ and $s = 1, 2, \cdots, S$, we have

$$
\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^M \Gamma^s_t \geq f^* + \sum_{j=1}^m g^*_j - \frac{1}{2\rho}\|z^1_0\|^2. \tag{104}
$$

Thus, the function Γ^s_t is bounded from below. Let Γ^* denote a lower bound of Γ^s_t.

Finally, telescoping (92) and (100) over t from 0 to $M-1$ and over s from 1 to S, we have

$$
\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^{M-1}(\sigma^H_{\min}\sum_{j=1}^m \|y^{s,t}_j - y^{s,t+1}_j\|^2 + \frac{L}{2b}\|x^s_t - \tilde{x}^s\|^2_2 + \chi_t\|x^s_t - x^s_{t+1}\|^2) \leq \frac{\Gamma^1_0 - \Gamma^*}{T}. \tag{105}
$$

where $T = MS$ and $\chi_t \geq \frac{\sqrt{231}\kappa_G L}{2\alpha} > 0$.

□

Theorem 3. *Suppose the sequence $\{(x^s_t, y^{s,t}_{[m]}, z^s_t)^M_{t=1}\}^S_{s=1}$ is generated from Algorithm 3, and let $\eta = \frac{\alpha\sigma_{\min}(G)}{5L}$ $(0 < \alpha \leq 1)$, $\rho = \frac{2\sqrt{231}\kappa_G L}{\sigma^A_{\min}\alpha}$ and*

$$
\nu_1 = m\big(\rho^2\sigma^B_{\max}\sigma^A_{\max} + \rho^2(\sigma^B_{\max})^2 + \sigma^2_{\max}(H)\big),\ \nu_2 = 3L^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2},\ \nu_3 = \frac{9L^2}{\sigma^A_{\min}\rho^2} + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2}.
$$

then we have

$$\min_{s,t} \mathbb{E}\big[dist(0, \partial L(x_t^s, y_{[m]}^{s,t}, z_t^s))^2\big] \leq \frac{\nu_{\max}}{T}\sum_{s=1}^{S}\sum_{t=0}^{M-1}\theta_t^s \leq \frac{2\nu_{\max}(\Gamma_0^1 - \Gamma^*)}{\gamma T}$$

where $\gamma = \min(\sigma_{\min}^H, \frac{L}{2}, \chi_t)$*,* $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ *and* Γ^* *is a lower bound of function* Γ_t^s*. It implies that the whole iteration number* $T = MS$ *satisfies*

$$T = \frac{2\nu_{\max}(\Gamma_0^1 - \Gamma^*)}{\epsilon\gamma},$$

then $(x_{t^*}^{s^*}, y_{[m]}^{s^*,t^*}, z_{t^*}^{s^*})$ *is an* ϵ*-approximate stationary point of* (1)*, where* $(t^*, s^*) = \arg\min_{t,s}\theta_t^s$*.*

Proof. We begin with defining an useful variable $\theta_t^s = \mathbb{E}\big[\|x_{t+1}^s - x_t^s\|^2 + \|x_t^s - x_{t-1}^s\|^2 + \frac{1}{b}(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2) + \sum_{j=1}^m \|y_j^{s,t} - y_j^{s,t+1}\|^2\big]$. By the step 7 of Algorithm 3, we have, for all $j \in [m]$

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \partial_{y_j} L(x, y_{[m]}, z))^2\big]_{s,t+1} &= \mathbb{E}\big[\text{dist}(0, \partial g_j(y_j^{s,t+1}) - B_j^T z_{t+1}^s)^2\big] \\
&= \|B_j^T z_t^s - \rho B_j^T(Ax_t^s + \sum_{i=1}^{j} B_i y_i^{s,t+1} + \sum_{i=j+1}^{m} B_i y_i^{s,t} - c) - H_j(y_j^{s,t+1} - y_j^{s,t}) - B_j^T z_{t+1}^s\|^2 \\
&= \|\rho B_j^T A(x_{t+1}^s - x_t^s) + \rho B_j^T \sum_{i=j+1}^{m} B_i(y_i^{s,t+1} - y_i^{s,t}) - H_j(y_j^{s,t+1} - y_j^{s,t})\|^2 \\
&\leq m\rho^2\sigma_{\max}^{B_j}\sigma_{\max}^{A}\|x_{t+1}^s - x_t^s\|^2 + m\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^{m}\sigma_{\max}^{B_i}\|y_i^{s,t+1} - y_i^{s,t}\|^2 \\
&\quad + m\sigma_{\max}^2(H_j)\|y_j^{s,t+1} - y_j^{s,t}\|^2 \\
&\leq m\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big)\theta_t^s, \qquad (106)
\end{aligned}$$

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

By the step 8 of Algorithm 3, we have

$$\begin{aligned}
\mathbb{E}[\text{dist}(0, \nabla_x L(x, y, z))]_{s,t+1} &= \mathbb{E}\|A^T z_{t+1}^s - \nabla f(x_{t+1}^s)\|^2 \\
&= \mathbb{E}\|v_t^s - \nabla f(x_{t+1}^s) - \frac{G}{\eta}(x_t^s - x_{t+1}^s)\|^2 \\
&= \mathbb{E}\|v_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) - \nabla f(x_{t+1}^s) - \frac{G}{\eta}(x_t^s - x_{t+1}^s)\|^2 \\
&\leq \frac{3L^2}{b}\|x_t^s - \tilde{x}^s\|^2 + 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\|x_t^s - x_{t+1}^s\|^2 \\
&\leq \big(3L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\big)\theta_t^s. \qquad (107)
\end{aligned}$$

By the step 9 of Algorithm 3, we have

$$\begin{aligned}\mathbb{E}[\text{dist}(0,\nabla_z L(x,y,z))]_{s,t+1} &= \mathbb{E}\|Ax^s_{t+1}+By^s_{t+1}-c\|^2 \\ &= \frac{1}{\rho^2}\mathbb{E}\|z^s_{t+1}-z^s_t\|^2 \\ &\leq \frac{9L^2}{\sigma^A_{\min}\rho^2 b}\big(\|x^s_t-\tilde{x}^s\|^2+\|x^s_{t-1}-\tilde{x}^s\|^2\big)+\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2}\|x^s_{t+1}-x^s_t\|^2 \\ &\quad + \frac{3(\sigma^2_{\max}(G)+3L^2\eta^2)}{\sigma^A_{\min}\eta^2\rho^2}\|x^s_t-x^s_{t-1}\|^2 \\ &\leq \big(\frac{9L^2}{\sigma^A_{\min}\rho^2}+\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2}\big)\theta^s_t.\end{aligned} \tag{108}$$

Let

$$\nu_1 = m\big(\rho^2\sigma^B_{\max}\sigma^A_{\max}+\rho^2(\sigma^B_{\max})^2+\sigma^2_{\max}(H)\big),\ \nu_2 = 3L^2+\frac{3\sigma^2_{\max}(G)}{\eta^2},\ \nu_3=\frac{9L^2}{\sigma^A_{\min}\rho^2}+\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho^2}.$$

Using (105), (106), (107) and (108), we have

$$\min_{s,t}\mathbb{E}\big[\text{dist}(0,\partial L(x^s_t,y^{s,t}_{[m]},z^s_t))^2\big] \leq \frac{\nu_{\max}}{T}\sum_{s=1}^{S}\sum_{t=0}^{M-1}\theta^s_t \leq \frac{2\nu_{\max}(\Gamma^1_0-\Gamma^*)}{\gamma T}, \tag{109}$$

where $\gamma=\min(\sigma^H_{\min},\frac{L}{2},\chi_t)$ with $\chi_t \geq \frac{\sqrt{231}\kappa_G L}{2\alpha} > 0$and $\nu_{\max}=\max(\nu_1,\nu_2,\nu_3)$.

Given $\eta=\frac{\alpha\sigma_{\min}(G)}{5L}$ $(0<\alpha\leq 1)$ and $\rho=\frac{2\sqrt{231}\kappa_G L}{\sigma^A_{\min}\alpha}$, since m is relatively small, it easy verifies that $\nu_{\max}=O(1)$ and $\gamma=O(1)$, which are independent on n and T. Thus, we obtain

$$\min_{t,s}\mathbb{E}\big[\text{dist}(0,\partial L(x^s_t,y^{s,t}_{[m]},z^s_t))^2\big] \leq O(\frac{1}{T}). \tag{110}$$

□

2.4. Theoretical Analysis of Non-convex SAGA-ADMM Algorithm

In the subsection, we first extend the existing nonconvex SAGA-ADM (Huang et al., 2016) to the multi-blocks setting for solving the problem (1), which is summarized in Algorithm 4. Then we study the convergence analysis of this non-convex SAGA-ADMM.

Algorithm 4 Nonconvex SAGA-ADMM Alogrithm

1: **Input:** $b, T, \eta > 0, \rho > 0$;
2: **Initialize:** x_0, $u_i^0 = x_0$ for $i \in \{1, 2, \cdots, n\}$, $\phi_0 = \frac{1}{n}\sum_{i=1}^n \nabla f_i(u_i^0)$, and y_j^0 for $j \in [m]$;
3: **for** $t = 0, 1, \cdots, T-1$ **do**
4: Uniformly random pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n\}$ with $|\mathcal{I}_t| = b$, and compute

$$v_t = \frac{1}{b}\sum_{i_t \in \mathcal{I}_t} \big(\nabla f_{i_t}(x_t) - \nabla f_{i_t}(u_{i_t}^t)\big) + \phi_t$$

with $\phi_t = \frac{1}{n}\sum_{i=1}^n \nabla f_i(u_i^t)$;
5: $y_j^{t+1} = \arg\min_{y_j} \big\{\mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_j, y_{[j+1:m]}^t, z_t) + \frac{1}{2}\|y_j - y_j^t\|^2_{H_j}\big\}$ with $H_j \succ 0$ for all $j \in [m]$;
6: $x_{t+1} = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[m]}^{t+1}, z_t, v_t\big)$;
7: $z_{t+1} = z_t - \rho(Ax_{t+1} + \sum_{j=1}^m B_j y_j^{t+1} - c)$;
8: $u_{i_t}^{t+1} = x_t$ for $i \in \mathcal{I}_t$ and $u_i^{t+1} = u_i^t$ for $i \notin \mathcal{I}_t$;
9: $\phi_{t+1} = \phi_t - \frac{1}{n}\sum_{i_t \in \mathcal{I}_t}\big(\nabla f_{i_t}(u_{i_t}^t) - \nabla f_{i_t}(u_{i_t}^{t+1})\big)$;
10: **end for**
11: **Output:** $\{x, y_{[m]}, z\}$ chosen uniformly random from $\{x_t, y_{[m]}^t, z_t\}_{t=1}^T$.

In Algorithm 4, we give

$$\begin{aligned}\hat{\mathcal{L}}_\rho(x, y_{[m]}^{t+1}, z_t, v_t) =& f(x_t) + v_t^T(x - x_t) + \frac{1}{2\eta}\|x - x_t\|^2_G + \sum_{j=1}^m g_j(y_j^{t+1}) \\ & - z_t^T(Ax + \sum_{j=1}^m B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax + \sum_{j=1}^m B_j y_j^{t+1} - c\|^2, \end{aligned} \tag{111}$$

where $\eta > 0$ and $G \succ 0$.

Lemma 8. *Suppose the sequence $\{x_t, y_{[m]}^t, z_t\}_{t=1}^T$ is generated by Algorithm 4. The following inequality holds*

$$\begin{aligned}\mathbb{E}\|z_{t+1} - z_t\|^2 \leq & \frac{9L^2}{\sigma^A_{\min} b}\frac{1}{n}\sum_{i=1}^n \big(\|x_t - u_i^t\|^2 + \|x_{t-1} - u_i^{t-1}\|^2\big) + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2}\mathbb{E}\|x_{t+1} - x_t\|^2 \\ & + \frac{3(\sigma^2_{\max}(G) + 3L^2\eta^2)}{\sigma^A_{\min}\eta^2}\mathbb{E}\|x_t - x_{t-1}\|^2. \end{aligned} \tag{112}$$

Proof. By the optimize condition of the the step 6 in Algorithm 4, we have

$$v_t + \frac{1}{\eta}G(x_{t+1} - x_t) - A^T z_t + \rho A^T(Ax_{t+1} + \sum_{j=1}^m B_j y_j^{t+1} - c) = 0. \tag{113}$$

Using the step 7 of Algorithm 4, then we have

$$A^T z_{t+1} = v_t + \frac{G}{\eta}(x_{t+1} - x_t). \tag{114}$$

It follows that

$$z_{t+1} = (A^T)^+\big(v_t + \frac{G}{\eta}(x_{t+1} - x_t)\big), \tag{115}$$

where $(A^T)^+$ is the pseudoinverse of A^T. Since A is a full column matrix, we have $(A^T)^+ = A(A^TA)^{-1}$. Using (115), then we have

$$\mathbb{E}\|z_{t+1}-z_t\|^2 = \mathbb{E}\|(A^T)^+\big(v_t + \frac{G}{\eta}(x_{t+1}-x_t) - v_{t-1} + \frac{G}{\eta}(x_t - x_{t-1})\big)\|^2$$
$$\leq \frac{1}{\sigma^A_{\min}}\big[3\mathbb{E}\|v_t - v_{t-1}\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\mathbb{E}\|x_{t+1}-x_t\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\mathbb{E}\|x_t - x_{t-1}\|^2\big]. \tag{116}$$

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

$$\mathbb{E}\|v_t - v_{t-1}\|^2 = \mathbb{E}\|v_t - \nabla f(x_t) + \nabla f(x_t) - \nabla f(x_{t-1}) + \nabla f(x_{t-1}) - v_{t-1}\|^2$$
$$\leq 3\mathbb{E}\|v_t - \nabla f(x_t)\|^2 + 3\mathbb{E}\|\nabla f(x_t) - \nabla f(x_{t-1})\|^2 + 3\mathbb{E}\|\nabla f(x_{t-1}) - v_{t-1}\|^2$$
$$\leq \frac{3L^2}{b}\frac{1}{n}\sum_{i=1}^n\big(\|x_t - u_i^t\|^2 + \|x_{t-1} - u_i^{t-1}\|^2\big) + 3L^2\mathbb{E}\|x_t - x_{t-1}\|^2, \tag{117}$$

where the second inequality holds by lemma 4 of (Reddi et al., 2016) and Assumption 1. Finally, combining the inequalities (116) with (117), we can obtain the above result. □

Lemma 9. *Suppose the sequence $\{x_t, y^t_{[m]}, z_t\}_{t=1}^T$ is generated from Algorithm 4, and define a* Lyapunov *function*

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y^t_{[m]}, z_t) + (\frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\rho\eta^2} + \frac{9L^2}{\sigma^A_{\min}\rho})\|x_t - x_{t-1}\|^2 + \frac{9L^2}{\sigma^A_{\min}\rho b}\frac{1}{n}\sum_{i=1}^n\|x_{t-1} - u_i^{t-1}\|^2 + c_t\frac{1}{n}\sum_{i=1}^n\|x_t - u_i^t\|^2\big],$$

where the positive sequence $\{c_t\}$ satisfies

$$c_t = \begin{cases} \frac{18L^2}{\sigma^A_{\min}\rho b} + \frac{L}{b} + (1-p)(1+\beta)c_{t+1},\ 0 \leq t \leq T-1, \\ 0,\ t \geq T, \end{cases}$$

where p denotes probability of an index i being in $\mathcal{I}_t$. Further, let $b = [n^{\frac{2}{3}}]$, $\eta = \frac{\alpha\sigma_{\min}(G)}{17L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{2031}\kappa_G}{\sigma^A_{\min}\alpha}$ we have

$$\frac{1}{T}\sum_{t=1}^T(\sigma^H_{\min}\sum_{j=1}^m\|y_j^t - y_j^{t+1}\|^2 + \chi_t\|x_t - x_{t+1}\|^2 + \frac{L}{2b}\frac{1}{n}\sum_{i=1}^n\|x_t - u_i^t\|^2) \leq \frac{\Omega_0 - \Omega^*}{T}, \tag{118}$$

where $\chi_t \geq \frac{\sqrt{2031}\kappa_G L}{2\alpha} > 0$ and Ω^* denotes a lower bound of Ω_t.

Proof. By the optimal condition of step 5 in Algorithm 4, we have, for $j \in [m]$

$$
\begin{aligned}
0 &= (y_j^t - y_j^{t+1})^T \big(\partial g_j(y_j^{t+1}) - B_j^T z_t + \rho B_j^T (Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) + H_j(y_j^{t+1} - y_j^t)\big) \\
&\leq g_j(y_j^t) - g_j(y_j^{t+1}) - (z_t)^T (B_j y_j^t - B_j y_j^{t+1}) + \rho (B_j y_j^t - B_j y_j^{t+1})^T (Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) - \|y_j^{t+1} - y_j^t\|_{H_j}^2 \\
&= g_j(y_j^t) - g_j(y_j^{t+1}) - (z_t)^T (Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{m} B_i y_i^k - c) + (z_t)^T (Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) \\
&\quad + \frac{\rho}{2} \|Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{m} B_i y_i^k - c\|^2 - \frac{\rho}{2} \|Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\|^2 - \frac{\rho}{2} \|B_j y_j^t - B_j y_j^{t+1}\|^2 - \|y_j^{t+1} - y_j^t\|_{H_j}^2 \\
&= \underbrace{f(x_t) + \sum_{i=1}^{j-1} g_i(y_i^{t+1}) + \sum_{i=j}^{m} g_i(y_i^t) - (z_t)^T (Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{m} B_i y_i^k - c) + \frac{\rho}{2} \|Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^{m} B_i y_i^k - c\|^2}_{\mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:m]}^t, z_t)} \\
&\quad - \underbrace{\big(f(x_t) + \sum_{i=1}^{j} g_i(y_i^{t+1}) + \sum_{i=j+1}^{m} g_i(y_i^t) - (z_t)^T (Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) + \frac{\rho}{2} \|Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\|^2\big)}_{\mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:m]}^t, z_t)} \\
&\quad - \frac{\rho}{2} \|B_j y_j^t - B_j y_j^{t+1}\|^2 - \|y_j^{t+1} - y_j^t\|_{H_j}^2 \\
&\leq \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:m]}^t, z_t) - \mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:m]}^t, z_t) - \sigma_{\min}(H_j) \|y_j^t - y_j^{t+1}\|^2, \qquad (119)
\end{aligned}
$$

where the first inequality holds by the convexity of function $g_j(y)$, 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 $(By_j^t - By_j^{t+1})^T (Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^k - c)$. Thus, we have, for all $j \in [m]$

$$
\mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:m]}^t, z_t) \leq \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:m]}^t, z_t) - \sigma_{\min}(H_j) \|y_j^t - y_j^{t+1}\|^2. \qquad (120)
$$

Telescoping inequality (120) over j from 1 to m, we obtain

$$
\mathcal{L}_\rho(x_t, y_{[m]}^{t+1}, z_t) \leq \mathcal{L}_\rho(x_t, y_{[m]}^t, z_t) - \sigma_{\min}^H \sum_{j=1}^{m} \|y_j^t - y_j^{t+1}\|^2. \qquad (121)
$$

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

Using Assumption 1, we have

$$
0 \leq f(x_t) - f(x_{t+1}) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \|x_{t+1} - x_t\|^2. \qquad (122)
$$

By the step 6 of Algorithm 4, we have

$$
0 = (x_t - x_{t+1})^T \big(v_t - A^T z_t + \rho A^T (Ax_{t+1} + \sum_{j=1}^{m} B_j y_j^{t+1} - c) + \frac{G}{\eta}(x_{t+1} - x_t)\big). \qquad (123)
$$

Combining (122) and (123), we have

$$
\begin{aligned}
0 \leq & 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 z_t + \rho A^T (Ax_{t+1} + \sum_{j=1}^m 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)) \\
& - (z_t)^T (Ax_t - Ax_{t+1}) + \rho (Ax_t - Ax_{t+1})^T (Ax_t + \sum_{j=1}^m 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)) - (z_t)^T (Ax_t + \sum_{j=1}^m B_j y_j^{t+1} - c) \\
& + (z_t)^T (Ax_{t+1} + \sum_{j=1}^m B_j y_j^{t+1} - c) + \frac{\rho}{2}\big(\|Ax_t + \sum_{j=1}^m B_j y_j^{t+1} - c\|^2 - \|Ax_{t+1} + \sum_{j=1}^m B_j y_j^{t+1} - c\|^2 - \|Ax_t - Ax_{t+1}\|^2\big) \\
= & \underbrace{f(x_t) + \sum_{j=1}^m g_j(y_j^{t+1}) - (z_t)^T (Ax_t + \sum_{j=1}^m B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax_t + \sum_{j=1}^m B_j y_j^{t+1} - c\|^2}_{\mathcal{L}_\rho(x_t, y_{[m]}^{t+1}, z_t)} \\
& - \underbrace{\big(f(x_{t+1}) + \sum_{j=1}^m g_j(y_j^{t+1}) - (z_t)^T (Ax_{t+1} + \sum_{j=1}^m B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^m B_j y_j^{t+1} - c\|^2\big)}_{\mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_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 \\
\leq & \mathcal{L}_\rho(x_t, y_{[m]}^{t+1}, z_t) - \mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_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)}{\leq} & \mathcal{L}_\rho(x_t, y_{[m]}^{t+1}, z_t) - \mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_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)}{\leq} & \mathcal{L}_\rho(x_t, y_{[m]}^{t+1}, z_t) - \mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_t) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 + \frac{L}{2b}\frac{1}{n}\sum_{i=1}^n \|x_t - u_i^t\|^2,
\end{aligned} \tag{124}
$$

where the equality (i) holds by applying the equality $(a-b)^T b = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ on the term $(Ax_t - Ax_{t+1})^T (Ax_{t+1} + \sum_{j=1}^m B_j y_j^{t+1} - c)$; the inequality (ii) follows by the inequality $a^T b \leq \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|a\|^2$, and the inequality (iii) holds by Lemma 4 of (Reddi et al., 2016). Thus, we obtain

$$
\begin{aligned}
\mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_t) \leq & \mathcal{L}_\rho(x_t, y_{[m]}^{t+1}, z_t) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 \\
& + \frac{L}{2b}\frac{1}{n}\sum_{i=1}^n \|x_t - u_i^t\|^2.
\end{aligned} \tag{125}
$$

By the step 7 in Algorithm 4, we have

$$
\begin{aligned}
\mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_{t+1}) - \mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_t) = & \frac{1}{\rho}\|z_{t+1} - z_t\|^2 \\
\leq & \frac{9L^2}{\sigma_{\min}^A \rho b}\frac{1}{n}\sum_{i=1}^n \big(\|x_t - u_i^t\|^2 + \|x_{t-1} - u_i^{t-1}\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x_{t+1} - x_t\|^2 \\
& + \frac{3(\sigma_{\max}^2(G) + 3L^2\eta^2)}{\sigma_{\min}^A \eta^2 \rho}\|x_t - x_{t-1}\|^2,
\end{aligned} \tag{126}
$$

where the first inequality follows by Lemma 8.

Combining (121), (125) and (126), we have

$$
\begin{aligned}
\mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_{t+1}) \leq \mathcal{L}_\rho(x_t, y_{[m]}^{t}, z_t) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 + \frac{L}{2b}\frac{1}{n}\sum_{i=1}^n \|x_t - u_i^t\|^2 \\
- \sigma_{\min}^H \sum_{j=1}^m \|y_j^t - y_j^{t+1}\|^2 + \frac{9L^2}{\sigma_{\min}^A \rho b}\frac{1}{n}\sum_{i=1}^n \big(\|x_t - u_i^t\|^2 + \|x_{t-1} - u_i^{t-1}\|^2\big) \\
+ \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x_{t+1} - x_t\|^2 + \frac{3(\sigma_{\max}^2(G) + 3L^2\eta^2)}{\sigma_{\min}^A \eta^2 \rho}\|x_t - x_{t-1}\|^2. \qquad (127)
\end{aligned}
$$

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

$$
\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[m]}^t, z_t) + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \rho \eta^2} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + \frac{9L^2}{\sigma_{\min}^A \rho b}\frac{1}{n}\sum_{i=1}^n \|x_{t-1} - z_i^{t-1}\|^2 + \frac{c_t}{n}\sum_{i=1}^n \|x_t - z_i^t\|^2\big]. \qquad (128)
$$

By the step 9 of Algorithm 4, we have

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|x_{t+1} - u_i^{t+1}\|^2 &= \frac{1}{n}\sum_{i=1}^n \big(p\|x_{t+1} - x_t\|^2 + (1-p)\|x_{t+1} - u_i^t\|^2\big) \\
&= \frac{p}{n}\sum_{i=1}^n \|x_{t+1} - x_t\|^2 + \frac{1-p}{n}\sum_{i=1}^n \|x_{t+1} - u_i^t\|^2 \\
&= p\|x_{t+1} - x_t\|^2 + \frac{1-p}{n}\sum_{i=1}^n \|x_{t+1} - u_i^t\|^2, \qquad (129)
\end{aligned}
$$

where p denotes probability of an index i being in $\mathcal{I}_t$. Here, we have

$$
p = 1 - (1 - \frac{1}{n})^b \geq 1 - \frac{1}{1 + b/n} = \frac{b/n}{1 + b/n} \geq \frac{b}{2n}, \qquad (130)
$$

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

$$
\begin{aligned}
\|x_{t+1} - u_i^t\|^2 &= \|x_{t+1} - x_t + x_t - u_i^t\|^2 \\
&= \|x_{t+1} - x_t\|^2 + 2(x_{t+1} - x_t)^T(x_t - u_i^t) + \|x_t - u_i^t\|^2 \\
&\leq \|x_{t+1} - x_t\|^2 + 2\big(\frac{1}{2\beta}\|x_{t+1} - x_t\|^2 + \frac{\beta}{2}\|x_t - u_i^t\|^2\big) + \|x_t - u_i^t\|^2 \\
&= (1 + \frac{1}{\beta})\|x_{t+1} - x_t\|^2 + (1+\beta)\|x_t - u_i^t\|^2, \qquad (131)
\end{aligned}
$$

where $\beta > 0$. Combining (129) with (131), we have

$$
\frac{1}{n}\sum_{i=1}^n \|x_{t+1} - u_i^{t+1}\|^2 \leq (1 + \frac{1-p}{\beta})\|x_{t+1} - x_t\|^2 + \frac{(1-p)(1+\beta)}{n}\sum_{i=1}^n \|x_t - u_i^t\|^2. \qquad (132)
$$

It follows that

$$
\begin{aligned}
\Omega_{t+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_{t+1}, y_{[m]}^{t+1}, z_{t+1}) + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_{t+1}-x_t\|^2 + \frac{9L^2}{\sigma_{\min}^A\rho b}\frac{1}{n}\sum_{i=1}^n\|x_t-u_i^t\|^2 + \frac{c_{t+1}}{n}\sum_{i=1}^n\|x_{t+1}-u_i^{t+1}\|^2\big] \\
&\leq \mathcal{L}_\rho(x_t, y_{[m]}^t, z_t) + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t-x_{t-1}\|^2 + \frac{9L^2}{\sigma_{\min}^A\rho b}\frac{1}{n}\sum_{i=1}^n\|x_{t-1}-u_i^{t-1}\|^2 \\
&\quad + \big(\frac{18L^2}{\sigma_{\min}^A\rho b} + \frac{L}{b} + (1-p)(1+\beta)c_{t+1}\big)\frac{1}{n}\sum_{i=1}^n\|x_t-u_i^t\|^2 - \sigma_{\min}^H\sum_{j=1}^m\|y_j^t-y_j^{t+1}\|^2 \\
&\quad - \frac{L}{2b}\frac{1}{n}\sum_{i=1}^n\|x_t-u_i^t\|^2 - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+\frac{1-p}{\beta})c_{t+1}\big)\|x_t-x_{t+1}\|^2 \\
&= \Omega_t - \sigma_{\min}^H\sum_{j=1}^m\|y_j^t-y_j^{t+1}\|^2 - \chi_t\|x_t-x_{t+1}\|^2 - \frac{L}{2b}\frac{1}{n}\sum_{i=1}^n\|x_t-u_i^t\|^2,
\end{aligned} \tag{133}
$$

where $c_t = \frac{18L^2}{\sigma_{\min}^A\rho b} + \frac{L}{b} + (1-p)(1+\beta)c_{t+1}$ and $\chi_t = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+\frac{1-p}{\beta})c_{t+1}$.
Let $c_T = 0$ and $\beta = \frac{b}{4n}$. Since $(1-p)(1+\beta) = 1+\beta-p-p\beta \leq 1+\beta-p$ and $p \geq \frac{b}{2n}$, it follows that

$$
c_t \leq c_{t+1}(1-\theta) + \frac{18L^2}{\sigma_{\min}^A\rho b} + \frac{L}{b}, \tag{134}
$$

where $\theta = p - \beta \geq \frac{b}{4n}$. Then recursing on t, for $0 \leq t \leq T-1$, we have

$$
c_t \leq \frac{1}{b}(\frac{18L^2}{\sigma_{\min}^A\rho} + L)\frac{1-\theta^{T-t}}{\theta} \leq \frac{1}{b\theta}(\frac{18L^2}{\sigma_{\min}^A\rho} + L) \leq \frac{4n}{b^2}(\frac{18L^2}{\sigma_{\min}^A\rho} + L). \tag{135}
$$

It follows that

$$
\begin{aligned}
\chi_t &= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+\frac{1-p}{\beta})c_{t+1} \\
&\geq \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+\frac{4n-2b}{b})\frac{4n}{b^2}(\frac{18L^2}{\sigma_{\min}^A\rho} + L) \\
&= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (\frac{4n}{b}-1)\frac{4n}{b^2}(\frac{18L^2}{\sigma_{\min}^A\rho} + L) \\
&\geq \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{16n^2}{b^3}(\frac{18L^2}{\sigma_{\min}^A\rho} + L) \\
&= \underbrace{\frac{\sigma_{\min}(G)}{\eta} - L - \frac{16n^2L}{b^3}}_{Q_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{288n^2L^2}{\sigma_{\min}^A\rho b^3}}_{Q_2}
\end{aligned} \tag{136}
$$

Let $b = [n^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{17L}$, we have $Q_1 \geq 0$. Further, let $\eta = \frac{\alpha\sigma_{\min}(G)}{17L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{2031}\kappa_G L}{\sigma_{\min}^A\alpha}$, we have

$$
\begin{aligned}
Q_2 &= \frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{288n^2L^2}{\sigma_{\min}^A\rho b^3} \\
&= \frac{\rho\sigma_{\min}^A}{2} - \frac{1734\kappa_G^2L^2}{\sigma_{\min}^A\alpha^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{288L^2}{\sigma_{\min}^A\rho} \\
&\geq \frac{\rho\sigma_{\min}^A}{4} + \underbrace{\frac{\rho\sigma_{\min}^A}{4} - \frac{2031\kappa_G^2L^2}{\sigma_{\min}^A\alpha^2\rho}}_{\geq 0} \\
&\geq \frac{\sqrt{2031}\kappa_G L}{2\alpha},
\end{aligned} \tag{137}
$$

where $\kappa_G \geq 1$. Thus, we have $\chi_t \geq \frac{\sqrt{2031}\kappa_G L}{2\alpha}$ for all t.

Since $\frac{L}{2b} > 0$ and $\chi_t > 0$, by (133), the function Ω_t is monotone decreasing. By the definition of function Ω_t, we have

$$\begin{aligned}
\Omega_t &\geq \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[m]}^t, z_t)\big] \\
&= f(x_t) + \sum_{j=1}^m g_j(y_j^t) - (z_t)^T(Ax_t + \sum_{j=1}^m B_j y_j^t - c) + \frac{\rho}{2}\|Ax_t + \sum_{j=1}^m B_j y_j^t - c\| \\
&= f(x_t) + \sum_{j=1}^m g_j(y_j^t) - \frac{1}{\rho}(z_t)^T(z_{t-1} - z_t) + \frac{1}{2\rho}\|z_t - z_{t-1}\|^2 \\
&= f(x_t) + \sum_{j=1}^m g_j(y_j^t) - \frac{1}{2\rho}\|z_{t-1}\|^2 + \frac{1}{2\rho}\|z_t\|^2 + \frac{1}{\rho}\|z_t - z_{t-1}\|^2 \\
&\geq f^* + \sum_{j=1}^m g_j^* - \frac{1}{2\rho}\|z_{t-1}\|^2 + \frac{1}{2\rho}\|z_t\|^2. \qquad (138)
\end{aligned}$$

Summing the inequality (138) over $t = 0, 1 \cdots, T$, we have

$$\frac{1}{T}\sum_{t=0}^T \Omega_t \geq f^* + \sum_{j=1}^m g_j^* - \frac{1}{2\rho}\|z_0\|^2. \qquad (139)$$

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

Finally, telescoping inequality (133) over t from 0 to T, we have

$$\frac{1}{T}\sum_{t=1}^T (\sigma_{\min}^H \sum_{j=1}^m \|y_j^t - y_j^{t+1}\|^2 + \chi_t\|x_t - x_{t+1}\|^2 + \frac{L}{2b}\frac{1}{n}\sum_{i=1}^n \|x_t - u_i^t\|^2) \leq \frac{\Omega_0 - \Omega^*}{T}, \qquad (140)$$

where $\chi_t \geq \frac{\sqrt{2031}\kappa_G L}{2\alpha} > 0$.

□

Theorem 4. *Suppose the sequence $\{x_t, y_{[m]}^t, z_t\}_{t=1}^T$ is generated from Algorithm 4, and let $b = [n^{\frac{2}{3}}]$, $\eta = \frac{\alpha\sigma_{\min}(G)}{17L}$ $(0 < \alpha \leq 1)$, $\rho = \frac{2\sqrt{2031}\kappa_G}{\sigma_{\min}^A \alpha}$ and*

$$\begin{aligned}
&\nu_1 = m\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big), \ \nu_2 = 3L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2} \\
&\nu_3 = \frac{9L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2},
\end{aligned}$$

then we have

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

where $\gamma = \min(\sigma_{\min}^H, L/2, \chi_t)$ with $\chi_t \geq \frac{\sqrt{2031}\kappa_G L}{2\alpha} > 0$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ and Ω^* is a lower bound of function Ω_t. It implies that the iteration number T satisfies

$$T = \frac{2\nu_{\max}}{\epsilon\gamma}(\Omega_0 - \Omega^*),$$

then $(x_{t^*}, y_{[m]}^{t^*}, z_{t^*})$ is an ϵ-approximate stationary point of (1), where $t^* = \arg\min_{1\leq t\leq T}\theta_t$.

Proof. We first define an useful variable $\theta_t = \mathbb{E}\big[\|x_{t+1}-x_t\|^2 + \|x_t - x_{t-1}\|^2 + \frac{1}{bn}\sum_{i=1}^n(\|x_t - u_i^t\|^2 + \|x_{t-1}-u_i^{t-1}\|^2) + \sum_{j=1}^m \|y_j^t - y_j^{t+1}\|^2\big]$. By the optimal condition of the step 5 in Algorithm 4, we have, for all $j \in [m]$

$$
\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \partial_{y_j} L(x, y_{[m]}, z))^2\big]_{t+1} &= \mathbb{E}\big[\text{dist}(0, \partial g_j(y_j^{t+1}) - B_j^T z_{t+1})^2\big] \\
&= \|B_j^T z_t - \rho B_j^T (Ax_t + \sum_{i=1}^{j} B_i y_i^{t+1} + \sum_{i=j+1}^{m} B_i y_i^t - c) - H_j(y_j^{t+1} - y_j^k) - B_j^T z_{t+1}\|^2 \\
&= \|\rho B_j^T A(x_{t+1} - x_t) + \rho B_j^T \sum_{i=j+1}^{m} B_i (y_i^{t+1} - y_i^t) - H_j(y_j^{t+1} - y_j^k)\|^2 \\
&\le m\rho^2 \sigma_{\max}^{B_j}\sigma_{\max}^A \|x_{t+1} - x_t\|^2 + m\rho^2\sigma_{\max}^{B_j} \sum_{i=j+1}^{m} \sigma_{\max}^{B_i}\|y_i^{t+1} - y_i^t\|^2 + m\sigma_{\max}^2(H_j)\|y_j^{t+1} - y_j^k\|^2 \\
&\le m\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big)\theta_t, \qquad (141)
\end{aligned}
$$

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 6 in Algorithm 4, we have

$$
\begin{aligned}
\mathbb{E}[\text{dist}(0, \nabla_x L(x, y_{[m]}, z))]_{t+1} &= \mathbb{E}\|A^T z_{t+1} - \nabla f(x_{t+1})\|^2 \\
&= \mathbb{E}\|v_t - \nabla f(x_{t+1}) - \frac{G}{\eta}(x_t - x_{t+1})\|^2 \\
&= \mathbb{E}\|v_t - \nabla f(x_t) + \nabla f(x_t) - \nabla f(x_{t+1}) - \frac{G}{\eta}(x_t - x_{t+1})\|^2 \\
&\le \frac{3L^2}{bn}\sum_{i=1}^n \|x_t - u_i^t\|^2 + 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\|x_t - x_{t+1}\|^2 \\
&\le \big(3L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\big)\theta_t. \qquad (142)
\end{aligned}
$$

By the step 7 of Algorithm 4, we have

$$
\begin{aligned}
\mathbb{E}[\text{dist}(0, \nabla_z L(x, y_{[m]}, z))]_{t+1} &= \mathbb{E}\|Ax_{t+1} + By_{t+1} - c\|^2 \\
&= \frac{1}{\rho^2}\mathbb{E}\|z_{t+1} - z_t\|^2 \\
&\le \frac{9L^2}{\sigma_{\min}^A \rho^2 b}\frac{1}{n}\sum_{i=1}^n \big(\|x_t - u_i^t\|^2 + \|x_{t-1} - u_i^{t-1}\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2\rho^2}\|x_{t+1} - x_t\|^2 \\
&\quad + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2} + \frac{9L^2)}{\sigma_{\min}^A\rho^2})\|x_t - x_{t-1}\|^2 \\
&\le (\frac{9L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2})\theta_t. \qquad (143)
\end{aligned}
$$

Let

$$
\nu_1 = m\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big),\ \nu_2 = 6L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2},\ \nu_3 = \frac{9L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}.
$$

Using (140), (141), (142) and (143), we have

$$
\min_{1\le t\le T} \mathbb{E}\big[\text{dist}(0, \partial L(x_t, y_{[m]}^t, z_t))^2\big] \le \frac{\nu_{\max}}{T}\sum_{t=1}^T \theta_t \le \frac{2\nu_{\max}(\Omega_0 - \Omega^*)}{\gamma T}, \qquad (144)
$$

where $\gamma = \min(\sigma_{\min}^H, \frac{L}{2}, \chi_t)$ and $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$.

Given $\eta = \frac{\alpha\sigma_{\min}(G)}{17L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{2\sqrt{2031}\kappa_G}{\sigma^A_{\min}\alpha}$, since m is relatively small, it easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(1)$, which are independent on n and T. Thus, we obtain

$$\min_{1\leq t\leq T} \mathbb{E}\big[\mathrm{dist}(0, \partial L(x_t, y^t_{[m]}, z_t))^2\big] \leq O(\frac{1}{T}). \tag{145}$$

□

References

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

Huang, F., Chen, S., and Lu, Z. Stochastic alternating direction method of multipliers with variance reduction for nonconvex optimization. *arXiv preprint arXiv:1610.02758*, 2016.

Reddi, S., Sra, S., Poczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pp. 1145–1153, 2016.

Zheng, S. and Kwok, J. T. Stochastic variance-reduced admm. *arXiv preprint arXiv:1604.07070*, 2016.