# Nonconvex Stochastic Nested Optimization via Stochastic ADMM

Zhongruo Wang

November 8, 2019

**Abstract**

We consider the stochastic nested composition optimization problem where the objective is a composition of two expected-value functions. We proposed the stochastic ADMM to solve this complicated objective. In order to find an $\epsilon$ stationary point where the expected norm of the subgradient of corresponding augmented Lagrangian is smaller than $\epsilon$, the total sample complexity of our method is $\mathcal{O}(\epsilon^{-3})$ for the online case and $\mathcal{O}\Big((2N_1 + N_2) + (2N_1 + N_2)^{1/2}\epsilon^{-2}\Big)$ for the finite sum case. The computational complexity is consistent with proximal version proposed in [Zhang and Xiao, 2019], but our algorithm can solve more general problem when the proximal mapping of the penalty is not easy to compute.

## 1 Introduction

Consider we solve the following optimization problem:

$$\min_{x\in\mathbf{R}^d, y\in\mathbf{R}^l} F(x) + \sum_{j=1}^m r_j(y) = \mathbf{E}_{\xi_2} f_{2,\xi_2}\Big(\mathbf{E}_{\xi_1} f_{1,\xi_1}(x)\Big) + \sum_{j=1}^m r_j(y_j) \text{ s.t. } Ax + \sum_{j=1}^m B_j y_j = c \tag{1}$$

An interesting special case is when $\xi_1, \xi_2$ follows a uniform distribution:

$$\min_{x\in\mathbf{R}^d, y\in\mathbf{R}^l} F(x) + \sum_{j=1}^m r_j(y) = \frac{1}{N_2}\sum_{j=1}^{N_2} f_{2,j}\Big(\frac{1}{N_1}\sum_{j=1}^{N_1} f_{1,j}(x)\Big) + \sum_{j=1}^m r_j(y_j) \text{ s.t. } Ax + \sum_{j=1}^m B_j y_j = c \tag{2}$$

## 2 Applications

## 3 Assumptions

The following assumptions are made for the further analysis of the algorithms:

1. $A \in \mathbf{R}^{p\times d}, B_j \in \mathbf{R}^{p\times l} \ \forall j, c \in \mathbf{R}^p$.

2. $A$ and $B$ has full column rank or full row rank.

3. $F(x)$ is $L_F$-smooth

4. $f_1 : \mathbf{R}^d \to \mathbf{R}^l$ and $f_2 : \mathbf{R}^l \to \mathbf{R}$ are two smooth vector mapping, and each realization of the random mapping $f_{i,\xi_i}$ is $\ell_1$-Lipschitz continuous and its Jacobian $f'_{i,\xi_i}$ are $L_i$-Lipschitz continuous.

5. $\mathbf{E}_{\xi_1}\|f_{1,\xi_1}(x) - f_1(x)\|_2^2 \le \delta^2$ for all $x \in \mathbf{dom}F(x)$

6. $\mathbf{E}_{\xi_i}\|\nabla f_{i,\xi_i}(x) - \nabla f_i(x)\|_2^2 \le \sigma_i^2$ for all $i$

7. $r(x)$ is a convex regularizer such as $\|\cdot\|_1, \|\cdot\|_2$

# 4  Motivation and Previous Works

1. When penalty is not simple as $\ell_1$ penalty, for example graph guided lasso and fussed lasso, we can't use simple proximal algorithms. Thus, perform operator splitting and using ADMM will be suitable for those kind of problems.

2. ADMM for general convex and strongly convex cases has been studied in [Yu and Huang, 2017]. In their fomulation, they assume a very special case on the penalty that $Ax + By = 0$ which is not quite general for most ADMM problems. Using ADMM to solve the same composite nonconvex composite nested objective hasn't been well studied.

3. Different variance reduced stochastic proximal methods have been studied in both convex and nonconvex cases. Same kind of proximal algorithms have also been studied for formulations of multiple level composite functions: [Zhang and Xiao, 2019],[Lin et al., 2018].

# 5  Contribution

In this work we will present a stochastic variance reduced ADMM algorithm to solve 2-level and multiple level composite stochastic problems for both finite sum and online case. We denote the sampling number to be $N$ and the augmented Lagrangian with penalty $\rho$ to be $\mathcal{L}_\rho$. In order to achieve $\mathbf{E}\|\partial L_\rho(x^R, y_{[m]}^R, z^R)\|_2^2 \leq \epsilon$ for a given threshold $\epsilon > 0$, for simple mini batch estimation, we can show that iteration complexity is $\mathcal{O}(\epsilon^{-2})$ and the total complexity is $\mathcal{O}(\epsilon^{-4})$ which is too costy; when using stochastic intergraded estimator like SARAH/SPIDER, we can show that the total sampling complexity is $\mathcal{O}(\epsilon^{-3})$ for the online case and $\mathcal{O}((2N_1 + N_2) + \sqrt{(2N_1 + N_2)}\epsilon^{-2})$ for the finite sum case.

# 6  Notations

- $\sigma_{\min}^A$ and $\sigma_{\max}^A$ denotes the smallest and largest eigenvalue of the matrix $A^T A$, $\sigma_{\min}(H_j)$ and $\sigma_{\max}(H_j)$ denotes the smallest and largest eigenvalue of $H_j^T H_j$ for all $j \in [m]$.

# 7  Algorithms

Consider the gradient of $F(x)$, from which we will have:

$$F'(x) = (\mathbf{E}_{\xi_1}[f'_{1,\xi_1}(x)])\mathbf{E}_{\xi_2}[f'_{2,\xi_2}(\mathbf{E}_{\xi_1} f_{1,\xi_1}(x))] \tag{3}$$

Now we want to use the abbreviation to denote the approximations:

$$Y_1^k \approx f_1(x^k), \quad Z_1^k \approx f'_1(x^k), \quad Z_2^k \approx f'_2(Y_1^k)$$

Then the overall estimator for the gradient $F'(x)$ is $v^t = (Z_1^t)^T Z_2^T$

To solve the problem by using stochastic ADMM, we first give the augmented Lagrangian function of the problem:

$$\mathcal{L}_\rho(x, y_{[m]}, z) = F(x) + \sum_{j=1}^m g_j(y_j) - \langle z, Ax + \sum_{j=1}^m B_j y_j - c \rangle + \frac{\rho}{2}\|Ax + \sum_{j=1}^m B_j y_j - c\|_2^2 \tag{4}$$

Due to the stochastic gradient of the function $F$ to update $x$, we use the approximate Lagrangian over $x_k$ with the estimated gradient $v_k$:

$$\hat{\mathcal{L}}_\rho(x, y_{[m]}, z_k, v_k) = F(x_k) + v_k^T(x - x_k) + \frac{1}{2\eta}\|x - x_k\|_G^2$$
$$+ \sum_{j=1}^m g_j(y_j^{k+1}) - \langle z_k, Ax + \sum_{j=1}^m B_j y_j - c \rangle + \frac{\rho}{2}\|Ax + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2 \tag{5}$$

In order to avoid computing the inverse of $\frac{G}{\eta} + A^T A$, we can set $G = r I_d - \rho \eta A^T A \succeq I_d$ with $r \geq \rho \eta \sigma_{\max}^A + 1$ to linearize the quadratic term $\frac{\rho}{2} \| A x + \sum_{j=1}^{m} B_j y_j - c \|_2^2$. Also, in order to compute the proximal operater for each $y_i$, we can set $H_j = \tau_j I_d - \rho B_j^T B_j \succeq I_d$ with $\tau_j \geq \rho \sigma_{\max}^{B_j} + 1$ for all $j \in [m]$ to linearize the term: $\frac{\rho}{2} \| A x_k + \sum_{i=1}^{j-1} B_j y_j + B_j y_j + \sum_{i=j+1}^{m} B_i y_i \|_2^2$. The question remains now is how to find a suitable gradient estimator for the composite function.

Now we are ready to define the $\epsilon$-staionary point of the solution:

**Definition 7.1.** *For any $\epsilon > 0$, the point $(x^*, y^*, \lambda^*)$ is said to be an $\epsilon$ stationary point of the nonconvex problem* (1) *if it holds that:*

$$\begin{cases} \mathbf{E} \| A x^* + B y^* - c \|_2^2 \leq \epsilon^2 \\ \mathbf{E} \| \nabla f(x^*) - A^T \lambda^* \|_2^2 \leq \epsilon^2 \\ \mathbf{E} \| \boldsymbol{dist}(B^T \lambda^*, \partial r(y^*))^2 \|_2^2 \leq \epsilon^2 \end{cases} \tag{6}$$

*where $\boldsymbol{dist}(y_0, \partial r(y)) = \inf \{ \| y_0 - z \| : z \in \partial g(y) \}$, $\partial r(y)$ denotes the subgradient of $r(y)$. If $\epsilon = 0$, the point $(x^*, y^*, \lambda^*)$ is said to be a stationary point.*

The above inequalities (6) are equivalent to $\mathbf{E} \| \boldsymbol{dist}(0, \partial L(x^*, y^*, \lambda^*)) \|_2^2 \leq \epsilon^2$, where:

$$\partial L(x, y, \lambda) = \begin{bmatrix} \partial L(x, y, \lambda) / \partial x \\ \partial L(x, y, \lambda) / \partial y \\ \partial L(x, y, \lambda) / \partial \lambda \end{bmatrix} \tag{7}$$

and $L(x, y, \lambda) = f(x) + g(y) - \langle \lambda, A x + B y - c \rangle$ is the Lagrangian function of the objective function (1). In the following the sections, we first consider about the mini-batch estimation on the gradient, we show that ADMM still convergence by using this simple implementations after suitable choice of parameters. After that, we consider use SARAH/SPIDER to estimate the nested gradient. By comparing the sampling complexity, we can show that SARAH/SPIDER based algorithm is more efficient than traditional mini-batch based algorithm.

## 8 Simple Mini-Batch Estimator

When facing the stochastic composite objective, one simple strategy is to estimate the composite gradient by using mini batch. We denote $f_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} f_i(x)$ to be the mini-batch estimation of a funtion $f$. Since we are computing the composite gradient, we will use mini batch strategy on computing the gradient and sampling the function value at each level. Here comes with the following algorithm.

---
**Algorithm 1:** Stochastic Nested ADMM with simple Mini Batch estimator
---
1 Initialization: Initial Point $x^0$, Batch size: $(\{S, s, B_1, B_2, b_1, b_2\})$, $q, \eta, \rho > 0$
2 **for** $k = 0$ *to* $K - 1$ **do**
3     Randomly sample batch $\mathcal{S}^k$ of $\xi_1$ with $|\mathcal{S}^k| = s$;
4     $Y^k = f_{1, \mathcal{S}_1^k}(x^k)$
5     Randomly sample batch $\mathcal{B}_1^k$ of $\xi_1$ with $|\mathcal{B}_1^k| = b_1$, and $\mathcal{B}_2^k$ with $|\mathcal{B}_2^k| = b_2$
6     $Z_1^k = f'_{1, \mathcal{B}_1^t}(x^k)$
7     $Z_2^k = f'_{2, \mathcal{B}_2^t}(Y^k)$
8     Calculated the nested gradient estimation: $v^k = (Z_1^k)^T Z_2^k$
9     $y_j^{k+1} = \arg\min_{y_i} \{ \mathcal{L}_\rho(x^k, y_{[j-1]}^{k+1}), y_j, y_{[j+1:m]}^k \} + \frac{1}{2} \| y_j - y_j^k \|_{H_j}^2$ for all $j \in [m]$
10     $x^{k+1} = \arg\min_x \hat{\mathcal{L}}_\rho(x, y_{[m]}^{k+1}, z^k v^k)$
11     $z^{k+1} = z^k - \rho(A x^{k+1} - \sum_{j=1}^{m} B_j y_j^{k+1} - c)$
12 **end**
13 **Output**: $(x, y_{[m]}, z)$ choosen uniformly random from $\{x_k, y_{[m]}^k, z_k\}_{k=1}^K$

---

From the algorithm we can see that $v_k$ is a biased estimation for the full gradient, we can still perform the analysis on the variance. Firstly, in each iteration $k$ from [Zhang and Xiao, 2019], we know that:

$$\|v^k - F'(x^k)\|_2^2 \leq 3\|Z_1^t\|_2^2 \left(\|Z_2^k - f_2'(y^k)\|_2^2 + L_2^2\|Y_1^k - f_1(x^k)\|_2^2\right) + 3\ell_2^2\|Z_1^k - f_1'(x^k)\|_2^2$$

By using mini batch estimator, we will have the variance on each estimator to be:

$$\mathbf{E}\|Z_2^k - f_2'(y^k)\|_2^2 \leq \frac{\sigma_2^2}{b_2}, \quad \mathbf{E}\|Y_1^k - f_1(x^k)\|_2^2 \leq \frac{\delta^2}{s} \quad \mathbf{E}\|Z_1^k - f_1'(x^k)\|_2^2 \leq \frac{\sigma_1^2}{b_1}$$

Also, we can have:

$$\|Z_1^k\| = \|f_{1,\xi_1}'\| + \|\frac{1}{b}\sum_{\xi\in\mathcal{B}_1^r}(f_{1,\xi_1}'(x^r) - f_{1,\xi_1}'(x^{r-1}))\| \leq \|f_{1,\xi_1}'\| + \|f_{1,\xi_1}'(x^r)\| + \|f_{1,\xi_1}'(x^{r-1})\| = 3\ell_1 \quad (8)$$

Now, the variance bound on the estimated gradient by conditioning on the batches is:

$$\mathbf{E}\|v^k - F'(x^k)\|_2^2 \leq \underbrace{\frac{27\ell_1^2\sigma_2^2}{b_2} + \frac{27\ell_1^2 L_2^2\delta^2}{s} + \frac{3\ell_2^2\sigma_1^2}{b_1}}_{\mathcal{C}} \quad (9)$$

Now, we are ready to analysis the convergence of the our proposed ADMM based on SARAH/SPIDER estimator.

**Lemma 8.1** (Bound on the dual variable). *Given the sequence $\{x^k, y_{[m]}^k, z^k\}_{k=1}^K$ is generated by Algorithm (1), we will have the bound on updating the dual variable $z^k$ to be:*

$$\mathbf{E}\|z^{k+1} - z^k\|_2^2 \leq \frac{18\mathcal{C}}{\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \left(\frac{9L^2}{\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2}\right)\|x^k - x^{k-1}\|_2^2 \quad (10)$$

*Proof.* By using the optimal condition of step 18 in the algorithm 2, we will have:

$$v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^T z_k + \rho A^T(Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) = 0 \quad (11)$$

By the updating rule on the dual variable, we will have:

$$A^T z^{k+1} = v_k + \frac{G}{\eta}(x^{k+1} - x^k) \quad (12)$$

It follows that:

$$z_{k+1} = (A^T)^+(v_k + \frac{G}{\eta}(x^{k+1} - x^k)) \quad (13)$$

where $(A^T)^+$ is the pseudoinverse of $A^T$.

Taking expectation conditioned on $\mathcal{S}^k, \mathcal{B}_1^k, \mathcal{B}_2^k$:

$$\begin{aligned}
&\mathbf{E}\|z_{k+1} - z_k\|_2^2 \\
=&\mathbf{E}\|(A^T)^+(v_k + \frac{G}{\eta}(x^{k+1} - x^k) - v_{k-1} - \frac{G}{\eta}(x^k - x^{k-1}))\|_2^2 \\
\leq&\frac{1}{\sigma_{\min}^A}\|v_k + \frac{G}{\eta}(x^{k+1} - x^k) - v_{k-1} - \frac{G}{\eta}(x^k - x^{k-1})\|_2^2 \\
\leq&\frac{1}{\sigma_{\min}^A}\left[3\mathbf{E}\|v_k - v_{k-1}\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x^k - x^{k-1}\|_2^2\right]
\end{aligned} \quad (14)$$

Now we want to prove the upper bound of $\mathbf{E}\|v_k - v_{k-1}\|_2^2$:

$$
\begin{aligned}
&\mathbf{E}\|v_k - v_{k-1}\|_2^2 \\
=&\mathbf{E}\|v_k - \nabla f(x^k) + \nabla f(x^k) - \nabla f(x^{k-1}) + \nabla f(x^{k-1}) - v_{k-1}\|_2^2 \\
\leq&3\mathbf{E}\|v_k - \nabla f(x^k)\|_2^2 + 3\|\nabla f(x^k) - \nabla f(x^{k-1})\|_2^2 + 3\mathbf{E}\|v_{k-1} - \nabla f(x^{k-1})\|_2^2 \\
\leq&6\mathcal{C} + 3L^2\|x_{k-1} - x_k\|_2^2
\end{aligned}
\tag{15}
$$

Where the last inequality follows from (9).

In the end, we will have the bound on updating the dual variable to be:

$$
\mathbf{E}\|z^{k+1} - z^k\|_2^2 \leq \frac{18\mathcal{C}}{\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \left(\frac{9L^2}{\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2}\right)\|x^k - x^{k-1}\|_2^2
\tag{16}
$$

$\square$

**Lemma 8.2** (Point convergence).

*Proof.* By the optimal condition of step 9 in algorithm 1, we will have:

$$
\begin{aligned}
0 =&(y_j^k - y_j^{k+1})^T(\partial g_j(y_j^{k+1}) - B_j^T z_k + \rho B_j^T(Ax^k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) + H_j(y_j^{k+1} - y_j^k)) \\[2mm]
\leq&g_j(y_j^k) - g_j(y_j^{k+1}) - (z_k)^T(B_j y_j^k - B_j y_j^{k+1}) + \rho(B_j y_j^k - B_j y_j^{k+1})^T\left(Ax^k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\right) - \|y_j^{k+1} - y_j^k\|_{H_j}^2 \\[2mm]
\leq&g_j(y_j^k) - g_j(y_j^{k+1}) - z_k^T(Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c) + z_k^T(Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) \\
&+ \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c\|_2^2 + \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\|_2^2 \\
&- \frac{\rho}{2}\|B_j y_j^k - B_j y_j^{k+1}\|_2^2 - \|y_j^{k+1} - y_j^k\|_{H_j}^2 \\[2mm]
\leq&\mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \frac{\rho}{2}\|B_j y_j^k - B_j y_j^{k+1}\|_2^2 - \|y_j^{k+1} - y_j^k\|_{H_j}^2 \\[2mm]
\leq&\mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2
\end{aligned}
\tag{17}
$$

Now, we will have the decrease bound on update the $y_j$ component is:

$$
\mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) \leq -\sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2
\tag{18}
$$

Since we know that $F$ is $L_F$-smooth, we will have:

$$
F(x^{k+1}) \leq F(x^k) + \langle \nabla F(x^k), x^{k+1} - x^k \rangle + \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2
$$

Now, using the optimal condition on updating the $x$ component in the algorithm, we will have

$$
0 = (x^k - x^{k+1})^T\left(v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^T z_k + \rho A^T(Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c)\right)
\tag{19}
$$

5

Combine two equation above, we will have:

$$
\begin{aligned}
0 \leq &\, f(x^k) - f(x^{k+1}) + \nabla f(x_k)^T(x^{k+1} - x^k) + \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2 \\
&+ (x^k - x^{k+1})^T\left(v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^T z_k + \rho A^T(Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c)\right)
\end{aligned}
$$

$$
\begin{aligned}
\leq &\, f(x^k) - f(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x^k - x^{k+1})^T(v_k - \nabla f(x^k)) \\
&- (z_k)^T(Ax^k - Ax^{k+1}) + \rho(Ax^k - Ax^{k+1})^T(Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c)
\end{aligned}
$$

$$
\begin{aligned}
\leq &\, f(x^k) - f(x^{k+1}) + \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x_k - x_{k+1})^T(v_k - \nabla f(x^k)) \\
&- z_k^T(Ax^k + \sum_{j=1}^m B_j y_j^{k+1} - c) + z_k^T(Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) \\
&+ \frac{\rho}{2}\|Ax^k + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2 - \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2 - \frac{\rho}{2}\|Ax^k - Ax^{k+1}\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
= &\, \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z^k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^k) \\
&+ \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x^k - x^{k+1})^T(v_k - \nabla f(x^k)) - \frac{\rho}{2}\|Ax^k - Ax^{k+1}\|_2^2
\end{aligned}
$$

$$
= \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z^k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^k) - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L_F}{2}\right)\|x^{k+1} - x^k\|_2^2 + \langle x^k - x^{k+1}, v_k - \nabla f(x_k)\rangle
$$

$$
= \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F\right)\|x_{k+1} - x_k\|_2^2 + \frac{1}{2L_F}\|v_k - \nabla f(x^k)\|_2^2
$$

$$
= \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F\right)\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}}{2L_F}
$$

$$\tag{20}$$

Thus, rearranging the equations, taking expectation on the batches $\mathcal{B}_1^k, \mathcal{B}_2^k$ and $\mathcal{S}^k$, we will have:

$$
\mathbf{E}\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^k) - \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z^k) \leq -\left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F\right)\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}}{2L_F} \tag{21}
$$

Now, using the update of $z$ in the algorithm, we will have:

$$
\begin{aligned}
&\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1}) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^k) \\
&= \frac{1}{\rho}\|z^{k+1} - z^k\|_2^2 \\
&= \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \left(\frac{9L^2}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\right)\|x^k - x^{k-1}\|_2^2
\end{aligned} \tag{22}
$$

Now, combining (18),(21) and (22), we will have:

$$\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1}) - \mathcal{L}_\rho(x^k, y_{[m]}^k, z^k)$$

$$
\leq -\sum_{j=1}^m \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2 - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F\right)\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}}{2L_F}
$$
$$
+ \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \left(\frac{9L_F^2}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\right)\|x^k - x^{k-1}\|_2^2 \tag{23}
$$

$$
\leq \left(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A}\right)\|x^{k-1} - x^k\|_2^2 - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F - \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\right)\mathbf{E}\|x^{k+1} - x^k\|_2^2
$$
$$
+ \frac{\mathcal{C}}{2L_F} + \frac{18\mathcal{C}}{\rho\sigma_{\min}^A}
$$

Now we defined a useful potential function:

$$
R_k = \mathbf{E}\mathcal{L}_\rho(x^k, y_{[m]}^k, z_k) + \left(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A}\right)\mathbf{E}\|x^k - x^{k-1}\|_2^2 \tag{24}
$$

Now we can show that

$$
R_{k+1} = \mathbf{E}\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1}) + \left(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A}\right)\mathbf{E}\|x^{k+1} - x^k\|_2^2
$$
$$
\leq \mathcal{L}_\rho(x^k, y_{[m]}^k, z^k) + \left(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A}\right)\mathbf{E}\|x^k - x^{k-1}\|_2^2
$$
$$
- \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F - \left(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A}\right)\right)\mathbf{E}\|x^{k+1} - x^k\|_2^2 - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|_2^2
$$
$$
+ \left(\frac{\mathcal{C}}{2L_F} + \frac{18\mathcal{C}}{\rho\sigma_{\min}^A}\right) \tag{25}
$$
$$
\leq R_k - \Lambda\mathbf{E}\|x^{k+1} - x^k\|_2^2 - \sigma_{\min}^H \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|_2^2 + \left(\frac{\mathcal{C}}{2L_f} + \frac{18\mathcal{C}}{\rho\sigma_{\min}^A}\right)
$$
$$
\leq R_k - \gamma\left(\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|_2^2\right) + \left(\frac{\mathcal{C}}{2L_f} + \frac{18\mathcal{C}}{\rho\sigma_{\min}^A}\right)
$$

In which:
$$
\Lambda = \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F - \left(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A}\right)\right) > 0, \quad \gamma = \min(\sigma_{\min}^H, \Lambda)
$$

When we choose $\eta$ ...

Based on the structure of the potential function $R_k$, we want to show that $R_k$ is lower bounded.

$$\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1})$$

$$=F(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \langle z^{k+1}, Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c \rangle + \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq F(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \langle (A^T)^+(v_k + \frac{G}{\eta}(x^{k+1}-x^k)), Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c \rangle + \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq F(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \langle (A^T)^+(v_k - \nabla F(x^k) + \nabla F(x^k) + \frac{G}{\eta}(x^{k+1}-x^k)), Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c \rangle$$

$$+ \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq F(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \frac{2}{\sigma_{\min}^A \rho}\|v_k - \nabla F(x^k)\|_2^2 - \frac{2}{\sigma_{\min}^A \rho}\|\nabla F(x^k)\|_2^2 - \frac{2\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x^{k+1}-x^k\|_2^2$$

$$+ \frac{\rho}{8}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq F(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \frac{2}{\sigma_{\min}^A \rho}\|v_k - \nabla F(x^k)\|_2^2 - \frac{2}{\sigma_{\min}^A \rho}\|\nabla F(x^k)\|_2^2 - \left(\frac{9L_F^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\right)\|x^{k+1}-x^k\|_2^2$$

$$(26)$$

In all,

$$R_{k+1} \geq \mathbf{E}f(x_{k+1}) + \sum_{j=1}^{m} \mathbf{E}g_j(x_{k+1}) - \frac{2\mathcal{C}}{\sigma_{\min}^A \rho} - \frac{2L_F^2}{\sigma_{\min}^A \rho} \geq f^* + \sum_{j=1}^{m} g_j^* - \frac{2\mathcal{C}}{\sigma_{\min}^A \rho} - \frac{2L_F^2}{\sigma_{\min}^A \rho} \qquad (27)$$

It follows that the potential function $R_k$ is bounded below. Let's denote the lower bound of $R_k$ is $R^*$. Now we sum up the (58) and averaging all the iterates from 0 to $K-1$ we will have:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(\|x^{k+1}-x^k\|_2^2 + \sum_{j=1}^{m}\|y_j^k - y_j^{k+1}\|_2^2\right) \leq \frac{1}{K\gamma}\left(R_0 - R_*\right) + \left(\frac{\mathcal{C}}{2L_f\gamma} + \frac{18\mathcal{C}}{\rho\sigma_{\min}^A\gamma}\right) \qquad (28)$$

Denote $\gamma = \max(\frac{1}{2L_f\Lambda}, \frac{18}{\rho\sigma_{\min}^A\Lambda})$. Since $\mathcal{C} = \frac{27\ell_1^2\sigma_2^2}{b_2} + \frac{27\ell_1^2 L_2^2\delta^2}{s} + \frac{3\ell_2^2\sigma_1^2}{b_1}$, in order to achieve $\epsilon^2$ stationary point solution, we can choose:

$$b_2 = \frac{54\gamma\ell_1^2\sigma_2^2}{\epsilon^2}, s = \frac{54\gamma\ell_1^2 L_2^2\sigma^2}{\epsilon^2}, b_1 = \frac{6\gamma\ell_2^2\sigma^2}{\epsilon^2}$$

From the above analysis we can see that the order of choice of the batch size is $\mathcal{O}(\epsilon^{-2})$ □

**Lemma 8.3** (Stationary Point Convergence)**.**

*Proof.* Consider the sequence $\theta_k = \mathbf{E}\left[\|x^{k+1}-x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2 + \sum_{i=1}^{m}\|y_i^{k+1} - y_i^k\|_2^2\right]$, where $\mathbf{E}$ denotes the expectation conditioned on the batch $\mathcal{S}_1^k, \mathcal{S}_2^k, \mathcal{B}^k$

Consider in the update of $y_i$ component, we will ha

$$\mathbf{E}[\mathbf{dist}(0, \partial_{y_i}\mathcal{L}(x, y_{[m]}, z))]_{k+1} = \mathbf{E}[\mathbf{dist}(0, \partial g_j(y_j^{k+1}) - B_j^T z^{k+1})]_{k+1}$$

$$=\mathbf{E}\|B_j^T z^k - \rho B_j^T(Ax^k + \sum_{i=1}^{j} B_j y_j^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) - H_j(y_j^{k+1} - y_j^k) - B_j^T z^{k+1}\|_2^2$$

$$=\mathbf{E}\|\rho B_j^T A(x^{k+1} - x^{k+1}) + \rho B_j^T \sum_{i=j+1}^{m} B_i(y_i^{k+1} - y_i^k) - H_j(y_j^{k+1} - y_j^k)\|_2^2 \tag{29}$$

$$\leq m\rho^2 \sigma_{\max}^{B_j}\sigma_{\max}^{A}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + m\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^{m}\sigma_{\max}^{B_i}\mathbf{E}\|y_i^{k+1} - y_i^k\|_2^2 + m\sigma_{\max}^2(H_j)\mathbf{E}\|y_j^{k+1} - y_j^k\|_2^2$$

$$\leq m(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H_j))\theta_k = \nu_1\theta_k$$

In the updating of the $x$-component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \nabla_x\mathcal{L}(x, y_{[m]}, z))^2]_{k+1} = \mathbf{E}[\|A^T z^{k+1} - \nabla f(x^{k+1})\|_2^2]$$

$$\leq \mathbf{E}\|v_k - \nabla f(x^{k+1}) - \frac{G}{\eta}(x^k - x^{k+1}))\|_2^2$$

$$\leq \mathbf{E}\|v_k - \nabla f(x^k) + \nabla f(x^k) - \nabla f(x^{k+1}) - \frac{G}{\eta}(x^k - x^{k+1}))\|_2^2 \tag{30}$$

$$\leq 3\mathcal{C} + 3(L_F^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\mathbf{E}\|x^k - x^{k+1}\|_2^2$$

$$\leq 3(L_F^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\theta_k + 3\mathcal{C} = \nu_2\theta_k + 3\mathcal{C}$$

In the updating of the $z$ component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \nabla_z\mathcal{L}(x, y_{[m]}, z))^2]_{k+1}$$

$$=\mathbf{E}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$=\frac{1}{\rho^2}\|z^{k+1} - z^k\|_2^2$$

$$\leq \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \left(\frac{9L^2}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\right)\|x^k - x^{k-1}\|_2^2 \tag{31}$$

$$\leq \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + \left(\frac{9L^2}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\right)\mathbf{E}\|x^{k+1} - x^k\|_2^2$$

$$\leq \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + \left(\frac{9L^2}{\rho\sigma_{\min}^A} + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\right)\theta_k = \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + \nu_3\theta_k$$

Since we know that:

$$\sum_{k=1}^{K-1}\theta_k = \sum_{k=1}^{K-1}\mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2 + \sum_{j=1}^{m}\|y_j^k - y_j^{k+1}\|_2^2$$

$$\leq 2\sum_{k=1}^{K-1}\left(\mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \sum_{j=1}^{m}\mathbf{E}\|y_j^k - y_j^{k+1}\|_2^2\right)$$

9

Now, consider $T$ is chosen uniformly from $1, 2, ..., K-1, K$, we will have the following bound:

$$\mathbf{E}\| \mathbf{dist}(0, \partial L(x^T, y_{[m]}^T, z^T))\|_2^2$$

$$\leq \frac{3\tilde{\nu}}{K} \sum_{k=1}^{K} \theta_k + \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + 3\mathcal{C}$$

$$\leq \frac{6\tilde{\nu}}{K} \sum_{k=1}^{K-1} \left( \mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \sum_{j=1}^{m} \mathbf{E}\|y_j^k - y_j^{k+1}\|_2^2 \right) + \left( \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + 3\mathcal{C} \right)$$

$$\leq \frac{6\tilde{\nu}}{K\gamma}(R_0 - R_*) + \left( \frac{\mathcal{C}}{2L_f\gamma} + \frac{18\mathcal{C}}{\rho\sigma_{\min}^A\gamma} \right) + \left( \frac{18\mathcal{C}}{\rho\sigma_{\min}^A} + 3\mathcal{C} \right)$$

$$(32)$$

with $\tilde{\nu} = \max(\nu_1, \nu_2, \nu_3)$. Given $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}(0 < \alpha < 1)$ and $\Lambda \geq \frac{\sqrt{78}L_F\kappa_G}{4\alpha}$, we will have:

$$\mathbf{E}\| \mathbf{dist}(0, \partial L(x^T, y_{[m]}^T, z^T))\|_2^2 \leq \mathcal{O}(\frac{1}{K}) + \mathcal{O}(\mathcal{C}) \tag{33}$$

$\square$

**Theorem 8.1** (Total Sampling complexity)**.** *Consider we want to achieve an $\epsilon$-stationary point solution, the total iteration complexity is $\mathcal{O}(\epsilon^{-2})$. In order to obtain the optimal epoch length, we choose $b_1, b_2, s \sim \mathcal{O}(\epsilon^{-2})$ to be the batch size in each iteration. In all, after $\mathcal{O}(\epsilon^{-2})$ iterations, the total sample complexity is $\mathcal{O}(\epsilon^{-4})$*

**Remark 8.1.** *From the above theorem, by using mini-batch estimation,we can still get the same $\mathcal{O}(1/K)$ iteration complexity as nonconvex ADMM, but in order to achieve $\epsilon$-stationary solution, the batch size will be in the same order as the total iteeration number.*

# 9  SARAH/SPIDER Estimator

Based on the inefficiency and the superior performance of SARAH/SPIDER [Fang et al., 2018] based algorithm, we will introduce how to use this new technique on estimating the composite (nested) gradient, which will leads to a more efficient algorithm with lower sampling complexity when dealing with those kind of problems.

---

**Algorithm 2:** Stochastic Nested ADMM with SARAH/SPIDER estimator

---

**1** Initialization: $x_0$, Batch size: $(\{S, s, B_1, B_2, b_1, b_2\})$, $q, \eta, \rho > 0$

**2** **for** $k = 0$ *to* $K - 1$ **do**

**3**      **if** $mod(k, q) == 0$ **then**

**4**          Randomly sample batch $\mathcal{S}^k$ of $\xi_1$ with $|\mathcal{S}^k| = S$;

**5**          $Y^k = f_{1,\mathcal{S}_1}(x^k)$

**6**          Randomly sample batch $\mathcal{B}_1^k$ of $\xi_1$ with $|\mathcal{B}_1^k| = B_1$, and $\mathcal{B}_2^k$ with $|\mathcal{B}_2^k| = B_2$

**7**          $Z_1^k = f'_{1,\mathcal{B}_1^k}(x^k)$

**8**          $Z_2^k = f'_{2,\mathcal{B}_2^k}(Y^k)$

**9**      **else**

**10**          Randomly sample batch $\mathcal{S}^k$ of $\xi_1$ with $|\mathcal{S}^k| = s$;

**11**          $Y^k = Y^{k-1} + f_{1,\mathcal{S}_1}(x^k) - f_{1,\mathcal{S}_1}(x^{k-1})$

**12**          Randomly sample batch $\mathcal{B}_1^k$ of $\xi_1$ with $|\mathcal{B}_1^k| = b_1$, and $\mathcal{B}_2^k$ with $|\mathcal{B}_2^k| = b_2$

**13**          $Z_1^k = Z_1^{k-1} + f'_{1,\mathcal{B}_1^t}(x^k) - f'_{1,\mathcal{B}_1^t}(x^{k-1})$

**14**          $Z_2^k = Z_2^{k-1} + f'_{2,\mathcal{B}_2^t}(x^k) - f'_{2,\mathcal{B}_2^t}(x^{k-1})$

**15**      **end**

**16**      Calculated the nested gradient estimation: $v^k = (Z_1^k)^T Z_2^k$

**17**      $y_j^{k+1} = \arg\min_{y_i}\{\mathcal{L}_\rho(x^k, y_{[j-1]}^{k+1}), y_j, y_{[j+1:m]}^k\} + \frac{1}{2}\|y_j - y_j^k\|_{H_j}^2$

**18**      $x^{k+1} = \arg\min_x \hat{\mathcal{L}}_\rho(x, y_{[m]}^{k+1}, z^k v^k)$

**19**      $z^{k+1} = z^k - \rho(Ax^{k+1} - \sum_{j=1}^m B_j y_j^{k+1} - c)$

**20** **end**

**21** **Output**: $(x, y_{[m]}, z)$ choosen uniformly random from $\{x_k, y_{[m]}^k, z_k\}_{k=1}^K$

---

Now we want to analysis the convergence of the algorithm. First, we want to show that under the choice of the suitable parameters, we can make sure that the gradient estimator is unbiased [Zhang and Xiao, 2019]. Throughout the paper, we will consider $n_k = \lceil k/q \rceil$ such that $(n_k - 1)q + 1 \le k \le n_k q - 1$.

**Lemma 9.1** ([Fang et al., 2018]). *Under assumption 2, the SPIDER generates stochastic gradient $v_k$ satisfies for all $(n_k - 1)q + 1 \le k \le n_k q - 1$*

$$\mathbf{E}\|v^k - \nabla f(x^k)\|_2^2 \le \sum_{i=(n_t-1)q}^{k-1} \frac{L^2}{|\mathcal{S}_2|}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \mathbf{E}\|v^{(n_k-1)q} - \nabla f(x^{(n_k-1)q})\|_2^2$$

Based on the SARAH/SPIDER estimator above, we can have the following upper bound on the variance of the estimation.

Firstly, from [Zhang and Xiao, 2019], we know that:

$$\|v^k - F'(x^k)\|_2^2 \le 3\|z_1^t\|_2^2\Big(\|z_2^k - f'_2(y^k)\|_2^2 + L_2^2\|y_1^k - f_1(x^k)\|_2^2\Big) + 3\ell_2^2\|z_1^k - f'_1(x^k)\|_2^2$$

Now, let's bound every term in the above inequality:

$$\|z_1^k\| = \|f'_{1,\xi_1}\| + \|\frac{1}{b}\sum_{\xi\in\mathcal{B}_1^r}(f'_{1,\xi_1}(x^r) - f'_{1,\xi_1}(x^{r-1}))\| \le \|f'_{1,\xi_1}\| + \|f'_{1,\xi_1}(x^r) + f'_{1,\xi_1}(x^{r-1})\| = 3\ell_1 \quad (34)$$

For all $(n_k - 1)q \leq r \leq k$, by using the SPIDER estimator, we will have:

$$\mathbf{E}\|Z_2^k - f_2'(Y^k)\|_2^2 \leq \mathbf{E}\|Z_2^0 - f_2'(Y_1^0)\|_2^2 + \frac{L_2^2}{b_2} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}[\|x^{r+1} - x^r\|_2^2] \leq \frac{\sigma_2^2}{B_2} + \ell_1^2 \frac{L_2^2}{b_2} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}\|x^{r+1} - x^r\|_2^2$$

$$(35)$$

$$\mathbf{E}\|Y_1^t - f_1(x^k)\|_2^2 \leq \mathbf{E}\|Y_1^0 - f_1(x^0)\|_2^2 + \frac{\ell_1^2}{s} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}[\|x^{r+1} - x^r\|_2^2] \leq \frac{\delta_1^2}{S} + \frac{\ell_1^2}{s} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}[\|x^{r+1} - x^r\|_2^2] \quad (36)$$

$$\mathbf{E}\|Z_1^t - f_1'(x^k)\|_2^2 \leq \mathbf{E}\|Z_1^0 - f_1'(x^0)\|_2^2 + \frac{L_1^2}{b_1} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}[\|x^{r+1} - x^r\|_2^2] \leq \frac{\sigma_1^2}{B_1} + \frac{L_1^2}{b_1} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}[\|x^{r+1} - x^r\|_2^2] \quad (37)$$

So we have the following conclusions after combining above inequalties:

- For the online case:

$$\|v^k - F'(x^k)\|_2^2 \leq \underbrace{\frac{27\ell_1^2\sigma_2^2}{B_2} + \frac{27\ell_1^2\delta_1^2}{S} + \frac{3\ell_2^2\sigma_1^2}{B_1}}_{\mathcal{C}_1} + \underbrace{\left(\frac{27\ell_1^4 L_2^2}{b_2} + \frac{27\ell_1^4}{s} + \frac{3\ell_2^2 L_1^2}{b_1}\right)}_{\mathcal{C}_2} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}[\|x^{r+1} - x^r\|_2^2]$$

$$(38)$$

- for the finite sum case, all the estimated variance will be vanished, we will have all the :

$$\|v^k - F'(x^k)\|_2^2 \leq \underbrace{\left(\frac{27\ell_1^4 L_2^2}{b_2} + \frac{27\ell_1^4}{s} + \frac{3\ell_2^2 L_1^2}{b_1}\right)}_{\mathcal{C}_2} \sum_{r=(n_k-1)q}^{k-1} \mathbf{E}[\|x^{r+1} - x^r\|_2^2] \quad (39)$$

### 9.0.1 Finite Sum Case

**Lemma 9.2** (Bound on the dual variable). *Given the sequence $\{x^k, y_{[m]}^k, z^k\}_{k=1}^K$ is generated by Algorithm (2), we will have the bound on updating the dual variable $z^k$ to be:*

$$\mathbf{E}\|z^{k+1} - z^k\|_2^2 \leq \frac{6\mathcal{C}_2}{\sigma_{\min}^A} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2}\|x^{k+1} - x^k\|_2^2 + \left(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2} + \frac{9L^2}{\sigma_{\min}^A}\right)\|x^{k-1} - x^k\|_2^2$$

$$(40)$$

*Proof.* By using the optimal condition of step 18 in the algorithm 2, we will have:

$$v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^T z_k + \rho A^T (Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) = 0 \quad (41)$$

By the updating rule on the dual variable, we will have:

$$A^T z^{k+1} = v_k + \frac{G}{\eta}(x^{k+1} - x^k) \quad (42)$$

It follows that:

$$z_{k+1} = (A^T)^+(v_k + \frac{G}{\eta}(x^{k+1} - x^k)) \quad (43)$$

where $(A^T)^+$ is the pseudoinverse of $A^T$.

Now we will have:

$$
\begin{aligned}
&\mathbf{E}\|z_{k+1} - z_k\|_2^2 \\
=&\mathbf{E}\|(A^T)^+(v_k + \frac{G}{\eta}(x^{k+1} - x^k) - v_{k_1} - \frac{G}{\eta}(x^k - x^{k-1}))\|_2^2 \\
\leq&\frac{1}{\sigma_{\min}^A}\|v_k + \frac{G}{\eta}(x^{k+1} - x^k) - v_{k_1} - \frac{G}{\eta}(x^k - x^{k-1})\|_2^2 \\
\leq&\frac{1}{\sigma_{\min}^A}\left[3\mathbf{E}\|v_k - v_{k-1}\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\mathbf{E}\|x^k - x^{k-1}\|_2^2\right]
\end{aligned}
\tag{44}
$$

Now we want to prove the upper bound of $\mathbf{E}\|v_k - v_{k-1}\|_2^2$:

$$
\begin{aligned}
&\mathbf{E}\|v_k - v_{k-1}\|_2^2 \\
=&\mathbf{E}\|v_k - \nabla f(x^k) + \nabla f(x^k) - \nabla f(x^{k-1}) + \nabla f(x^{k-1}) - v_{k-1}\|_2^2 \\
\leq&3\mathbf{E}\|v_k - \nabla f(x^k)\|_2^2 + 3\mathbf{E}\|\nabla f(x^k) - \nabla f(x^{k-1})\|_2^2 + 3\mathbf{E}\|v_{k-1} - \nabla f(x^{k-1})\|_2^2 \\
\leq&6\mathcal{C}_2\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + 3L^2\mathbf{E}\|x_{k-1} - x_k\|_2^2
\end{aligned}
\tag{45}
$$

In the end, we will have the bound on updating the dual variable to be:

$$
\mathbf{E}\|z^{k+1} - z^k\|_2^2 \leq \frac{6\mathcal{C}_2}{\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2}\|x^{k+1} - x^k\|_2^2 + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2} + \frac{9L^2}{\sigma_{\min}^A})\|x^{k-1} - x^k\|_2^2
\tag{46}
$$

$\square$

**Lemma 9.3** (Point convergence). *Consider the sequence $\{x^k, y_{[m]}^k, z^k\}_{k=1}^K$ is generated from algorithm (2). Define a potential function $R_k$ as follows:*

$$
R_k = \mathcal{L}_\rho(x^k, y_{[m]}^k, z_k) + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A})\|x^{k-1} - x^k\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2
\tag{47}
$$

*Denote $R_*$ is the lower bound of $R_k$. We will have:*

$$
\frac{1}{K}\sum_{k=0}^K\left(\|x^{k+1} - x^k\|_2^2 + \sum_{j=1}^m\|y_j^k - y_j^{k+1}\|_2^2\right) \leq \frac{1}{K\gamma}\left(\mathbf{E}[R_0] - R_*\right)
\tag{48}
$$

*by setting*

$$
\mathcal{C}_2 = \frac{L_F}{q}, b_2 = \frac{27\ell_1^4 L_2^2 q}{L_F^2}, s = \frac{27\ell_1^4 q}{L_F^2}, b_1 = \frac{3\ell_2^2 L_1^2 q}{L_F^2}, \rho \geq \frac{\sqrt{78}L_F\kappa_G}{4\alpha}, \eta = \frac{2\alpha\sigma_{\min}(G)}{3L}(0 \leq \alpha \leq 1)
$$

*Proof.* By the optimal condition of step 9 in algorithm 2, we will have:

$$0 = (y_j^k - y_j^{k+1})^T (\partial g_j(y_j^{k+1}) - B_j^T z_k + \rho B_j^T (Ax^k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) + H_j(y_j^{k+1} - y_j^k))$$

$$\leq g_j(y_j^k) - g_j(y_j^{k+1}) - (z_k)^T (B_j y_j^k - B_j y_j^{k+1}) + \rho(B_j y_j^k - B_j y_j^{k+1})^T \left( Ax^k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c \right) - \|y_j^{k+1} - y_j^k\|_{H_j}^2$$

$$\leq g_j(y_j^k) - g_j(y_j^{k+1}) - z_k^T (Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c) + z_k^T (Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c)$$

$$+ \frac{\rho}{2} \|Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c\|_2^2 + \frac{\rho}{2} \|Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\|_2^2$$

$$- \frac{\rho}{2} \|B_j y_j^k - B_j y_j^{k+1}\|_2^2 - \|y_j^{k+1} - y_j^k\|_{H_j}^2$$

$$\leq \mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \frac{\rho}{2} \|B_j y_j^k - B_j y_j^{k+1}\|_2^2 - \|y_j^{k+1} - y_j^k\|_{H_j}^2$$

$$\leq \mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \sigma_{\min}(H_j) \|y_j^k - y_j^{k+1}\|_2^2$$

$$(49)$$

Now, we will have the decrease bound on update the $y_j$ component is:

$$\mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) \leq -\sigma_{\min}(H_j) \|y_j^k - y_j^{k+1}\|_2^2 \qquad (50)$$

Since we know that $f$ is $L_F$-smooth, we will have:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L_F}{2} \|x^{k+1} - x^k\|_2^2$$

Now, using the optimal condition on updating the $x$ component in the algorithm, we will have

$$0 = (x^k - x^{k+1})^T \left( v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^T z_k + \rho A^T (Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c) \right) \qquad (51)$$

Combine two equation above, we will have:

$$0 \leq f(x^k) - f(x^{k+1}) + \nabla f(x_k)^T(x^{k+1} - x^k) + \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2$$

$$+ (x^k - x^{k+1})^T\left(v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^Tz_k + \rho A^T(Ax^{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c)\right)$$

$$\leq f(x^k) - f(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x_k - x_{k+1})^T(v_k - \nabla f(x^k))$$

$$- (z_k)^T(Ax_k - Ax_{k+1}) + \rho(Ax_k - Ax_{k+1})^T(Ax_{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c)$$

$$\leq f(x^k) - f(x^{k+1}) + \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x_k - x_{k+1})^T(v_k - \nabla f(x^k))$$

$$- z_k^T(Ax^k + \sum_{j=1}^m B_jy_j^{k+1} - c) + z_k^T(Ax^{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c)$$

$$+ \frac{\rho}{2}\|Ax^k + \sum_{j=1}^m B_jy_j^{k+1} - c\|_2^2 - \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^m B_jy_j^{k+1} - c\|_2^2 - \frac{\rho}{2}\|Ax^k - Ax^{k+1}\|_2^2$$

$$= \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k)$$

$$+ \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x_k - x_{k+1})^T(v_k - \nabla f(x^k)) - \frac{\rho}{2}\|Ax^k - Ax^{k+1}\|_2^2$$

$$= \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L_F}{2}\right)\|x_{k+1} - x_k\|_2^2 + \langle x_k - x_{k+1}, v_k - \nabla f(x_k)\rangle$$

$$= \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F\right)\|x_{k+1} - x_k\|_2^2 + \frac{1}{2L_F}\|v_k - \nabla f(x^k)\|_2^2$$

$$= \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F\right)\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}_2}{2L_F}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

$$\tag{52}$$

$$\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^k) - \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z^k) \leq -\left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F\right)\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}_2}{2L_F}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x_{i+1} - x_i\|_2^2$$

$$\tag{53}$$

Now, using the update of $z$ in the algorithm, we will have:

$$\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1}) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^k)$$

$$= \frac{1}{\rho}\|z^{k+1} - z^k\|_2^2$$

$$= \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\|x^{k+1} - x^k\|_2^2 + \left(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A}\right)\|x^{k-1} - x^k\|_2^2$$

$$\tag{54}$$

15

Now, combining (50),(53) and (54), we will have:

$$\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1}) - \mathcal{L}_\rho(x^k, y_{[m]}^k, z^{k+1})$$

$$\leq -\sum_{j=1}^m \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2 - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L\Big)\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}_2}{2L}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

$$+ \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\|x^{k+1} - x^k\|_2^2 + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A})\|x^{k-1} - x^k\|_2^2$$

$$\leq -\sum_{j=1}^m \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2 + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A})\|x^{k-1} - x^k\|_2^2$$

$$- \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2}\Big)\|x^{k+1} - x^k\|_2^2$$

$$+ \Big(\frac{\mathcal{C}_2}{2L} + \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

(55)

Now we defined a useful potential function:

$$R_k = \mathcal{L}_\rho(x^k, y_{[m]}^k, z_k) + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A})\|x^{k-1} - x^k\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 \quad (56)$$

Now we can show that

$$R_{k+1} = \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_{k+1}) + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A})\|x^{k+1} - x^k\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

$$\leq \mathcal{L}_\rho(x^k, y_{[m]}^k, z^k) + (\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A})\|x^k - x^{k-1}\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

$$- \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\|x^{k+1} - x^k\|_2^2 - \sigma_{\min}^H\sum_{j=1}^m\|y_j^k - y_j^{k+1}\|_2^2$$

$$+ \Big(\frac{\mathcal{C}_2}{2L_F} + \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

$$\leq R_k - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\|x^{k+1} - x^k\|_2^2 - \sigma_{\min}^H\sum_{j=1}^m\|y_j^k - y_j^{k+1}\|_2^2$$

$$+ \Big(\frac{\mathcal{C}_2}{2L_F} + \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

(57)

Let $(n_k - 1)q \leq l \leq n_kq - 1$, telescoping (57) over $k$ from $(n_k - 1)q$ to $k$ and take the expectation, we will

have:

$$\mathbf{E}[R_{k+1}]$$

$$\leq \mathbf{E}[R_{(n_k-1)q}] - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\right)\sum_{l=(n_k-1)q}^{k}\|x^{l+1}-x^l\|_2^2$$

$$- \sigma_{\min}^H \sum_{l=(n_k-1)q}^{k}\sum_{j=1}^{m}\|y_j^l - y_j^{l+1}\|_2^2 + \left(\frac{\mathcal{C}_2}{2L} + \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\right)\sum_{l=(n_k-1)q}^{k}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1}-x^i\|_2^2$$

$$\leq \mathbf{E}[R_{(n_k-1)q}] - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A} - \frac{\mathcal{C}_2 q}{2L_F} + \frac{6\mathcal{C}_2 q}{\rho\sigma_{\min}^A}\right)\sum_{l=(n_k-1)q}^{k}\|x^{l+1}-x^l\|_2^2$$

$$- \sigma_{\min}^H \sum_{l=(n_k-1)q}^{k}\sum_{j=1}^{m}\|y_j^l - y_j^{l+1}\|_2^2$$

$$(58)$$

Consider we have $\mathcal{C}_2 q = L_F^2$, let's denote

$$\Lambda = \Big(\underbrace{\frac{\sigma_{\min}(G)}{\eta} - \frac{3L_F}{2}}_{\Lambda_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2L_F^2}{\rho\sigma_{\min}^A q}}_{\Lambda_2}\Big)$$

Now, choosing $0 \leq \eta \leq \frac{2\sigma_{\min}(G)}{3L}$, we will have $\Lambda_1 > 0$.

Further, let $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L_F}(0 < \alpha < 1)$, since $b > 1 > \alpha^2, \kappa_G = \frac{\sigma_{\max}^A}{\sigma_{\min}^A} > 1$, we will have:

$$\begin{aligned}
\Lambda_2 &= \frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2L_F^2}{\rho\sigma_{\min}^A}\\
&= \frac{\rho\sigma_{\min}^A}{2} - \frac{27L_F^2\kappa_G^2}{2\sigma_{\min}^A\rho\alpha^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2L_F^2}{\rho\sigma_{\min}^A q}\\
&\geq \frac{\rho\sigma_{\min}^A}{2} - \frac{27L_F^2\kappa_G^2}{2\sigma_{\min}^A\rho\alpha^2} - \frac{9L_F^2\kappa_G^2}{\rho\sigma_{\min}^A\alpha^2} - \frac{2L_F^2\kappa_G^2}{\rho\sigma_{\min}^A\alpha^2}\\
&= \frac{\rho\sigma_{\min}^A}{2} - \frac{49L_F^2\kappa_G^2}{2\sigma_{\min}^A\rho\alpha^2}\\
&= \frac{\rho\sigma_{\min}^A}{4} + \frac{\rho\sigma_{\min}^A}{4} - \frac{49L_F^2\kappa_G^2}{2\sigma_{\min}^A\rho\alpha^2}
\end{aligned}$$

$$(59)$$

From the above result, we just need to choose the penalty $\rho \geq \frac{\sqrt{98}L_F\kappa_G}{\sigma_{\min}^A\alpha}$. Upon the result we have, we can argue that:

$$\Lambda \geq \frac{\sqrt{98}L_F\kappa_G}{4\alpha}$$

Also, by choosing $\mathcal{C}_2 = L_F^2/q$, we will have:

$$\left(\frac{27\ell_1^4 L_2^2}{b_2} + \frac{27\ell_1^4}{s} + \frac{3\ell_2^2 L_1^2}{b_1}\right) = \frac{L_F^2}{q}$$

We can have that:

$$b_2 = \frac{27\ell_1^4 L_2^2 q}{L_F^2}, s = \frac{27\ell_1^4 q}{L_F^2}, b_1 = \frac{3\ell_2^2 L_1^2 q}{L_F^2}$$

Since we know that $L_F^2 \sim \mathcal{O}(\ell_1^4 + \ell_2^2)$, we can argue that $b_1, s, b_2 \sim \mathcal{O}(q)$.

17

Based on the structure of the potential function $R_k$, we want to show that $R_k$ is lower bounded.

$$
\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1})
$$
$$
= f(x^{k+1}) + \sum_{i=1}^m g_j(y_j^{k+1}) - \langle z^{k+1}, Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c \rangle + \frac{\rho}{2} \|Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2
$$
$$
\geq f(x^{k+1}) + \sum_{i=1}^m g_j(y_j^{k+1}) - \langle (A^T)^+(v_k + \frac{G}{\eta}(x^{k+1} - x^k)), Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c \rangle + \frac{\rho}{2} \|Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2
$$

$$
\geq f(x^{k+1}) + \sum_{i=1}^m g_j(y_j^{k+1}) - \langle (A^T)^+(v_k - \nabla f(x^k) + \nabla f(x^k) + \frac{G}{\eta}(x^{k+1} - x^k)), Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c \rangle
$$
$$
+ \frac{\rho}{2} \|Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2
$$

$$
\geq f(x^{k+1}) + \sum_{i=1}^m g_j(y_j^{k+1}) - \frac{2}{\sigma_{\min}^A \rho} \|v_k - \nabla f(x^k)\|_2^2 - \frac{2}{\sigma_{\min}^A} \|\nabla f(x^k)\|_2^2 - \frac{2\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} \|x^{k+1} - x^k\|_2^2
$$
$$
+ \frac{\rho}{8} \|Ax^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2
$$

$$
\geq f(x^{k+1}) + \sum_{i=1}^m g_j(y_j^{k+1}) - \frac{2L_F^2}{\sigma_{\min}^A q\rho} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1} - x^i\|_2^2 - \frac{2L_F^2}{\sigma_{\min}^A \rho} - \frac{2\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} \|x^{k+1} - x^k\|_2^2
$$
$$
\geq f(x^{k+1}) + \sum_{i=1}^m g_j(y_j^{k+1}) - \frac{2L_F^2}{\sigma_{\min}^A q\rho} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1} - x^i\|_2^2 - \frac{2L_F^2}{\sigma_{\min}^A \rho} - \left( \frac{9L_F^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} \right) \|x^{k+1} - x^k\|_2^2
$$
$$
\tag{60}
$$

In all,
$$
R_{k+1} \geq f(x_{k+1}) + \sum_{j=1}^m g_j(x_{k+1}) - \frac{2L_F^2}{\sigma_{\min}^A \rho} \geq f^* + \sum_{j=1}^m g_j^* - \frac{2L_F^2}{\sigma_{\min}^A \rho} \tag{61}
$$

It follows that the potential function $R_k$ is bounded below. Let's denote the lower bound of $R_k$ is $R^*$. Now we sum up the (58) over all the iterates from 0 to $K$, we will have:

$$
\mathbf{E}[R_k] - \mathbf{E}[R_0] \leq - \sum_{i=0}^{K-1} (\Lambda \|x^{i+1} - x^i\|_2^2 + \sigma_{\min}^H \sum_{j=1}^m \|y_j^i - y_j^{i+1}\|_2^2) \tag{62}
$$

Finally, we will have the iteration bound to be:

$$
\frac{1}{K} \sum_{k=0}^K \left( \|x^{k+1} - x^k\|_2^2 + \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|_2^2 \right) \leq \frac{1}{K\gamma} \left( R_0 - R_* \right) \tag{63}
$$

In which $\gamma = \min(\Lambda, \sigma_{\min}^H)$. $\qquad\square$

**Lemma 9.4** (Stationary point convergence). *Suppose the sequence $\{x^k, y_{[m]}^k, z^k\}$ is generated from Algorithm (2), there exists a constant $\tilde{\nu}$ such that, with $T$ sampling uniformly from $1, ..., K$, we will have:*

$$
\mathbf{E}\| \mathbf{dist}(0, \partial L(x^T, y_{[m]}^T, z^T)) \|_2^2 \leq \frac{9\tilde{\nu}}{K\gamma}(R_0 - R_*) \tag{64}
$$

18

*Proof.* Consider the sequence $\theta_k = \mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2 + \frac{1}{q}\sum_{i=(n_k-1)q}^{k}\|x^{i+1} - x^i\|_2^2 + \sum_{j=1}^{m}\|y_j^k - y_j^{k+1}\|_2^2]$.

Consider in the update of $y_i$ component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \partial_{y_i}\mathcal{L}(x, y_{[m]}, z))]_{k+1} = \mathbf{E}[\mathbf{dist}(0, \partial g_j(y_j^{k+1}) - B_j^T z^{k+1})]_{k+1}$$

$$=\mathbf{E}\|B_j^T z^k - \rho B_j^T(Ax^k + \sum_{i=1}^{j} B_j y_j^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) - H_j(y_j^{k+1} - y_j^k) - B_j^T z^{k+1}\|_2^2$$

$$=\mathbf{E}\|\rho B_j^T A(x^{k+1} - x^{k+1}) + \rho B_j^T \sum_{i=j+1}^{m} B_i(y_i^{k+1} - y_i^k) - H_j(y_j^{k+1} - y_j^k)\|_2^2 \tag{65}$$

$$\leq m\rho^2 \sigma_{\max}^{B_j}\sigma_{\max}^A \mathbf{E}\|x^{k+1} - x^k\|_2^2 + m\rho^2 \sigma_{\max}^{B_j}\sum_{i=j+1}^{m}\sigma_{\max}^{B_i}\mathbf{E}\|y_i^{k+1} - y_i^k\|_2^2 + m\sigma_{\max}^2(H_j)\mathbf{E}\|y_j^{k+1} - y_j^k\|_2^2$$

$$\leq m(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H_j))\theta_k = \nu_1\theta_k$$

In the updating of the $x$-component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \nabla_x\mathcal{L}(x, y_{[m]}, z))^2]_{k+1} = \mathbf{E}[\|A^T z^{k+1} - \nabla f(x^{k+1})\|_2^2]$$

$$\leq \mathbf{E}\|v_k - \nabla f(x^{k+1}) - \frac{G}{\eta}(x^k - x^{k+1}))\|_2^2$$

$$\leq \mathbf{E}\|v_k - \nabla f(x^k) + \nabla f(x^k) - \nabla f(x^{k+1}) - \frac{G}{\eta}(x^k - x^{k+1}))\|_2^2 \tag{66}$$

$$\leq \sum_{i=(n_k-1)q}^{k-1} \frac{L_F^2}{2}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + 3(L_F^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\|x^k - x^{k+1}\|_2^2$$

$$\leq 3(L_F^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\theta_k = \nu_2\theta_k$$

In the updating of the $z$ component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \nabla_z\mathcal{L}(x, y_{[m]}, z))^2]_{k+1}$$

$$=\mathbf{E}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$=\frac{1}{\rho^2}\|z^{k+1} - z^k\|_2^2 \tag{67}$$

$$\leq \frac{6C_2}{\rho^2\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\rho^2\sigma_{\min}^A\eta^2}\|x^{k+1} - x^k\|_2^2 + (\frac{3\sigma_{\max}^2(G)}{\rho^2\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho^2\sigma_{\min}^A})\|x^{k-1} - x^k\|_2^2$$

$$\leq (\frac{9L^2}{\rho^2\sigma_{\min}^A} + \frac{3\sigma_{\max}^2 G}{\rho^2\sigma_A^2\eta^2})\theta_k = \nu_3\theta_k$$

Since we know that:

$$\sum_{k=1}^{K-1}\theta_k = \sum_{k=1}^{K-1}\mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2 + \frac{1}{q}\sum_{i=(n_k-1)q}^{k}\|x^{i+1} - x^i\|_2^2 + \sum_{j=1}^{m}\|y_j^k - y_j^{k+1}\|_2^2]$$

$$\leq 3\left(\sum_{k=1}^{K-1}\mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \sum_{k=1}^{K-1}\sum_{j=1}^{m}\mathbf{E}\|y_j^k - y_j^{k+1}\|_2^2\right)$$

19

Now, consider $T$ is chosen uniformly from $1, 2, ..., K-1, K$, we will have the following bound:

$$\mathbf{E}\|\,\mathbf{dist}(0, \partial L(x^R, y^R_{[m]}, z^R))\|_2^2$$

$$\leq \frac{\tilde{\nu}}{K} \sum_{k=1}^{K} \theta_k \leq \frac{9\tilde{\nu}}{K} \Big( \sum_{k=1}^{K-1} \mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \sum_{k=1}^{K-1} \sum_{j=1}^{m} \mathbf{E}\|y_j^k - y_j^{k+1}\|_2^2 \Big) \tag{68}$$

$$\leq \frac{9\tilde{\nu}}{K\gamma}(R_0 - R_*)$$

with $\tilde{\nu} = \max(\nu_1, \nu_2, \nu_3)$. Given $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}(0 < \alpha < 1)$ and $\Lambda \geq \frac{\sqrt{78}L_F\kappa_G}{4\alpha}$, we will have:

$$\mathbf{E}\|\,\mathbf{dist}(0, \partial L(x^T, y^T_{[m]}, z^T))\|_2^2 \leq \mathcal{O}(\frac{1}{K}) \tag{69}$$

$\square$

**Theorem 9.1** (Total Sampling complexity). *Consider we want to achieve an $\epsilon$-stationary point solution, the total iteration complexity is $\mathcal{O}(\epsilon^{-2})$. In order to obtain the optimal epoch length, we choose $q = (2N_1 + N_2)^{\frac{1}{2}}$ to be the size of the inner loop and $b_1, b_2, s \sim \mathcal{O}((2N_1 + N_2)^{\frac{1}{2}})$ accordingly. After $\mathcal{O}(\epsilon^{-2})$ iterations, the total sample complexity is $\mathcal{O}\Big((2N_1 + N_2) + (2N_1 + N_2)^{\frac{1}{2}}\epsilon^{-2}\Big)$*

### 9.0.2 Online Case

**Lemma 9.5** (Bound on the dual variable). *Given the sequence $\{x^k, y^k_{[m]}, z^k\}_{k=1}^{K}$ is generated by Algorithm (2), we will have the bound on updating the dual variable $z^k$ to be:*

$$\mathbf{E}\|z^{k+1} - z^k\|_2^2 \leq \frac{6\mathcal{C}_1}{\sigma_{\min}^A} + \frac{6\mathcal{C}_2}{\sigma_{\min}^A} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2}\|x^{k+1} - x^k\|_2^2 + \Big(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2} + \frac{9L^2}{\sigma_{\min}^A}\Big)\|x^{k-1} - x^k\|_2^2 \tag{70}$$

*Proof.* By using the proof strategy in equation eq. (46), we will have:

$$\mathbf{E}\|z^{k+1} - z^k\|_2^2 \leq \frac{6\mathcal{C}_1}{\sigma_{\min}^A} + \frac{6\mathcal{C}_2}{\sigma_{\min}^A} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2}\|x^{k+1} - x^k\|_2^2 + \Big(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2} + \frac{9L^2}{\sigma_{\min}^A}\Big)\|x^{k-1} - x^k\|_2^2 \tag{71}$$

$\square$

**Lemma 9.6** (Point Convergence). *Consider the sequence $\{x^k, y^k_{[m]}, z^k\}_{k=1}^{K}$ is generated from algorithm (2). Define a potential function $R_k$ as follows:*

$$R_k = \mathcal{L}_\rho(x^k, y^k_{[m]}, z_k) + \Big(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A \eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A}\Big)\|x^{k-1} - x^k\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1} - x^i\|_2^2 \tag{72}$$

*Denote $R_*$ is the lower bound of $R_k$. We will have:*

$$\frac{1}{K} \sum_{k=0}^{K} \Big(\|x^{k+1} - x^k\|_2^2 + \sum_{j=1}^{m} \|y_j^k - y_j^{k+1}\|_2^2\Big) \leq \frac{1}{K\gamma}\Big(\mathbf{E}[R_0] - R_*\Big) + \Big(\frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma_{\min}^A}\Big) \tag{73}$$

*by setting*

$$\mathcal{C}_2 = \frac{L_F}{q}, b_2 = \frac{27\ell_1^4 L_2^2 q}{L_F^2}, s = \frac{27\ell_1^4 q}{L_F^2}, b_1 = \frac{3\ell_2^2 L_1^2 q}{L_F^2}, \rho \geq \frac{\sqrt{78}L_F\kappa_G}{4\alpha}, \eta = \frac{2\alpha\sigma_{\min}(G)}{3L}(0 \leq \alpha \leq 1)$$

*Proof.* By the optimal condition of step 9 in algorithm 2, we will have:

$$0 =(y_j^k - y_j^{k+1})^T(\partial g_j(y_j^{k+1}) - B_j^T z_k + \rho B_j^T(Ax^k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) + H_j(y_j^{k+1} - y_j^k))$$

$$\leq g_j(y_j^k) - g_j(y_j^{k+1}) - (z_k)^T(B_j y_j^k - B_j y_j^{k+1}) + \rho(B_j y_j^k - B_j y_j^{k+1})^T\left(Ax^k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\right) - \|y_j^{k+1} - y_j^k\|_{H_j}^2$$

$$\leq g_j(y_j^k) - g_j(y_j^{k+1}) - z_k^T(Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c) + z_k^T(Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c)$$

$$+ \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j-1} B_i y_i^{k+1} + \sum_{i=j}^{m} B_i y_i^k - c\|_2^2 + \frac{\rho}{2}\|Ax_k + \sum_{i=1}^{j} B_i y_i^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c\|_2^2$$

$$- \frac{\rho}{2}\|B_j y_j^k - B_j y_j^{k+1}\|_2^2 - \|y_j^{k+1} - y_j^k\|_{H_j}^2$$

$$\leq \mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \frac{\rho}{2}\|B_j y_j^k - B_j y_j^{k+1}\|_2^2 - \|y_j^{k+1} - y_j^k\|_{H_j}^2$$

$$\leq \mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2$$

$$\tag{74}$$

Now, we will have the decrease bound on update the $y_j$ component is:

$$\mathcal{L}_\rho(x^k, y_j^{k+1}, y_{[j+1:m]}^k, z_k) - \mathcal{L}_\rho(x^k, y_{j-1}^{k+1}, y_{[j:m]}^k, z_k) \leq -\sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2 \tag{75}$$

Since we know that $F$ is $L_F$-smooth, we will have:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2$$

Now, using the optimal condition on updating the $x$ component in the algorithm, we will have

$$0 = (x^k - x^{k+1})^T\left(v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^T z_k + \rho A^T(Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c)\right) \tag{76}$$

Combine two equation above, we will have:

$$
\begin{aligned}
0 \leq\ & f(x^k) - f(x^{k+1}) + \nabla f(x_k)^T (x^{k+1} - x^k) + \frac{L_F}{2} \|x^{k+1} - x^k\|_2^2 \\
& + (x^k - x^{k+1})^T \left( v_k + \frac{G}{\eta}(x^{k+1} - x^k) - A^T z_k + \rho A^T (A x^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c) \right)
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & f(x^k) - f(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x_k - x_{k+1})^T (v_k - \nabla f(x^k)) \\
& - (z_k)^T (A x_k - A x_{k+1}) + \rho(A x_k - A x_{k+1})^T (A x_{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c)
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & f(x^k) - f(x^{k+1}) + \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x_k - x_{k+1})^T (v_k - \nabla f(x^k)) \\
& - z_k^T \left( A x^k + \sum_{j=1}^m B_j y_j^{k+1} - c \right) + z_k^T \left( A x^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c \right) \\
& + \frac{\rho}{2}\|A x^k + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2 - \frac{\rho}{2}\|A x^{k+1} + \sum_{j=1}^m B_j y_j^{k+1} - c\|_2^2 - \frac{\rho}{2}\|A x^k - A x^{k+1}\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
=\ & \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) \\
& + \frac{L_F}{2}\|x^{k+1} - x^k\|_2^2 - \frac{1}{\eta}\|x^k - x^{k+1}\|_G^2 + (x_k - x_{k+1})^T (v_k - \nabla f(x^k)) - \frac{\rho}{2}\|A x^k - A x^{k+1}\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
=\ & \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left( \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho \sigma_{\min}^A}{2} - \frac{L_F}{2} \right)\|x_{k+1} - x_k\|_2^2 + \langle x_k - x_{k+1}, v_k - \nabla f(x_k) \rangle
\end{aligned}
$$

$$
\begin{aligned}
=\ & \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left( \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho \sigma_{\min}^A}{2} - L_F \right)\|x_{k+1} - x_k\|_2^2 + \frac{1}{2L_F}\|v_k - \nabla f(x^k)\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
=\ & \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z_k) - \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_k) - \left( \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho \sigma_{\min}^A}{2} - L_F \right)\|x_{k+1} - x_k\|_2^2 \\
& + \frac{\mathcal{C}_2}{2L_F} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x_{i+1} - x_i\|_2^2 + \frac{\mathcal{C}_1}{2L_F}
\end{aligned}
$$

$$(77)$$

$$
\begin{aligned}
& \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^k) - \mathcal{L}_\rho(x^k, y_{[m]}^{k+1}, z^k) \\
& \leq - \left( \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho \sigma_{\min}^A}{2} - L_F \right)\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}_2}{2L_F} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x_{i+1} - x_i\|_2^2 + \frac{\mathcal{C}_1}{2L_F}
\end{aligned}
$$

$$(78)$$

Now, using the update of $z$ in the algorithm, we will have:

$$\mathcal{L}_\rho(x^{k+1}, y^{k+1}_{[m]}, z^{k+1}) - \mathcal{L}_\rho(x^{k+1}, y^{k+1}_{[m]}, z^k)$$

$$= \frac{1}{\rho}\|z^{k+1} - z^k\|_2^2$$

$$= \frac{6\mathcal{C}_1}{\rho\sigma^A_{\min}} + \frac{6\mathcal{C}_2}{\rho\sigma^A_{\min}} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2}\|x^{k+1} - x^k\|_2^2 + \left(\frac{3\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2} + \frac{9L_F^2}{\rho\sigma^A_{\min}}\right)\|x^{k-1} - x^k\|_2^2 \tag{79}$$

Now, combining (75),(78) and (79), we will have:

$$\mathcal{L}_\rho(x^{k+1}, y^{k+1}_{[m]}, z^{k+1}) - \mathcal{L}_\rho(x^k, y^k_{[m]}, z^{k+1})$$

$$\leq -\sum_{j=1}^m \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2 - \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L_F\right)\|x^{k+1} - x^k\|_2^2 + \frac{\mathcal{C}_2}{2L_F}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{\mathcal{C}_1}{2L_F}$$

$$+ \frac{6\mathcal{C}_1}{\rho\sigma^A_{\min}} + \frac{6\mathcal{C}_2}{\rho\sigma^A_{\min}}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2}\|x^{k+1} - x^k\|_2^2 + \left(\frac{3\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2} + \frac{9L^2}{\rho\sigma^A_{\min}}\right)\|x^{k-1} - x^k\|_2^2$$

$$\leq -\sum_{j=1}^m \sigma_{\min}(H_j)\|y_j^k - y_j^{k+1}\|_2^2 + \left(\frac{3\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2} + \frac{9L^2}{\rho\sigma^A_{\min}}\right)\|x^{k-1} - x^k\|_2^2$$

$$- \left(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma^A_{\min}}{2} - L - \frac{3\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2}\right)\|x^{k+1} - x^k\|_2^2$$

$$+ \left(\frac{\mathcal{C}_2}{2L} + \frac{6\mathcal{C}_2}{\rho\sigma^A_{\min}}\right)\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2$$

$$+ \frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma^A_{\min}} \tag{80}$$

Now we defined a useful potential function:

$$R_k = \mathcal{L}_\rho(x^k, y^k_{[m]}, z_k) + \left(\frac{3\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2} + \frac{9L_F^2}{\rho\sigma^A_{\min}}\right)\|x^{k-1} - x^k\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma^A_{\min}}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 \tag{81}$$

23

Now we can show that

$$
\begin{aligned}
R_{k+1} =& \mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z_{k+1}) + \Big(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho\sigma_{\min}^A}\Big)\|x^{k+1} - x^k\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k}\mathbf{E}\|x^{i+1} - x^i\|_2^2 \\
\leq& \mathcal{L}_\rho(x^k, y_{[m]}^k, z^k) + \Big(\frac{3\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} + \frac{9L_F^2}{\rho\sigma_{\min}^A}\Big)\|x^k - x^{k-1}\|_2^2 + \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 \\
& - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\|x^{k+1} - x^k\|_2^2 - \sigma_{\min}^H\sum_{j=1}^m\|y_j^k - y_j^{k+1}\|_2^2 \\
& + \Big(\frac{\mathcal{C}_2}{2L_F} + \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma_{\min}^A} \\
\\
\leq& R_k - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\|x^{k+1} - x^k\|_2^2 - \sigma_{\min}^H\sum_{j=1}^m\|y_j^k - y_j^{k+1}\|_2^2 \\
& + \Big(\frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma_{\min}^A}
\end{aligned}
\tag{82}
$$

Let $(n_k - 1)q \leq l \leq n_k q - 1$, telescoping the (82) over $k$ from $(n_k - 1)q$ to $k$ and take the expectation, we will have:

$$
\begin{aligned}
& \mathbf{E}[R_{k+1}] \\
\leq& \mathbf{E}[R_{(n_k-1)q}] - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\sum_{l=(n_k-1)q}^{k}\|x^{l+1} - x^l\|_2^2 \\
& - \sigma_{\min}^H\sum_{l=(n_k-1)q}^{k}\sum_{j=1}^m\|y_j^l - y_j^{l+1}\|_2^2 + \Big(\frac{\mathcal{C}_2}{2L} + \frac{6\mathcal{C}_2}{\rho\sigma_{\min}^A}\Big)\sum_{l=(n_k-1)q}^{k}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma_{\min}^A} \\
\\
\leq& \mathbf{E}[R_{(n_k-1)q}] - \Big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L_F - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2\mathcal{C}_2}{\rho\sigma_{\min}^A} - \frac{\mathcal{C}_2 q}{2L_F} + \frac{6\mathcal{C}_2 q}{\rho\sigma_{\min}^A}\Big)\sum_{l=(n_k-1)q}^{k}\|x^{l+1} - x^l\|_2^2 \\
& - \sigma_{\min}^H\sum_{l=(n_k-1)q}^{k}\sum_{j=1}^m\|y_j^l - y_j^{l+1}\|_2^2 + \Big(\frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma_{\min}^A}\Big)
\end{aligned}
\tag{83}
$$

Consider we have $\mathcal{C}_2 q = L_F^2$, let's denote

$$
\Lambda = \Big(\underbrace{\frac{\sigma_{\min}(G)}{\eta} - \frac{3L_F}{2}}_{\Lambda_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\rho\sigma_{\min}^A\eta^2} - \frac{9L_F^2}{\rho\sigma_{\min}^A} - \frac{2L_F^2}{\rho\sigma_{\min}^A q}}_{\Lambda_2}\Big)
$$

Now, choosing $0 \leq \eta \leq \frac{2\sigma_{\min}(G)}{3L}$, we will have $\Lambda_1 > 0$.

Further, let $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L_F}$ $(0 < \alpha < 1)$, since $b > 1 > \alpha^2, \kappa_G = \frac{\sigma^A_{\max}}{\sigma^A_{\min}} > 1$, we will have:

$$
\begin{aligned}
\Lambda_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\sigma^2_{\max}(G)}{\rho\sigma^A_{\min}\eta^2} - \frac{9L^2_F}{\rho\sigma^A_{\min}} - \frac{2L^2_F}{\rho\sigma^A_{\min}} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{27L^2_F\kappa^2_G}{2\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2_F}{\rho\sigma^A_{\min}} - \frac{2L^2_F}{\rho\sigma^A_{\min}q} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{27L^2_F\kappa^2_G}{2\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2_F\kappa^2_G}{\rho\sigma^A_{\min}\alpha^2} - \frac{2L^2_F\kappa^2_G}{\rho\sigma^A_{\min}\alpha^2} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{49L^2_F\kappa^2_G}{2\sigma^A_{\min}\rho\alpha^2} \\
&= \frac{\rho\sigma^A_{\min}}{4} + \frac{\rho\sigma^A_{\min}}{4} - \frac{49L^2_F\kappa^2_G}{2\sigma^A_{\min}\rho\alpha^2}
\end{aligned}
\tag{84}
$$

From the above result, we just need to choose the penalty $\rho \geq \frac{\sqrt{78}L_F\kappa_G}{\sigma^A_{\min}\alpha}$. Upon the result we have, we can argue that:

$$
\Lambda \geq \frac{\sqrt{78}L_F\kappa_G}{4\alpha}
$$

Also, by choosing $\mathcal{C}_2 = L^2_F/q$, we will have:

$$
\left( \frac{27\ell^4_1 L^2_2}{b_2} + \frac{27\ell^4_1}{s} + \frac{3\ell^2_2 L^2_1}{b_1} \right) = \frac{L^2_F}{q}
$$

We can have that:

$$
b_2 = \frac{27\ell^4_1 L^2_2 q}{L^2_F}, s = \frac{27\ell^4_1 q}{L^2_F}, b_1 = \frac{3\ell^2_2 L^2_1 q}{L^2_F}
$$

Since we know that $L^2_F \sim \mathcal{O}(\ell^4_1 + \ell^2_2)$, we can argue that $b_1, s, b_2 \sim \mathcal{O}(q)$.

Based on the structure of the potential function $R_k$, we want to show that $R_k$ is lower bounded.

$$\mathcal{L}_\rho(x^{k+1}, y_{[m]}^{k+1}, z^{k+1})$$

$$= f(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \langle z^{k+1}, Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c \rangle + \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq f(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \langle (A^T)^+ (v_k + \frac{G}{\eta}(x^{k+1}-x^k)), Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c \rangle + \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq f(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \langle (A^T)^+ (v_k - \nabla f(x^k) + \nabla f(x^k) + \frac{G}{\eta}(x^{k+1}-x^k)), Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c \rangle$$
$$+ \frac{\rho}{2}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq f(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \frac{2}{\sigma_{\min}^A \rho}\|v_k - \nabla f(x^k)\|_2^2 - \frac{2}{\sigma_{\min}^A}\|\nabla f(x^k)\|_2^2 - \frac{2\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x^{k+1}-x^k\|_2^2$$
$$+ \frac{\rho}{8}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$\geq f(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \frac{2L_F^2}{\sigma_{\min}^A q\rho} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1}-x^i\|_2^2 - \frac{2\mathcal{C}_1}{\sigma_{\min}^A \rho} - \frac{2\ell_1^2 \ell_2^2}{\sigma_{\min}^A \rho} - \frac{2\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x^{k+1}-x^k\|_2^2$$

$$\geq f(x^{k+1}) + \sum_{i=1}^{m} g_j(y_j^{k+1}) - \frac{2L_F^2}{\sigma_{\min}^A q\rho} \sum_{i=(n_k-1)q}^{k-1} \mathbf{E}\|x^{i+1}-x^i\|_2^2 - \frac{2\mathcal{C}_1}{\sigma_{\min}^A \rho} - \frac{2\ell_1^2 \ell_2^2}{\sigma_{\min}^A \rho} - \left(\frac{9L_F^2}{\sigma_{\min}^A \rho} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\right)\|x^{k+1}-x^k\|_2^2$$
$$\tag{85}$$

In all,

$$R_{k+1} \geq f(x_{k+1}) + \sum_{j=1}^{m} g_j(x_{k+1}) - \frac{2\ell_1^2 \ell_2^2}{\sigma_{\min}^A \rho} - \frac{2\mathcal{C}_1}{\sigma_{\min}^A \rho} \geq f^* + \sum_{j=1}^{m} g_j^* - \frac{2\ell_1^2 \ell_2^2}{\sigma_{\min}^A \rho} - \frac{2\mathcal{C}_1}{\sigma_{\min}^A \rho} \tag{86}$$

It follows that the potential function $R_k$ is bounded below. Let's denote the lower bound of $R_k$ is $R^*$. Now we sum up the (58) over all the iterates from 0 to $K$, we will have:

$$\mathbf{E}[R_k] - \mathbf{E}[R_0] \leq -\sum_{i=0}^{K-1}\left(\Lambda\|x^{i+1}-x^i\|_2^2 + \sigma_{\min}^H \sum_{j=1}^{m}\|y_j^i - y_j^{i+1}\|_2^2\right) + \left(\frac{\mathcal{C}_1 K}{2L_F} + \frac{6\mathcal{C}_1 K}{\rho\sigma_{\min}^A}\right) \tag{87}$$

Finally, we will have the iteration bound to be:

$$\boxed{\frac{1}{K}\sum_{k=0}^{K}\left(\|x^{k+1}-x^k\|_2^2 + \sum_{j=1}^{m}\|y_j^k - y_j^{k+1}\|_2^2\right) \leq \frac{1}{K\gamma}\left(\mathbf{E}[R_0] - \mathbf{E}[R_K]\right) + \left(\frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma_{\min}^A}\right)} \tag{88}$$

In which $\gamma = \min(\Lambda, \sigma_{\min}^H)$ and $\Lambda \geq \frac{\sqrt{78}L_F\kappa_G}{4\alpha}$. $\qquad\square$

**Lemma 9.7** (Stationary point convergence). *Suppose the sequence $\{x^k, y_{[m]}^k, z^k\}$ is generated from Algorithm (2), there exists a constant $\tilde{\nu}$ such that, with $T$ sampling uniformly from $1, ..., K$, we will have:*

$$\mathbf{E}\|\, dist(0, \partial L(x^T, y_{[m]}^T, z^T))\|_2^2 \leq \frac{9\tilde{\nu}}{K\gamma}(R_0 - R_*) + \frac{9\nu_{\max}}{\gamma}\left(\frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho\sigma_{\min}^A}\right) + 3\mathcal{C}_1 + \frac{6\mathcal{C}_1}{\rho^2 \sigma_{\min}^A} \tag{89}$$

26

*Proof.* Consider the sequence $\theta_k = \mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2 + \frac{1}{q}\sum_{i=(n_k-1)q}^{k}\|x^{i+1} - x^i\|_2^2 + \sum_{j=1}^m \|y_j^k - y_j^{k+1}\|_2^2]$.

Consider in the update of $y_i$ component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \partial_{y_i}\mathcal{L}(x, y_{[m]}, z))]_{k+1} = \mathbf{E}[\mathbf{dist}(0, \partial g_j(y_j^{k+1}) - B_j^T z^{k+1})]_{k+1}$$

$$=\mathbf{E}\|B_j^T z^k - \rho B_j^T(Ax^k + \sum_{i=1}^{j} B_j y_j^{k+1} + \sum_{i=j+1}^{m} B_i y_i^k - c) - H_j(y_j^{k+1} - y_j^k) - B_j^T z^{k+1}\|_2^2$$

$$=\mathbf{E}\|\rho B_j^T A(x^{k+1} - x^{k+1}) + \rho B_j^T \sum_{i=j+1}^{m} B_i(y_i^{k+1} - y_i^k) - H_j(y_j^{k+1} - y_j^k)\|_2^2 \tag{90}$$

$$\leq m\rho^2 \sigma_{\max}^{B_j}\sigma_{\max}^A \mathbf{E}\|x^{k+1} - x^k\|_2^2 + m\rho^2 \sigma_{\max}^{B_j}\sum_{i=j+1}^{m}\sigma_{\max}^{B_i}\mathbf{E}\|y_i^{k+1} - y_i^k\|_2^2 + m\sigma_{\max}^2(H_j)\mathbf{E}\|y_j^{k+1} - y_j^k\|_2^2$$

$$\leq m(\rho^2 \sigma_{\max}^B \sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H_j))\theta_k = \nu_1 \theta_k$$

In the updating of the $x$-component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \nabla_x \mathcal{L}(x, y_{[m]}, z))^2]_{k+1} = \mathbf{E}[\|A^T z^{k+1} - \nabla f(x^{k+1})\|_2^2]$$

$$\leq \mathbf{E}\|v_k - \nabla f(x^{k+1}) - \frac{G}{\eta}(x^k - x^{k+1}))\|_2^2$$

$$\leq \mathbf{E}\|v_k - \nabla f(x^k) + \nabla f(x^k) - \nabla f(x^{k+1}) - \frac{G}{\eta}(x^k - x^{k+1}))\|_2^2 \tag{91}$$

$$\leq \sum_{i=(n_k-1)q}^{k-1}\frac{3L_F^2}{q}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + 3\mathcal{C}_1 + 3(L_F^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\|x^k - x^{k+1}\|_2^2$$

$$\leq 3(L_F^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\theta_k + 3\mathcal{C}_1 = \nu_2 \theta_k + 3\mathcal{C}_1$$

In the updating of the $z$ component, we will have:

$$\mathbf{E}[\mathbf{dist}(0, \nabla_z \mathcal{L}(x, y_{[m]}, z))^2]_{k+1}$$

$$=\mathbf{E}\|Ax^{k+1} + \sum_{j=1}^{m} B_j y_j^{k+1} - c\|_2^2$$

$$=\frac{1}{\rho^2}\|z^{k+1} - z^k\|_2^2$$

$$\leq \frac{6\mathcal{C}_1}{\rho^2 \sigma_{\min}^A} + \frac{6\mathcal{C}_2}{\rho^2 \sigma_{\min}^A}\sum_{i=(n_k-1)q}^{k-1}\mathbf{E}\|x^{i+1} - x^i\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\rho^2\sigma_{\min}^A\eta^2}\|x^{k+1} - x^k\|_2^2 + (\frac{3\sigma_{\max}^2(G)}{\rho^2\sigma_{\min}^A\eta^2} + \frac{9L^2}{\rho^2\sigma_{\min}^A})\|x^{k-1} - x^k\|_2^2$$

$$\leq (\frac{9L^2}{\rho^2\sigma_{\min}^A} + \frac{3\sigma_{\max}^2 G}{\rho^2\sigma_A^2\eta^2})\theta_k + \frac{6\mathcal{C}_1}{\rho^2\sigma_{\min}^A} = \nu_3\theta_k + \frac{6\mathcal{C}_1}{\rho^2\sigma_{\min}^A}$$

$$\tag{92}$$

Since we know that:

$$\sum_{k=1}^{K-1}\theta_k = \sum_{k=1}^{K-1}\mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \|x^k - x^{k-1}\|_2^2 + \frac{1}{q}\sum_{i=(n_k-1)q}^{k}\|x^{i+1} - x^i\|_2^2 + \sum_{j=1}^{m}\|y_j^k - y_j^{k+1}\|_2^2$$

$$\leq 3\left(\sum_{k=1}^{K-1}\mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \sum_{k=1}^{K-1}\sum_{j=1}^{m}\mathbf{E}\|y_j^k - y_j^{k+1}\|_2^2\right)$$

27

Now, consider $T$ is chosen uniformly from $1, 2, ..., K-1, K$, we will have the following bound:

$$
\mathbf{E}\|\operatorname{\mathbf{dist}}(0, \partial L(x^T, y_{[m]}^T, z^T))\|_2^2
$$

$$
\leq \frac{3\nu_{\max}}{K} \sum_{k=1}^{K} \theta_k + 3\mathcal{C}_1 + \frac{6\mathcal{C}_1}{\rho^2 \sigma_{\min}^A}
$$

$$
\leq \frac{9\nu_{\max}}{K} \left( \sum_{k=1}^{K-1} \mathbf{E}[\|x^{k+1} - x^k\|_2^2 + \sum_{k=1}^{K-1} \sum_{j=1}^{m} \mathbf{E}\|y_j^k - y_j^{k+1}\|_2^2 \right) + 3\mathcal{C}_1 + \frac{6\mathcal{C}_1}{\rho^2 \sigma_{\min}^A} \tag{93}
$$

$$
\leq \frac{9\nu_{\max}}{K\gamma}(R_0 - R_*) + \frac{9\nu_{\max}}{\gamma} \left( \frac{\mathcal{C}_1}{2L_F} + \frac{6\mathcal{C}_1}{\rho \sigma_{\min}^A} \right) + 3\mathcal{C}_1 + \frac{6\mathcal{C}_1}{\rho^2 \sigma_{\min}^A}
$$

Given $\eta = \frac{2\alpha\sigma_{\min}(G)}{3L}$ $(0 < \alpha < 1)$ and $\Lambda \geq \frac{\sqrt{78}L_F\kappa_G}{4\alpha}$, with $T$ chosen uniformly from $1, 2, ..., K-1, K$, we will have:

$$
\mathbf{E}\|\operatorname{\mathbf{dist}}(0, \partial L(x^R, y_{[m]}^R, z^R))\|_2^2 \leq \mathcal{O}(\frac{1}{K}) + \mathcal{O}(\mathcal{C}_1) \tag{94}
$$

$\square$

**Theorem 9.2** (Total Sampling complexity). *Consider we want to achieve an $\epsilon$-stationary point solution, the total iteration complexity is $\mathcal{O}(\epsilon^{-2})$. We choose $\mathcal{C}_1 \sim \mathcal{O}(\epsilon^2)$ such that $B_1, B_2, S \sim \mathcal{O}(\epsilon^{-2})$. We choose $b_1, b_2, s \sim \mathcal{O}(\epsilon^{-1})$. The size of optimal epoch will be the same order as $b_1, b_2$. After $\mathcal{O}(\epsilon^{-2})$ iterations, the total sample complexity is $\mathcal{O}\left(\epsilon^{-3}\right)$*

# References

[Fang et al., 2018] Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699.

[Lin et al., 2018] Lin, T., Fan, C., Wang, M., and Jordan, M. I. (2018). Improved oracle complexity for stochastic compositional variance reduced gradient. *arXiv preprint arXiv:1806.00458*.

[Yu and Huang, 2017] Yu, Y. and Huang, L. (2017). Fast stochastic variance reduced admm for stochastic composition optimization. *arXiv preprint arXiv:1705.04138*.

[Zhang and Xiao, 2019] Zhang, J. and Xiao, L. (2019). Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*.