Zeroth-Order Stochastic Alternating Direction Method of Multipliers for Nonconvex Nonsmooth Optimization

Abstract

Alternating direction method of multipliers (ADMM) is a popular optimization tool for the composite and constrained problems in machine learning. However, in many machine learning problems such as black-box attacks and bandit feedback, ADMM could fail because the explicit gradients of these problems are difficult or infeasible to obtain. Zeroth-order (gradient-free) methods can solve these problems due to that the objective function values are only required in the optimization. Recently, though there exist a few zeroth-order ADMM methods, they build on the convexity of objective function. Clearly, these existing zeroth-order methods are limited in many applications. In the paper, thus, we propose a class of fast zeroth-order stochastic ADMM methods (*i.e.*, ZO-SVRG-ADMM and ZO-SAGA-ADMM) for solving nonconvex problems with multiple nonsmooth penalties, based on the coordinate smoothing gradient estimator. Moreover, we prove that both the ZO-SVRG-ADMM and ZO-SAGA-ADMM have convergence rate of $O(1/T)$, where T denotes the number of iteration. In particular, our methods not only reach the best convergence rate $O(1/T)$ for the nonconvex optimization, but also are able to effectively solve many machine learning problems with multiple complex regularized penalties. Finally, we conduct experiments of the black-box binary classification and the structured adversarial attack on black-box deep neural network to validate the efficiency of our algorithms.

1 Introduction

Alternating direction method of multipliers (ADMM [Gabay and Mercier, 1976; Boyd *et al.*, 2011]) is a popular optimization tool for solving the composite and constrained problems in machine learning. In particular, ADMM can efficiently optimize some complicated structure problems such as the graph-guided fused lasso [Kim *et al.*, 2009] and the overlapping group lasso, which are too complicated for the other popular optimization methods such as proximal gradient methods [Beck and Teboulle, 2009]. Thus, ADMM has been widely studied in recent years [Boyd *et al.*, 2011]. For the big data optimization, the stochastic ADMM method [Ouyang *et al.*, 2013] has been proposed. Due to variances of the stochastic gradient, however, these methods suffer from a slow convergence rate. To speedup the convergence, recently, some faster stochastic ADMM methods [Suzuki, 2014; Zheng and Kwok, 2016a] have been proposed by using the variance reduced techniques. In fact, ADMM is also highly successful in solving various nonconvex problems such as tensor decomposition and learning neural networks [Taylor *et al.*, 2016]. [Zheng and Kwok, 2016b] also accordingly have proposed the nonconvex stochastic ADMM method by using the variance reduced (VR) technique.

Currently, most of the ADMM methods need to compute repeatedly the gradient of the loss function over the iterations. However, in many machine learning problems, the explicit expression for gradient of objective function is difficult or infeasible to obtain. For example, in black-box situations, only prediction results (*i.e.*, function values) are provided [Chen *et al.*, 2017; Liu *et al.*, 2018c]. In bandit settings [Agarwal *et al.*, 2010], the player only receives partial feedback in terms of loss function values, so it is impossible to obtain the gradient of the full loss function. Clearly, the classic optimization methods, based on the first-order gradient or second-order information, are not competent to these problems. Thus, zeroth-order (gradient-free) optimization methods [Duchi *et al.*, 2015; Nesterov and Spokoiny, 2017] are developed by only using the function values to optimize these problems.

In the paper, we focus on using the zeroth-order methods to solve the following nonconvex nonsmooth problem:

$$\min_{x,\{y_j\}_{j=1}^k} F(x, y_{[k]}) =: \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \sum_{j=1}^{k} \psi_j(y_j) \quad (1)$$

$$\text{s.t. } Ax + \sum_{j=1}^{k} B_j y_j = c,$$

where $y_{[k]} = \{y_1, \cdots, y_k\}$, $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) : \mathbb{R}^d \to \mathbb{R}$ is a *nonconvex* and smooth function, and $\psi_j(y_j) : \mathbb{R}^q \to \mathbb{R}$ is a convex and possibly *nonsmooth* function for all $j \in [k]$, $k \geq 1$. In machine learning, function $f(x)$ can be used for the empirical loss, $\sum_{j=1}^{k} \psi_j(y_j)$ can be used for not only single structure penalty (*e.g.,* sparse) but also superposition

Table 1: Convergence properties comparison of the zeroth-order ADMM algorithms and other ones. C, NC, S, NS and mNS are the abbreviations of convex, non-convex, smooth, non-smooth and the sum of multiple non-smooth functions, respectively. T is the whole iteration number; d is the dimension of data and n denotes the sample size. Gaussian Smoothing Gradient Estimator (**GauSGE**), Uniform Smoothing Gradient Estimator (**UniSGE**) and Coordinate Smoothing Gradient Estimator (**CooSGE**).

Algorithm	Reference	Gradient estimator	Problem	Convergence rate
ZOO-ADMM	[Liu *et al.*, 2018b]	GauSGE	C(S) + C(NS)	$O(\sqrt{1/T})$
ZO-GADM	[Gao *et al.*, 2018]	UniSGE	C(S) + C(NS)	$O(\sqrt{1/T})$
RSPGF	[Ghadimi *et al.*, 2016]	GauSGE	S(NC) + NS(C)	$O(\sqrt{1/T})$
ZO-ProxSVRG ZO-ProxSAGA	[Huang *et al.*, 2019]	CooSGE	NC(S) + C(NS)	$O(1/T)$
ZO-SVRG-ADMM ZO-SAGA-ADMM	Ours	CooSGE	NC(S) + C(mNS)	$O(1/T)$

structures penalties (*e.g.*, sparse + group sparse). Due to the flexibility in splitting the objective function into loss $f(x)$ and regularizers $\psi_j(y_j)$ for all $j \in [k]$, ADMM is an efficient method to solve the above constricted problem. However, in the problem (1), we only access the objective values rather than the explicit function $F(x, y_{[k]})$, thus the classic ADMM methods are unsuitable for this problem.

Recently, [Gao *et al.*, 2018; Liu *et al.*, 2018b] proposed the zeroth-order stochastic ADMM methods, which only use the objective values to optimize, but these zeroth-order stochastic methods build on the convexity of objective function, so they are limited in many applications. In fact, in machine learning, there exist many nonconvex problems such as neural network training and tensor decomposition. Although, recently, some zeroth-order algorithms [Liu *et al.*, 2018c; Ghadimi *et al.*, 2016; Huang *et al.*, 2019] have been proposed for some simple non-convex problems, these methods are not suitable for the problem (1).

In the paper, thus, we propose a class of fast zeroth-order stochastic ADMM methods to solve the problem (1) based on the coordinate smoothing gradient estimator [Liu *et al.*, 2018c], which only uses the objective function values to optimize. In particular, the proposed zeroth-order stochastic methods build on the VR techniques, *i.e.*, SVRG [Johnson and Zhang, 2013] and SAGA [Defazio *et al.*, 2014], respectively. Moreover, we study the convergence properties of the proposed methods. Table 1 shows the convergence properties of the proposed methods and other related ones.

1.1 Challenges and Contributions

Although both SVRG and SAGA show good performances in the first-order and second-order methods, applying these techniques to the nonconvex zeroth-order ADMM method is *not trivial*. There are at least two **challenges**:

- Both SVRG and SAGA rely on the assumption that the stochastic gradient is an *unbiased* estimate of the true full gradient, which doesn't hold in zeroth-order algorithms;
- Due to failure of the Féjer monotonicity of iteration, the convergence analysis of the nonconvex ADMM is generally quite difficult [Wang *et al.*, 2015]. With using the inexact stochastic gradient, this difficulty becomes greater in the nonconvex zeroth-order ADMM methods.

In the paper, thus, we will fill this gap between the nonconvex zeroth-order ADMM and variance reduction methods. In summary, our major **contributions** are given below:

1) For solving the problem (1), we propose a class of fast zeroth-order stochastic ADMM methods (*i.e.*, ZO-SVRG-ADMM and ZO-SAGA-ADMM) based on the coordinate smoothing gradient estimator and the variance reduction techniques of SVRG and SAGA, respectively.
2) We prove that both the ZO-SVRG-ADMM and ZO-SAGA-ADMM have convergence rate of $O(\frac{1}{T})$ for nonconvex nonsmooth optimization. In particular, our methods not only reach the existing best convergence rate $O(\frac{1}{T})$ for the nonconvex optimization, but also are able to effectively solve many machine learning problems with multiple complex regularized penalties.
3) Extensive experiments conducted on black-box binary classification and structured adversarial attack on black-box deep neural networks validate efficiency of the proposed algorithms.

2 Related Works

Zeroth-order (gradient-free) optimization is a powerful optimization tool for solving many machine learning problems, where the gradient of objective function is not available or computationally prohibitive. For example, zeroth-order optimization methods have been applied to bandit feedback analysis [Agarwal *et al.*, 2010] and black-box attacks on deep neural networks (DDNs) [Chen *et al.*, 2017; Liu *et al.*, 2018c]. Recently, thus, the zeroth-order methods are widely studied. For example, [Nesterov and Spokoiny, 2017] have proposed several random zeroth-order methods by using Gaussian smoothing gradient estimator. Zeroth-order mirror descent method have been proposed in [Duchi *et al.*, 2015]. To solve the convex problems with the nonsmooth regularization, [Gao *et al.*, 2018; Liu *et al.*, 2018b] have proposed the zeroth-order online and stochastic ADMM methods.

So far, the above zeroth-order algorithms mainly build on the convexity of problems. In fact, zeroth-order algorithm is also highly successful in solving various nonconvex problems such as black-box adversarial attack to deep neural network [Liu *et al.*, 2018c; Liu *et al.*, 2018a]. Thus, recently [Liu *et al.*, 2018c; Liu *et al.*, 2018a] have begun to propose some zeroth-order stochastic methods for the nonconvex optimization. In addition, [Ghadimi *et al.*, 2016;

Huang *et al.*, 2019] have proposed the zeroth-order proximal stochastic gradient methods for the nonconvex problems with a simple nonsmooth regularization. However, these zeroth-order methods are not effective for solving some complex machine learning problems such as the overlapping group structured adversarial attack to black-box DDNs.

2.1 Notations

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

3 Preliminaries

In the section, we begin with restating a standard ϵ-approximate stationary point of the problem (1), as in [Jiang *et al.*, 2016; Zheng and Kwok, 2016b].

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

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

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

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

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

Next, we make some mild assumptions regarding problem (1) as follows:

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

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

which is equivalent to

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

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

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

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

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

4 Fast Zeroth-Order Stochastic ADMMs

In the section, we propose a class of zeroth-order stochastic ADMM methods to solve the problem (1). First, we define the augmented Lagrangian function of the problem (1) as follows:

$$\mathcal{L}_\rho(x, y_{[k]}, \lambda) = f(x) + \sum_{j=1}^k \psi_j(y_j) - \langle \lambda, Ax + \sum_{j=1}^k B_j y_j - c\rangle + \frac{\rho}{2}\|Ax + \sum_{j=1}^k B_j y_j - c\|^2, \tag{3}$$

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

In the problem (1), the explicit expression of objective function $f_i(x)$ is not available, and only the function value of $f_i(x)$ is available. To avoid computing explicit gradient, thus, we use the zeroth-order gradient estimator to estimate the gradient. Specifically, we use the coordinate smoothing function [Gu *et al.*, 2018; Liu *et al.*, 2018c] to estimate the gradients as follows: for $i \in [n]$,

$$\hat{\nabla} f_i(x) = \sum_{j=1}^d \frac{1}{2\mu_j}\big(f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j)\big)e_j, \tag{4}$$

where μ_j is a coordinate-wise smoothing parameter, and e_j is a standard basis vector with 1 at its j-th coordinate, and 0 otherwise.

Algorithm 1 Nonconvex ZO-SVRG-ADMM

Input: b, m, T, $S = [T/m]$, $\eta > 0$ and $\rho > 0$;
Initialize: $\tilde{x}^1 = x_0^1$, $y_j^{0,1}$ for $j \in [k]$ and λ_0^1;
for $s = 1, 2, \cdots, S$ **do**
  $\hat{\nabla} f(\tilde{x}^s) = \frac{1}{n}\sum_{i=1}^n \hat{\nabla} f_i(\tilde{x}^s)$;
  for $t = 0, 1, \cdots, m-1$ **do**
    Uniformly randomly pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n\}$, and $|\mathcal{I}_t| = b$;
    Using (4) to estimate stochastic gradient $\hat{g}_t^s = \hat{\nabla} f_{\mathcal{I}_t}(x_t^s) - \hat{\nabla} f_{\mathcal{I}_t}(\tilde{x}^s) + \hat{\nabla} f(\tilde{x}^s)$;
    $y_j^{s,t+1} = \arg\min_{y_j} \mathcal{L}_\rho(x_t^s, y_{[j-1]}^{s,t+1}, y_j, y_{[j+1:k]}^{s,t}, \lambda_t^s) + \frac{1}{2}\|y_j - y_j^{s,t}\|_{H_j}^2$, for all $j \in [k]$;
    $x_{t+1}^s = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[k]}^{s,t+1}, \lambda_t^s, \hat{g}_t^s\big)$;
    $\lambda_{t+1}^s = \lambda_t^s - \rho(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c)$;
  end for
  $\tilde{x}^{s+1} = x_0^{s+1} = x_m^s$, $y_j^{s+1,0} = y_j^{s,m}$ for all $j \in [k]$, $\lambda_0^{s+1} = \lambda_m^s$;
end for
Output: Iterate $\{x, y_{[k]}, \lambda\}$ chosen uniformly random from $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$.

Based on the above estimated gradients, we propose a zeroth-order ADMM (ZO-ADMM) method to solve the problem (1) by executing the following iterations, for $t =$

$0, 1, 2, \cdots$:

$$\begin{cases} y_j^{t+1} = \arg\min_{y_j} \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_j, y_{[j+1:k]}^t, \lambda_t) \\ \qquad + \frac{1}{2}\|y_j - y_j^t\|_{H_j}^2, \ \forall j \in [k] \\ x_{t+1} = \arg\min_x \hat{\mathcal{L}}_\rho(x, y_{t+1}, \lambda_t, \hat{g}_t) \\ \lambda_{t+1} = \lambda_t - \rho(Ax_{t+1} + By_{t+1} - c), \end{cases} \tag{5}$$

where

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

where $\hat{\nabla} f(x) = \frac{1}{n}\sum_{i=1}^n \hat{\nabla} f_i(x)$ and $\eta > 0$ is a step size. To adopt the following proximal operator to update y:

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

set $H_j = r_j I - \rho B_j^T B_j \succ I$ with $r_j > \rho\sigma_{\max}(B_j^T B_j) + 1$ to linearize the term $\|Ax_t + \sum_{j=1}^k B_j y_j - c\|^2$. At the same time, considering the matrix $A^T A$ is large, set $G = rI - \rho\eta A^T A \succ I$ with $r > \rho\eta\sigma_{\max}(A^T A) + 1$ to linearize the term $\|Ax + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2$.

Algorithm 2 Nonconvex ZO-SAGA-ADMM

1: **Input:** $b, T, \eta > 0$ and $\rho > 0$;
2: **Initialize:** $z_i^0 = x_0$ for $i \in \{1, 2, \cdots, n\}$, $\hat{\phi}_0 = \frac{1}{n}\sum_{i=1}^n \nabla f_i(z_i^0)$, y_j^0 for $j \in [k]$ and λ_0;
3: **for** $t = 0, 1, \cdots, T-1$ **do**
4: Uniformly randomly pick a mini-batch $\mathcal{I}_t$ (with replacement) from $\{1, 2, \cdots, n\}$, and $|\mathcal{I}_t| = b$;
5: Using (4) to estimate stochastic gradient $\hat{g}_t = \frac{1}{b}\sum_{i_t \in \mathcal{I}_t}\big(\nabla f_{i_t}(x_t) - \nabla f_{i_t}(z_{i_t}^t)\big) + \hat{\phi}_t$ with $\hat{\phi}_t = \frac{1}{n}\sum_{i=1}^n \nabla f_i(z_i^t)$;
6: $y_j^{t+1} = \arg\min_{y_j} \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_j, y_{[j+1:k]}^t, \lambda_t) + \frac{1}{2}\|y_j - y_j^t\|_{H_j}^2$, for all $j \in [k]$;
7: $x_{t+1} = \arg\min_x \hat{\mathcal{L}}_\rho\big(x, y_{[k]}^{t+1}, \lambda_t, \hat{g}_t\big)$;
8: $\lambda_{t+1} = \lambda_t - \rho(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c)$;
9: $z_{i_t}^{t+1} = x_t$ for $i \in \mathcal{I}_t$ and $z_i^{t+1} = z_i^t$ for $i \notin \mathcal{I}_t$;
10: $\hat{\phi}_{t+1} = \hat{\phi}_t - \frac{1}{n}\sum_{i_t \in \mathcal{I}_t}\big(\nabla f_{i_t}(z_{i_t}^t) - \nabla f_{i_t}(z_{i_t}^{t+1})\big)$;
11: **end for**
12: **Output:** Iterate $\{x, y_{[k]}, \lambda\}$ chosen uniformly random from $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$.

In the problem (1), not only the noisy gradient of $f_i(x)$ is not available, but also the sample size n is very large. In the paper, thus, we propose a class of fast zeroth-order stochastic ADMM methods (*i.e.*, ZO-SVRG-ADMM and ZO-SAGA-ADMM) to solve the problem (1), based on the variance reduced techniques of SVRG and SAGA.

Algorithm 1 shows the algorithmic framework of ZO-SVRG-ADMM. In Algorithm 1, we use the estimated stochastic gradient $\hat{g}_t^s = \hat{\nabla} f_{\mathcal{I}_t}(x_t^s) - \hat{\nabla} f_{\mathcal{I}_t}(\tilde{x}^s) + \hat{\nabla} f(\tilde{x}^s)$ with $\hat{\nabla} f_{\mathcal{I}_t}(x_t^s) = \frac{1}{b}\sum_{i_t \in \mathcal{I}_t} \hat{\nabla} f_{i_t}(x_t^s)$. We have $\mathbb{E}[\hat{g}_t^s] = \hat{\nabla} f(x_t^s) \neq \nabla f(x_t^s)$, *i.e.*, this stochastic gradient is a **biased** estimate of the true full gradient. Although the SVRG has shown a great promise, it relies upon the assumption that the stochastic gradient is an **unbiased** estimate of the true full gradient. Thus, adapting the similar ideas of SVRG to zeroth-order ADMM optimization is not a trivial task.

Algorithm 2 describes the algorithmic framework of ZO-SAGA-ADMM. In Algorithm 2, we use the estimated stochastic gradient $\hat{g}_t = \frac{1}{b}\sum_{i_t \in \mathcal{I}_t}\big(\hat{\nabla} f_{i_t}(x_t) - \nabla f_{i_t}(z_{i_t}^t)\big) + \hat{\phi}_t$ with $\hat{\phi}_t = \frac{1}{n}\sum_{i=1}^n \hat{\nabla} f_i(z_i^t)$. Similarly, we have $\mathbb{E}_{\mathcal{I}_t}[\hat{g}_t] = \hat{\nabla} f(x_t^s) \neq \nabla f(x_t)$, *i.e.*, this stochastic gradient is a **biased** estimate of the true full gradient.

5 Convergence Analysis

In the section, we will study the convergence properties of the proposed algorithms (ZO-SVRG-ADMM and ZO-SAGA-ADMM). For notational simplicity, let

$$\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big),$$
$$\nu_3 = \frac{18L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2},\ \nu_2 = 6L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2},\ \nu_4 = \frac{L}{4} + \frac{9L^2}{\sigma_{\min}^A}.$$

5.1 Convergence Analysis of ZO-SVRG-ADMM

In the subsection, we will study the convergence properties of the ZO-SVRG-ADMM.

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

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

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

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

It follows that

$$\frac{1}{T}\sum_{s=1}^S\sum_{t=0}^{m-1}(\sigma_{\min}^H\sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|_2^2 + \chi_t\|x_{t+1}^s - x_t^s\|^2) \le \frac{R_0^1 - R^*}{T} + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4},$$

where R^* denotes a lower bound of R_t^s and $\chi_t = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1 + 1/\beta)c_{t+1}$.

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

Theorem 1. *Suppose the sequence $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ is generated from Algorithm 1. Let $m = [n^{\frac{1}{3}}]$, $b = [d^{1-l}n^{\frac{2}{3}}]$, $l \in \{0, \frac{1}{2}, 1\}$, $\eta = \frac{\alpha\sigma_{\min}(G)}{9d^lL}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G d^lL}{\sigma_{\min}^A\alpha}$, then we have*

$$\min_{s,t} \mathbb{E}\big[dist(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \leq O(\frac{d^{2l}}{T}) + O(d^{2+2l}\mu^2),$$

where $\gamma = \min(\sigma_{\min}^H, \chi_t, L)$ with $\chi_t \geq \frac{3\sqrt{71\kappa_A}\kappa_G d^lL}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ and R^* is a lower bound of function R_t^s. It follows that suppose the smoothing parameter μ and the whole number of iteration $T = mS$ satisfy

$$\frac{1}{\mu^2} \geq \frac{2d^{2+2l}}{\epsilon}\max\big\{\nu_1\nu_4 + \frac{3L^2}{2}, \nu_2\nu_4 + \frac{9L^2}{\sigma_{\min}^A\rho^2}, \nu_3\nu_4\big\},$$
$$T = \frac{4\nu_{\max}(R_0^1 - R^*)}{\epsilon\gamma},$$

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

Remark 1. *Theorem 1 shows that given $m = [n^{\frac{1}{3}}]$, $b = [d^{1-l}n^{\frac{2}{3}}]$, $l \in \{0, \frac{1}{2}, 1\}$, $\eta = \frac{\alpha\sigma_{\min}(G)}{9d^lL}$ $(0 < \alpha \leq 1)$, $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G d^lL}{\sigma_{\min}^A\alpha}$ and $\mu = \frac{d}{\sqrt{T}}$, the ZO-SVRG-ADMM has $O(\frac{d^{2l}}{T})$ of convergence rate.*

5.2 Convergence Analysis of ZO-SAGA-ADMM

In the subsection, we will provide the convergence analysis of the ZO-SAGA-ADMM.

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

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\tfrac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\rho\eta^2} + \tfrac{9L^2}{\sigma_{\min}^A\rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A\rho b}\frac{1}{n}\sum_{i=1}^n\|x_{t-1} - z_i^{t-1}\|^2 + c_t\frac{1}{n}\sum_{i=1}^n\|x_t - z_i^t\|^2\big],$$

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

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

It follows that

$$\frac{1}{T}\sum_{t=1}^T(\sigma_{\min}^H\sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2 + \frac{Ld}{b}\frac{1}{n}\sum_{i=1}^n\|x_t - z_i^t\|^2 + \chi_t\|x_t - x_{t+1}\|^2) \leq \frac{\Omega_0 - \Omega^*}{T} + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4},$$

where $\chi_t = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1 + \frac{1-p}{\beta})c_{t+1}$ and Ω^* denotes a lower bound of Ω_t.

Let $\theta_t = \|x_{t+1} - x_t\|^2 + \|x_t - x_{t-1}\|^2 + \frac{d}{bn}\sum_{i=1}^n(\|x_t - z_i^t\|^2 + \|x_{t-1} - z_i^{t-1}\|^2) + \sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2$.

Theorem 2. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2. Let $b = n^{\frac{2}{3}}d^{\frac{1-l}{3}}$, $l \in \{0, \frac{1}{2}, 1\}$, $\eta = \frac{\alpha\sigma_{\min}(G)}{33d^lL}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{791\kappa_A}\kappa_G d^lL}{\sigma_{\min}^A\alpha}$ then we have*

$$\min_{1\leq t\leq T}\mathbb{E}\big[dist(0, \partial L(x_t, y_{[k]}^t, \lambda_t))^2\big] \leq O(\frac{d^{2l}}{T}) + O(d^{2+2l}\mu^2),$$

where $\gamma = \min(\sigma_{\min}^H, \chi_t, L)$ with $\chi_t \geq \frac{3\sqrt{791\kappa_A}\kappa_G d^lL}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ and Ω^* is a lower bound of function Ω_t. It follows that suppose the parameters μ and T satisfy

$$\frac{1}{\mu^2} \geq \frac{2d^{2+2l}}{\epsilon}\max\big\{\nu_1\nu_4 + \frac{3L^2}{2}, \nu_2\nu_4 + \frac{9L^2}{\sigma_{\min}^A\rho^2}, \nu_3\nu_4\big\},$$
$$T = \frac{4\kappa_{\max}}{\epsilon\gamma}(\Omega_0 - \Omega^*),$$

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

Remark 2. *Theorem 2 shows that $b = n^{\frac{2}{3}}d^{\frac{1-l}{3}}$, $l \in \{0, \frac{1}{2}, 1\}$, $\eta = \frac{\alpha\sigma_{\min}(G)}{33d^lL}$ $(0 < \alpha \leq 1)$, $\rho = \frac{6\sqrt{791\kappa_A}\kappa_G d^lL}{\sigma_{\min}^A\alpha}$ and $\mu = \frac{d}{\sqrt{T}}$, the ZO-SAGA-ADMM has the $O(\frac{d^{2l}}{T})$ of convergence rate.*

All related proofs are accessible in https://github.com/IJCAI-2019/IJCAI-2019-supp-code.

6 Experiments

In this section, we compare our algorithms (ZO-SVRG-ADMM, ZO-SAGA-ADMM) with the ZO-ProxSVRG, ZO-ProxSAGA [Huang *et al.*, 2019], the deterministic zeroth-order ADMM (ZO-ADMM), and zeroth-order stochastic ADMM (ZO-SGD-ADMM) without VR on two applications: **robust** black-box binary classification and **structured adversarial attacks** on black-box deep neural networks (DNNs).

Table 2: Real Datasets for Black-Box Binary Classification

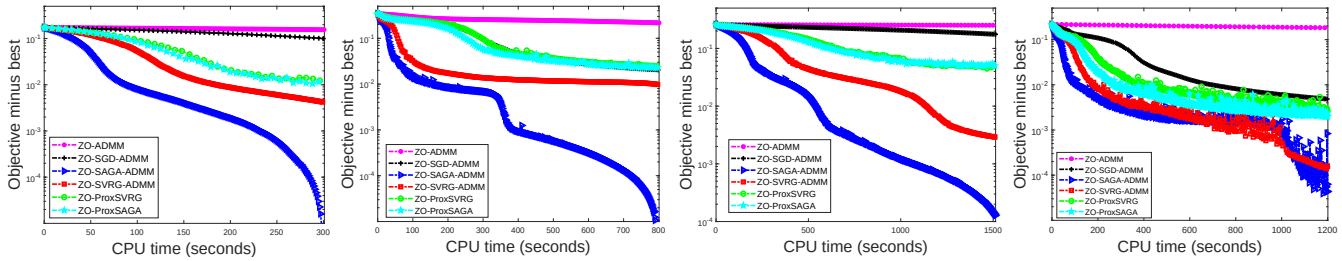
datasets	#samples	#features	#classes
20news	16,242	100	2
a9a	32,561	123	2
w8a	64,700	300	2
covtype.binary	581,012	54	2

6.1 Robust Black-Box Binary Classification

In this experiment, we apply the proposed algorithms to solve the robust black-box binary classification task with graph-guided fused lasso. Specifically, given a set of training samples $(a_i, l_i)_{i=1}^n$, where $a_i \in \mathbb{R}^d$ and $l_i \in \{-1, +1\}$, we find the optimal parameter $x \in \mathbb{R}^d$ by solving the problem:

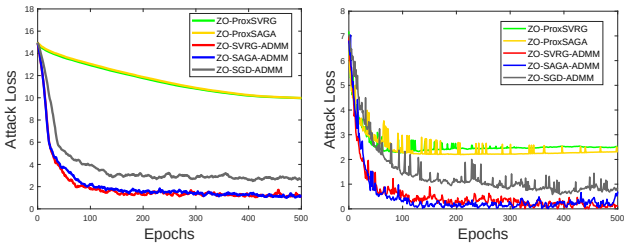
$$\min_{x\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^n f_i(x) + \tau_1\|x\|_1 + \tau_2\|\tilde{A}x\|_1, \tag{8}$$

where $f_i(x)$ is the black-box loss function, that only returns the function value given an input. Here, we specify the loss



(a) 20news (b) a9a (c) w8a (d) covtype.binary

Figure 1: Objective value gaps *versus* CPU time on benchmark datasets.



(a) *MNIST* (b) *CIFAR-10*

Figure 2: Attack loss on adversarial attacks black-box DNNs.

function $f_i(x) = \frac{\sigma^2}{2}\big(1 - \exp(-\frac{(l_i - a_i^T x)^2}{\sigma^2})\big)$, which is the *nonconvex* robust correntropy induced loss [He *et al.*, 2011]. $\tilde{A}$ decodes the sparsity pattern of graph obtained by sparse precision matrix estimation [Friedman *et al.*, 2008]. In the experiment, we give mini-batch size $b = 20$, smoothing parameter $\mu = \frac{1}{d\sqrt{t}}$ and penalty parameters $\tau_1 = \tau_1 = 10^{-5}$.

In the experiment, we use some public real datasets[1], which are summarized in Table 2. For each dataset, we use half of the samples as training data, and the rest as testing data. Figure 1 shows that the objective values of the proposed methods faster decrease than the other methods, as the CPU time increases. In particular, our methods show the better performances than the zeroth-order proximal stochastic methods.

6.2 Structured Attacks on Black-Box DNNs

In this experiment, we use our algorithms to generate adversarial examples to attack the pre-trained DNN models, whose parameters are hidden from us and only its outputs are accessible. Moreover, we consider the problem: "What possible structures could adversarial perturbations have to fool black-box DNNs ?" Thus, we use the zeroth-order algorithms to find an universal structured adversarial perturbation $x \in \mathbb{R}^d$ that could fool the samples $\{a_i \in \mathbb{R}^d,\ l_i \in \mathbb{N}\}_{i=1}^n$, which can be regarded as the following problem:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \max\big\{F_{l_i}(a_i + x) - \max_{j\neq l_i} F_j(a_i + x), 0\big\} + \tau_1 \sum_{p=1}^{P}\sum_{q=1}^{Q} \|x_{\mathcal{G}_{p,q}}\|_2 + \tau_2\|x\|_2^2, \quad (9)$$

[1] *20news* is from https://cs.nyu.edu/~roweis/data.html; others are from www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

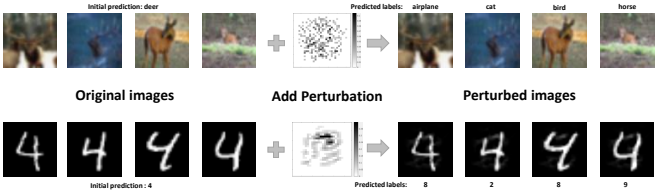


Figure 3: Group sparsity of perturbation is represented by heatmap.

where the overlapping groups $\{\mathcal{G}_{p,q}\}$, $p = 1, \cdots, P$, $q = 1, \cdots, Q$ generate from dividing an image into sub-groups of pixels, and $F(a)$ represents the final layer output before softmax of neural network. Here, following [Xu *et al.*, 2018], we use the overlapping lasso to obtain structured perturbations.

In the experiment, we use the pre-trained DNN models on MNIST and CIFAR-10 as the target black-box models, which can attain 99.4% and 80.8% test accuracy, respectively. For MNIST, we select 20 samples from a target class and set batch size $b = 4$; For CIFAR-10, we select 30 samples and set $b = 5$. In the experiment, we set $\mu = \frac{1}{d\sqrt{t}}$, where $d = 28\times 28$ and $d = 3\times 32\times 32$ for MNIST and CIFAR-10, respectively. Moreover, we set the penalty parameters $\tau_1 = 1$ and $\tau_2 = 2$. For both datasets, the kernel size for overlapped group lasso is set to 3×3 and the stride is 1.

Figure 2 shows that attack losses (*i.e.* the first term of the problem (9)) of our methods faster decrease than the other methods, as the number of iteration increases. Figure 3 shows that our algorithms can learn some group-wise sparse and interpretable structure perturbations, which can successfully attack the corresponding pre-trained DNN models.

7 Conclusions

In the paper, we have proposed fast ZO-SVRG-ADMM and ZO-SAGA-ADMM methods based on the coordinate smoothing gradient estimator, which only use the objective function values to optimization. Moreover, we prove that the proposed methods have convergence rate of $O(\frac{1}{T})$. In particular, our methods not only reach the existing best convergence rate $O(\frac{1}{T})$ for the nonconvex optimization, but also are able to effectively solve many machine learning problems with the complex nonsmooth regularizations. Our methods obtain state-of-the-art performances on both the black-box binary classification and structured adversarial attack on black-box DNNs.

References

[Agarwal *et al.*, 2010] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40. Citeseer, 2010.

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

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

[Chen *et al.*, 2017] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017.

[Defazio *et al.*, 2014] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

[Duchi *et al.*, 2015] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE TIT*, 61(5):2788–2806, 2015.

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

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

[Gao *et al.*, 2018] Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018.

[Ghadimi *et al.*, 2016] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[Gu *et al.*, 2018] Bin Gu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Faster derivative-free stochastic algorithm for shared memory machines. In *ICML*, pages 1807–1816, 2018.

[He *et al.*, 2011] Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE TPAMI*, 33(8):1561–1576, 2011.

[Huang *et al.*, 2019] Feihu Huang, Bin Gu, Zhouyuan Huo, Songcan Chen, and Heng Huang. Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In *AAAI*, 2019.

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

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

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

[Liu *et al.*, 2018a] Liu Liu, Minhao Cheng, Cho-Jui Hsieh, and Dacheng Tao. Stochastic zeroth-order optimization via variance reduction method. *CoRR*, abs/1805.11811, 2018.

[Liu *et al.*, 2018b] Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In *AISTATS*, volume 84, pages 288–297, 2018.

[Liu *et al.*, 2018c] Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In *NIPS*, pages 3731–3741, 2018.

[Nesterov and Spokoiny, 2017] Yurii Nesterov and Vladimir G. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

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

[Suzuki, 2014] Taiji Suzuki. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *ICML*, pages 736–744, 2014.

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

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

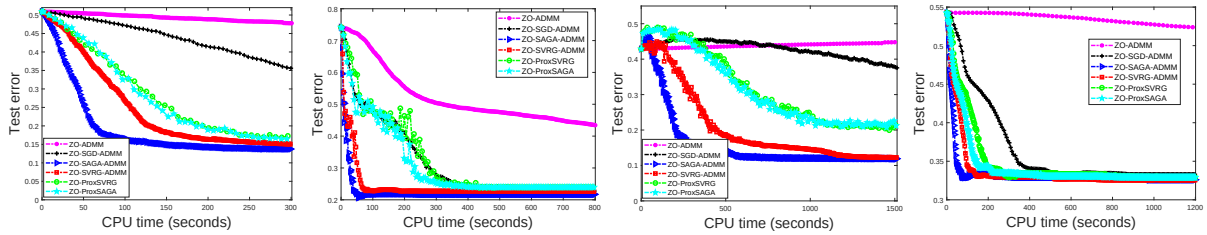
[Xu *et al.*, 2018] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. *arXiv preprint arXiv:1808.01664*, 2018.

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

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

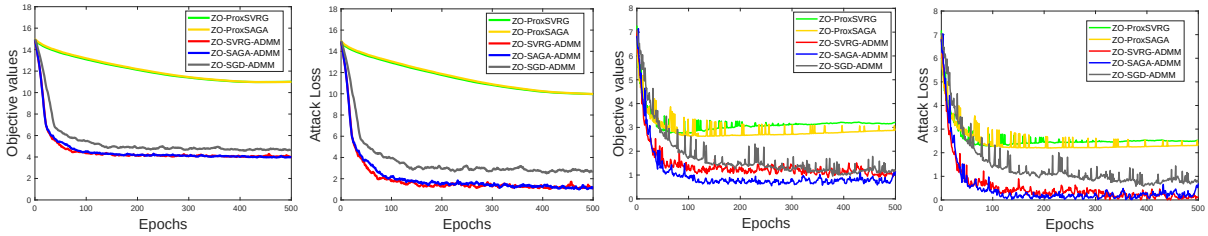
A Supplementary Materials

In this section, we begin with additional experimental results. Figure 4 shows that test loss of our algorithms faster decrease than the other methods in solving black-box binary classification problems, as the CPU time increases. Figure 5 shows that objective values and attack loss of our algorithms faster decrease than the comparative methods in attacking pre-trained black-box DNNs task. From Figure 6, we find that our algorithms can learn some group-wise sparse and interpretable structure perturbations, which can successfully attack the corresponding pre-trained DNN models.



(a) 20news (b) a9a (c) w8a (d) covtype.binary

Figure 4: Test loss *versus* CPU time on benchmark datasets in black-box binary classification task.



(a) *MNIST* (b) *MNIST* (c) *CIFAR-10* (d) *CIFAR-10*

Figure 5: Objective values and Attack loss *versus* CPU time on MNIST and CIFAR-10 datasets.

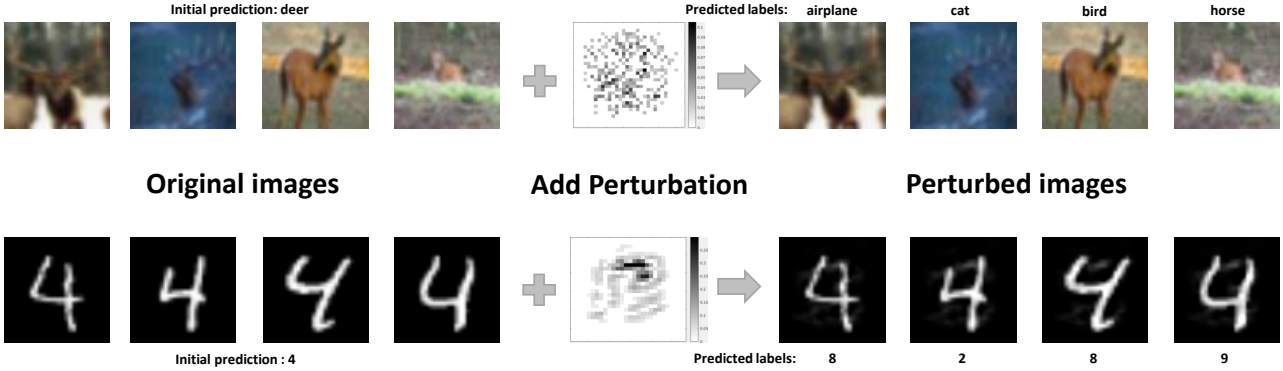


Figure 6: Group sparsity of perturbation is represented by heatmap.

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

Notations: To make the paper easier to follow, we give the following notations:

- $[k] = \{1, 2, \cdots, k\}$ and $[j : k] = \{j, j+1, \cdots, k\}$ for all $1 \leq j \leq k$.
- $\|\cdot\|$ denotes the vector ℓ_2 norm and the matrix spectral norm, respectively.
- $\|x\|_G = \sqrt{x^T G x}$, where G is a positive definite matrix.
- $\sigma^A_{\min}$ and $\sigma^A_{\max}$ denotes the minimum and maximum eigenvalues of $A^T A$, respectively; the conditional number $\kappa_A = \frac{\sigma^A_{\max}}{\sigma^A_{\min}}$.
- $\sigma^{B_j}_{\max}$ denotes the maximum eigenvalues of $B_j^T B_j$ for all $j \in [k]$, and $\sigma^B_{\max} = \max_{j=1}^k \sigma^{B_j}_{\max}$.
- $\sigma_{\min}(G)$ and $\sigma_{\max}(G)$ denote the minimum and maximum eigenvalues of matrix G, respectively; the conditional number $\kappa_G = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)}$.

- $\sigma_{\min}(H_j)$ and $\sigma_{\max}(H_j)$ denote the minimum and maximum eigenvalues of matrix H_j for all $j \in [k]$, respectively; $\sigma_{\min}(H) = \min_{j=1}^{k} \sigma_{\min}(H_j)$ and $\sigma_{\max}(H) = \max_{j=1}^{k} \sigma_{\max}(H_j)$.
- μ denotes the smoothing parameter of the gradient estimator.
- η denotes the step size of updating variable x.
- L denotes the Lipschitz constant of $\nabla f(x)$.
- b denotes the mini-batch size of stochastic gradient.
- T, m and S are the total number of iterations, the number of iterations in the inner loop, and the number of iterations in the outer loop, respectively.

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

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

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

$$\begin{aligned}\mathbb{E}\|\lambda_{t+1}^s - \lambda_t^s\|^2 \leq & \frac{18L^2 d}{\sigma_{\min}^A b}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2}\|x_{t+1}^s - x_t^s\|^2 \\ & + \big(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2} + \frac{9L^2}{\sigma_{\min}^A}\big)\|x_t^s - x_{t-1}^s\|^2 + \frac{9L^2 d^2 \mu^2}{\sigma_{\min}^A}. \end{aligned} \tag{10}$$

Proof. Using the optimal condition for the step 9 of Algorithm 1, we have

$$\hat{g}_t^s + \frac{1}{\eta} G(x_{t+1}^s - x_t^s) - A^T \lambda_t^s + \rho A^T (A x_{t+1}^s + \sum_{j=1}^{k} B_j y_j^{s,t+1} - c) = 0, \tag{11}$$

By the step 11 of Algorithm 1, we have

$$A^T \lambda_{t+1}^s = \hat{g}_t^s + \frac{1}{\eta} G(x_{t+1}^s - x_t^s). \tag{12}$$

Since

$$A^T(\lambda_{t+1}^s - \lambda_t^s) = \hat{g}_t^s - \hat{g}_{t-1}^s + \frac{G}{\eta}(x_{t+1}^s - x_t^s) - \frac{G}{\eta}(x_t^s - x_{t-1}^s), \tag{13}$$

then we have

$$\|\lambda_{t+1}^s - \lambda_t^s\|^2 \leq \frac{1}{\sigma_{\min}^A}\big[3\|\hat{g}_t^s - \hat{g}_{t-1}^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_{t+1}^s - x_t^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_t^s - x_{t-1}^s\|^2\big]. \tag{14}$$

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

$$\begin{aligned}\|\hat{g}_t^s - \hat{g}_{t-1}^s\|^2 &= \|\hat{g}_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) - \nabla f(x_{t-1}^s) + \nabla f(x_{t-1}^s) - \hat{g}_{t-1}^s\|^2 \\ &\leq 3\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 + 3\|\nabla f(x_t^s) - \nabla f(x_{t-1}^s)\|^2 + 3\|\nabla f(x_{t-1}^s) - \hat{g}_{t-1}^s\|^2 \\ &\leq \frac{6L^2 d}{b}\|x_t^s - \tilde{x}^s\|^2 + \frac{3L^2 d^2 \mu^2}{2} + \frac{6L^2 d}{b}\|x_{t-1}^s - \tilde{x}^s\|^2 + \frac{3L^2 d^2 \mu^2}{2} + 3\|\nabla f(x_t^s) - \nabla f(x_{t-1}^s)\|^2 \\ &\leq \frac{6L^2 d}{b}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + 3L^2\|x_t^s - x_{t-1}^s\|^2 + 3L^2 d^2 \mu^2, \end{aligned} \tag{15}$$

where the second inequality holds by Lemma 1 of [Huang *et al.*, 2019] and the third inequality holds by Assumption 1. Finally, combining (14) and (15), we obtain the above result. □

Lemma 4. *Suppose the sequence* $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ *is generated from Algorithm 1, and define a* Lyapunov *function:*

$$R_t^s = \mathbb{E}\big[\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + \big(\frac{3\kappa_A \sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} + \frac{9L^2}{\sigma_{\min}^A \rho}\big)\|x_t^s - x_{t-1}^s\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A \rho b}\|x_{t-1}^s - \tilde{x}^s\|^2 + c_t\|x_t^s - \tilde{x}^s\|^2\big] \tag{16}$$

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

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

It follows that

$$\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}(\sigma_{\min}^{H}\sum_{j=1}^{k}\|y_j^{s,t}-y_j^{s,t+1}\|^2+\frac{Ld}{b}\|x_t^s-\tilde{x}^s\|_2^2+\chi_t\|x_{t+1}^s-x_t^s\|^2)\le\frac{R_0^1-R^*}{T}+\frac{9L^2d^2\mu^2}{\sigma_{\min}^{A}\rho}+\frac{Ld^2\mu^2}{4}. \tag{17}$$

where R^* denotes a lower bound of R_t^s and $\chi_t=\frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^{A}}{2}-L-\frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^{A}\eta^2\rho}-\frac{9L^2}{\sigma_{\min}^{A}\rho}-(1+1/\beta)c_{t+1}$.

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

$$\begin{aligned}
0&=(y_j^{s,t}-y_j^{s,t+1})^T\big(\partial\psi_j(y_j^{s,t+1})-B^T\lambda_t^s+\rho B^T(Ax_t^s+\sum_{i=1}^{j}B_iy_i^{s,t+1}+\sum_{i=j+1}^{k}B_iy_i^{s,t}-c)+H_j(y_j^{s,t+1}-y_j^{s,t})\big)\\
&\le\psi_j(y_j^{s,t})-\psi_j(y_j^{s,t+1})-(\lambda_t^s)^T(B_jy_j^{s,t}-B_jy_j^{s,t+1})+\rho(By_j^{s,t}-By_j^{s,t+1})^T(Ax_t^s+\sum_{i=1}^{j}B_iy_i^{s,t+1}+\sum_{i=j+1}^{k}B_iy_i^{s,t}-c)\\
&\quad-\|y_j^{s,t+1}-y_j^{s,t}\|_{H_j}^2\\
&=\psi_j(y_j^{s,t})-\psi_j(y_j^{s,t+1})-(\lambda_t^s)^T(Ax_t^s+\sum_{i=1}^{j-1}B_iy_i^{s,t+1}+\sum_{i=j}^{k}B_iy_i^{s,t}-c)+(\lambda_t^s)^T(Ax_t^s+\sum_{i=1}^{j}B_iy_i^{s,t+1}+\sum_{i=j+1}^{k}B_iy_i^{s,t}-c)\\
&\quad+\frac{\rho}{2}\|Ax_t^s+\sum_{i=1}^{j-1}B_iy_i^{s,t+1}+\sum_{i=j}^{k}B_iy_i^{s,t}-c\|^2-\frac{\rho}{2}\|Ax_t^s+\sum_{i=1}^{j}B_iy_i^{s,t+1}+\sum_{i=j+1}^{k}B_iy_i^{s,t}-c\|^2-\|y_j^{s,t+1}-y_j^{s,t}\|_{H_j}^2\\
&\quad-\frac{\rho}{2}\|B_jy_j^{s,t}-B_jy_j^{s,t+1}\|^2\\
&=\underbrace{f(x_t^s)+\sum_{i=1}^{j-1}\psi_i(y_i^{s,t+1})+\sum_{i=j}^{k}\psi_i(y_i^{s,t})-(\lambda_t^s)^T(Ax_t^s+\sum_{i=1}^{j-1}B_iy_i^{s,t+1}+\sum_{i=j}^{k}B_iy_i^{s,t}-c)+\frac{\rho}{2}\|Ax_t^s+\sum_{i=1}^{j-1}B_iy_i^{s,t+1}+\sum_{i=j}^{k}B_iy_i^{s,t}-c\|^2}_{\mathcal{L}_\rho(x_t^s,y_{[j-1]}^{s,t+1},y_{[j:k]}^{s,t},\lambda_t^s)}\\
&\quad-\underbrace{\big(f(x_t^s)+\sum_{i=1}^{j}\psi_i(y_i^{s,t+1})+\sum_{i=j+1}^{k}\psi_i(y_i^{s,t})-(\lambda_t^s)^T(Ax_t^s+\sum_{i=1}^{j}B_iy_i^{s,t+1}+\sum_{i=j+1}^{k}B_iy_i^{s,t}-c)+\frac{\rho}{2}\|Ax_t^s+\sum_{i=1}^{j}B_iy_i^{s,t+1}+\sum_{i=j+1}^{k}B_iy_i^{s,t}-c\|^2\big)}_{\mathcal{L}_\rho(x_t^s,y_{[j]}^{s,t+1},y_{[j+1:k]}^{s,t},\lambda_t^s)}\\
&\quad-\|y_j^{s,t+1}-y_j^{s,t}\|_{H_j}^2-\frac{\rho}{2}\|B_jy_j^{s,t}-B_jy_j^{s,t+1}\|^2\\
&\le\mathcal{L}_\rho(x_t^s,y_{[j-1]}^{s,t+1},y_{[j:k]}^{s,t},\lambda_t^s)-\mathcal{L}_\rho(x_t^s,y_{[j]}^{s,t+1},y_{[j+1:k]}^{s,t},\lambda_t^s)-\sigma_{\min}(H_j)\|y_j^{s,t}-y_j^{s,t+1}\|^2,
\end{aligned}\tag{18}$$

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

$$\mathcal{L}_\rho(x_t^s,y_{[j]}^{s,t+1},y_{[j+1:k]}^{s,t},\lambda_t^s)\le\mathcal{L}_\rho(x_t^s,y_{[j-1]}^{s,t+1},y_{[j:k]}^{s,t},\lambda_t^s)-\sigma_{\min}(H_j)\|y_j^{s,t}-y_j^{s,t+1}\|^2. \tag{19}$$

Telescoping inequality (19) over j from 1 to k, we obtain

$$\mathcal{L}_\rho(x_t^s,y_{[k]}^{s,t+1},\lambda_t^s)\le\mathcal{L}_\rho(x_t^s,y_{[k]}^{s,t},\lambda_t^s)-\sigma_{\min}^{H}\sum_{j=1}^{k}\|y_j^{s,t}-y_j^{s,t+1}\|^2, \tag{20}$$

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

By Assumption 1, we have

$$0\le f(x_t^s)-f(x_{t+1}^s)+\nabla f(x_t^s)^T(x_{t+1}^s-x_t^s)+\frac{L}{2}\|x_{t+1}^s-x_t^s\|^2. \tag{21}$$

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

$$0=(x_t^s-x_{t+1}^s)^T\big(\hat{g}_t^s-A^T\lambda_t^s+\rho A^T(Ax_{t+1}^s+\sum_{j=1}^{k}B_jy_j^{s,t+1}-c)+\frac{G}{\eta}(x_{t+1}^s-x_t^s)\big). \tag{22}$$

Combining (21) and (22), 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(\hat{g}_t^s - A^T\lambda_t^s + \rho A^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{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(\hat{g}_t^s - \nabla f(x_t^s)) \\
& - (\lambda_t^s)^T(Ax_t^s - Ax_{t+1}^s) + \rho(Ax_t^s - Ax_{t+1}^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) \\
\overset{(i)}{=}\ & f(x_t^s) - f(x_{t+1}^s) + \frac{L}{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(\hat{g}_t^s - \nabla f(x_t^s)) - (\lambda_t^s)^T(Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) \\
& + (\lambda_t^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\big(\|Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 - \|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 - \|Ax_t^s - Ax_{t+1}^s\|^2\big) \\
=\ & \underbrace{f(x_t^s) + \sum_{j=1}^k \psi_j(x_{t+1}^s) - (\lambda_t^s)^T(Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_t^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2}_{\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s)} \\
& \underbrace{- \big(f(x_{t+1}^s) + \sum_{j=1}^k \psi_j(x_{t+1}^s) - (\lambda_t^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2\big)}_{\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s)} \\
& + \frac{L}{2}\|x_t^s - x_{t+1}^s\|^2 + (x_t^s - x_{t+1}^s)^T(\hat{g}_t^s - \nabla f(x_t^s)) - \frac{1}{\eta}\|x_t^s - x_{t+1}^s\|_G^2 - \frac{\rho}{2}\|Ax_t^s - Ax_{t+1}^s\|^2 \\
\leq\ & \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - \frac{L}{2})\|x_t^s - x_{t+1}^s\|^2 + (x_t^s - x_{t+1}^s)^T(\hat{g}_t^s - \nabla f(x_t^s)) \\
\overset{(ii)}{\leq}\ & \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 + \frac{1}{2L}\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 \\
\overset{(iii)}{\leq}\ & \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 + \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|^2 + \frac{Ld^2\mu^2}{4},
\end{aligned}
\tag{23}
$$

where the equality (i) holds by applying the equality $(a-b)^Tb = \frac{1}{2}(\|a\|^2 - \|b\|^2 - \|a-b\|^2)$ on the term $(Ax_t^s - Ax_{t+1}^s)^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c)$, the inequality (ii) holds by the inequality $a^Tb \leq \frac{L}{2}\|a\|^2 + \frac{1}{2L}\|b\|^2$, and the inequality (iii) holds by Lemma 1 of [Huang *et al.*, 2019]. Thus, we obtain

$$
\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) \leq \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t+1}, \lambda_t^s) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 + \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|^2 + \frac{Ld^2\mu^2}{4}. \tag{24}
$$

Using the step 10 in Algorithm 1, we have

$$
\begin{aligned}
\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) - \mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_t^s) &= \frac{1}{\rho}\|\lambda_{t+1}^s - \lambda_t^s\|^2 \\
&\leq \frac{18L^2d}{\sigma_{\min}^A b\rho}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2 \\
&\quad + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho}.
\end{aligned}
\tag{25}
$$

Combining (20), (24) and (25), we have

$$\begin{aligned}\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) \le \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) - \sigma_{\min}^H \sum_{j=1}^k \|y_j^{s,t} - y_j^{s,t+1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t^s - x_{t+1}^s\|^2 \\ + \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|^2 + \frac{18L^2d}{\sigma_{\min}^A b\rho}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2 \\ + (\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4}. \end{aligned} \tag{26}$$

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

$$R_t^s = \mathbb{E}\big[\mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A\rho b}\|x_{t-1}^s - \tilde{x}^s\|^2 + c_t\|x_t^s - \tilde{x}^s\|^2\big]. \tag{27}$$

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

$$\begin{aligned}\|x_{t+1}^s - x_t^s + x_t^s - \tilde{x}^s\|^2 &= \|x_{t+1}^s - x_t^s\|^2 + 2(x_{t+1}^s - x_t^s)^T(x_t^s - \tilde{x}^s) + \|x_t^s - \tilde{x}^s\|^2 \\ &\le \|x_{t+1}^s - x_t^s\|^2 + 2\big(\frac{1}{2\beta}\|x_{t+1}^s - x_t^s\|^2 + \frac{\beta}{2}\|x_t^s - \tilde{x}^s\|^2\big) + \|x_t^s - \tilde{x}^s\|^2 \\ &= (1 + 1/\beta)\|x_{t+1}^s - x_t^s\|^2 + (1+\beta)\|x_t^s - \tilde{x}^s\|^2, \end{aligned} \tag{28}$$

where the above inequality holds by the Cauchy-Schwarz inequality with $\beta > 0$. Combining (27) with (28), then we obtain

$$\begin{aligned}R_{t+1}^s &= \mathbb{E}\big[\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_{t+1}^s - x_t^s\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A b\rho}\|x_t^s - \tilde{x}^s\|^2 + c_{t+1}\|x_{t+1}^s - \tilde{x}^s\|^2\big] \\ &\le \mathcal{L}_\rho(x_t^s, y_{[k]}^{s,t}, \lambda_t^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t^s - x_{t-1}^s\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A b\rho}\|x_{t-1}^s - \tilde{x}^s\|^2 \\ &\quad + \big(\frac{36\kappa_A L^2 d}{\sigma_{\min}^A b\rho} + \frac{2Ld}{b} + (1+\beta)c_{t+1}\big)\|x_t^s - \tilde{x}^s\|^2 - \sigma_{\min}^H\sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2 \\ &\quad - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+1/\beta)c_{t+1}\big)\|x_t^s - x_{t+1}^s\|^2 \\ &\quad - \frac{3(\kappa_A - 1)\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_t^s - x_{t+1}^s\|^2 - \frac{18(\kappa_A - 1)L^2 d}{\sigma_{\min}^A b\rho}\|x_{t-1}^s - \tilde{x}^s\|^2 - \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|^2 + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4} \\ &\le R_t^s - \sigma_{\min}^H\sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2 - \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|^2 - \chi_t\|x_t^s - x_{t+1}^s\|^2 + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4}, \end{aligned} \tag{29}$$

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

Next, we will prove the relationship between R_1^{s+1} and R_m^s. Due to $x_0^{s+1} = x_m^s = \tilde{x}^{s+1}$, we have

$$\hat{g}_0^{s+1} = \hat{\nabla} f_{\mathcal{I}}(x_0^{s+1}) - \hat{\nabla} f_{\mathcal{I}}(x_0^{s+1}) + \hat{\nabla} f(x_0^{s+1}) = \hat{\nabla} f(x_0^{s+1}) = \hat{\nabla} f(x_m^s). \tag{30}$$

It follows that

$$\begin{aligned}\mathbb{E}\|\hat{g}_0^{s+1} - \hat{g}_m^s\|^2 &= \mathbb{E}\|\hat{\nabla} f(x_m^s) - \hat{\nabla} f_{\mathcal{I}}(x_m^s) + \hat{\nabla} f_{\mathcal{I}}(\tilde{x}^s) - \hat{\nabla} f(\tilde{x}^s)\|^2 \\ &= \|\hat{\nabla} f_{\mathcal{I}}(x_m^s) - \hat{\nabla} f_{\mathcal{I}}(\tilde{x}^s) - \mathbb{E}_{\mathcal{I}}[\hat{\nabla} f_{\mathcal{I}}(x_m^s) - \hat{\nabla} f_{\mathcal{I}}(\tilde{x}^s)]\|^2 \\ &\le \|\hat{\nabla} f_{\mathcal{I}}(x_m^s) - \hat{\nabla} f_{\mathcal{I}}(\tilde{x}^s)\|^2 \\ &\le \frac{1}{b}\sum_{i\in\mathcal{I}}\|\hat{\nabla} f_i(x_m^s) - \hat{\nabla} f_i(\tilde{x}^s)\|^2 \\ &= \frac{1}{b}\sum_{i\in\mathcal{I}}\|\hat{\nabla} f_i(x_m^s) - \hat{\nabla} f_i(\tilde{x}^s)\|^2 \\ &\le \frac{L^2 d}{b}\|x_m^s - \tilde{x}^s\|^2, \end{aligned} \tag{31}$$

where the first inequality holds by the inequality $\mathbb{E}\|\zeta - \mathbb{E}\zeta\|^2 = \mathbb{E}\|\zeta\|^2 - \|\mathbb{E}\zeta\|^2$; the third inequality holds by the definition of zeroth-order gradient (4).

By Lemma 3, we have

$$\begin{aligned}
\|\lambda_1^{s+1} - \lambda_m^s\|^2 &\leq \frac{1}{\sigma_{\min}^A}\|\hat{g}_0^{s+1} - \hat{g}_m^s + \frac{G}{\eta}(x_1^{s+1} - x_0^{s+1}) + \frac{G}{\eta}(x_m^s - x_{m-1}^s)\|^2 \\
&= \frac{1}{\sigma_{\min}^A}\|\hat{\nabla} f(x_m^s) - \hat{g}_m^s + \frac{G}{\eta}(x_1^{s+1} - x_m^s) + \frac{G}{\eta}(x_m^s - x_{m-1}^s)\|^2 \\
&\leq \frac{1}{\sigma_{\min}^A}\big(3\|\hat{\nabla} f(x_m^s) - \hat{g}_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_m^s - x_{m-1}^s\|^2\big) \\
&\leq \frac{1}{\sigma_{\min}^A}\big(3\|\hat{\nabla} f(x_m^s) - \hat{g}_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_m^s - x_{m-1}^s\|^2\big) \\
&\leq \frac{1}{\sigma_{\min}^A}\big(\frac{3L^2 d}{b}\|x_m^s - \tilde{x}^s\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_m^s - x_{m-1}^s\|^2\big). \qquad (32)
\end{aligned}$$

Since $x_m^s = x_0^{s+1}$, $y_j^{s,m} = y_j^{s+1,0}$ for all $j \in [k]$ and $\lambda_m^s = \lambda_0^{s+1}$, by (20), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) &\leq \mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,0}, \lambda_0^{s+1}) - \sigma_{\min}^H \sum_{j=1}^k \|y_j^{s+1,0} - y_j^{s+1,1}\|^2 \\
&= \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) - \sigma_{\min}^H \sum_{j=1}^k \|y_j^{s,m} - y_j^{s+1,1}\|^2. \qquad (33)
\end{aligned}$$

By (24), we have

$$\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) \leq \mathcal{L}_\rho(x_0^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_0^{s+1} - x_1^{s+1}\|^2 + \frac{Ld^2\mu^2}{4}. \qquad (34)$$

By (25), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) &\leq \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) + \frac{1}{\rho}\|\lambda_1^{s+1} - \lambda_0^{s+1}\|^2 \\
&\leq \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_0^{s+1}) + \frac{1}{\sigma_{\min}^A \rho}\big(\frac{3L^2 d}{b}\|x_m^s - \tilde{x}^s\|_2^2 \\
&\quad + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_m^s - x_{m-1}^s\|^2\big). \qquad (35)
\end{aligned}$$

where the second inequality holds by (32).

Combining (33), (34) with (35), we have

$$\begin{aligned}
\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) &\leq \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) - \sigma_{\min}^H \sum_{j=1}^k \|y_j^{s,m} - y_j^{s+1,1}\|^2 - (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_0^{s+1} - x_1^{s+1}\|^2 \\
&\quad + \frac{1}{\sigma_{\min}^A \rho}\big(\frac{3L^2 d}{b}\|x_m^s - \tilde{x}^s\|_2^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_1^{s+1} - x_m^s\|^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\|x_m^s - x_{m-1}^s\|^2\big) + \frac{Ld^2\mu^2}{4}. \qquad (36)
\end{aligned}$$

Therefore, we have

$$
\begin{aligned}
R_1^{s+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_1^{s+1} - x_0^{s+1}\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A b\rho}\|x_0^{s+1} - \tilde{x}^{s+1}\|^2 + c_1\|x_1^{s+1} - \tilde{x}^{s+1}\|^2\big] \\
&= \mathcal{L}_\rho(x_1^{s+1}, y_{[k]}^{s+1,1}, \lambda_1^{s+1}) + \big(\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho} + c_1\big)\|x_1^{s+1} - x_0^{s+1}\|^2 \\
&\le \mathcal{L}_\rho(x_m^s, y_{[k]}^{s,m}, \lambda_m^s) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_m^s - x_{m-1}^s\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A\rho b}\|x_{m-1}^s - \tilde{x}^s\|_2^2 + (\frac{36\kappa_A L^2 d}{\sigma_{\min}^A\rho b} + \frac{2Ld}{b})\|x_m^s - \tilde{x}^s\|_2^2 \\
&\quad - \sigma_{\min}^H\sum_{j=1}^k\|y_j^{s,m} - y_j^{s+1,1}\|^2 - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - c_1\big)\|x_1^{s+1} - x_m^s\|_2^2 \\
&\quad - \frac{9L^2}{\sigma_{\min}^A\rho}\|x_m^s - x_{m-1}^s\|_2^2 - \frac{18\kappa_A L^2 d}{\sigma_{\min}^A\rho b}\|x_{m-1}^s - \tilde{x}^s\|_2^2 - \frac{3(\kappa_A-1)\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}\|x_1^{s+1} - x_m^s\|_2^2 \\
&\quad - (\frac{33\kappa_A L^2 d}{\sigma_{\min}^A\rho b} + \frac{3(\kappa_A-1)L^2 d}{\sigma_{\min}^A\rho b} + \frac{2Ld}{b})\|x_m^s - \tilde{x}^s\|_2^2 + \frac{Ld^2\mu^2}{4} \\
&\le R_m^s - \sigma_{\min}^H\sum_{j=1}^k\|y_j^{s,m} - y_j^{s+1,1}\|^2 - \frac{Ld}{b}\|x_m^s - \tilde{x}^s\|_2^2 - \chi_m\|x_1^{s+1} - x_m^s\|^2 + \frac{Ld^2\mu^2}{4},
\end{aligned} \tag{37}
$$

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

By (12), we have

$$
\lambda_{t+1} = (A^T)^+(\hat{\nabla}f(x_t) + \frac{G}{\eta}(x_{t+1} - x_t)), \tag{38}
$$

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

Then we have

$$
\begin{aligned}
&\mathcal{L}_\rho(x_{t+1}^s, y_{[k]}^{s,t+1}, \lambda_{t+1}^s) = f(x_{t+1}^s) + \sum_{j=1}^k\psi_j(y_j^{s,t+1}) - \lambda_{t+1}^T(Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c) + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&= f(x_{t+1}^s) + \sum_{j=1}^k\psi_j(y_j^{s,t+1}) - \langle(A^T)^+(\hat{g}_t^s + \frac{G}{\eta}(x_{t+1}^s - x_t^s)), Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\rangle + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&= f(x_{t+1}^s) + \sum_{j=1}^k\psi_j(y_j^{s,t+1}) - \langle(A^T)^+(\hat{g}_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) + \frac{G}{\eta}(x_{t+1}^s - x_t^s)), Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\rangle \\
&\quad + \frac{\rho}{2}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&\ge f(x_{t+1}^s) + \sum_{j=1}^k\psi_j(y_j^{s,t+1}) - \frac{5\kappa_A}{2\sigma_{\min}^A\rho}\|\hat{g}_t^s - \nabla f(x_t^s)\|^2 - \frac{5\kappa_A}{2\sigma_{\min}^A\rho}\|\nabla f(x_t^s)\|^2 - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2 \\
&\quad + \frac{\rho}{5}\|Ax_{t+1}^s + \sum_{j=1}^k B_j y_j^{s,t+1} - c\|^2 \\
&\ge f(x_{t+1}^s) + \sum_{j=1}^k\psi_j(y_j^{s,t+1}) - \frac{5\kappa_A L^2 d}{\sigma_{\min}^A\rho b}\|x_t^s - \tilde{x}^s\|_2^2 - \frac{5\kappa_A L^2 d^2\mu^2}{4\sigma_{\min}^A\rho} - \frac{5\kappa_A\delta^2}{2\sigma_{\min}^A\rho} - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1}^s - x_t^s\|^2
\end{aligned} \tag{39}
$$

where the first inequality is obtained by applying $\langle a, b\rangle \le \frac{1}{2\beta}\|a\|^2 + \frac{\beta}{2}\|b\|^2$ to the terms $\langle(A^T)^+(\hat{\nabla}f(x_t) - \nabla f(x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$, $\langle(A^T)^+\nabla f(x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ and $\langle(A^T)^+\frac{G}{\eta}(x_{t+1} - x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ with $\beta = \frac{\rho}{5}$, respectively; the second inequality follows by Lemma 1 of [Huang *et al.*, 2019] and Assumption 2. Using the definition

of function R_t^s and Assumption 3, we have

$$R_{t+1}^s \geq f^* + \sum_{j=1}^{k} \psi_j^* - \frac{5\kappa_A L^2 d^2 \mu^2}{4\sigma_{\min}^A \rho} - \frac{5\kappa_A \delta^2}{2\sigma_{\min}^A \rho}, \text{ for } t = 0, 1, 2, \cdots \tag{40}$$

Thus, the function R_t^s is bounded from below. Let R^* denotes a lower bound of R_t^s.

Finally, telescoping (29) and (37) over t from 0 to $m-1$ and over s from 1 to S, we have

$$\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}(\sigma_{\min}^H \sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|_2^2 + \chi_t\|x_{t+1}^s - x_t^s\|^2) \leq \frac{R_0^1 - R^*}{T} + \frac{9L^2 d^2 \mu^2}{\sigma_{\min}^A \rho} + \frac{Ld^2\mu^2}{4}, \tag{41}$$

where $T = mS$.

□

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

$$\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big),\ \nu_2 = 6L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}$$
$$\nu_3 = \frac{18L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2},\ \nu_4 = \frac{L}{4} + \frac{9L^2}{\sigma_{\min}^A}.$$

Theorem 3. *Suppose the sequence $\{(x_t^s, y_{[k]}^{s,t}, \lambda_t^s)_{t=1}^m\}_{s=1}^S$ is generated from Algorithm 1. Let $m = [n^{\frac{1}{3}}]$, $b = [d^{1-l}n^{\frac{2}{3}}]$, $l \in \{0, \frac{1}{2}, 1\}$, $\eta = \frac{\alpha\sigma_{\min}(G)}{9d^lL}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G d^l L}{\sigma_{\min}^A \alpha}$, then we have*

$$\min_{s,t}\mathbb{E}\big[dist(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \leq O(\frac{d^{2l}}{T}) + O(d^{2+2l}\mu^2),$$

where $\gamma = \min(\sigma_{\min}^H, \chi_t, L)$ with $\chi_t \geq \frac{3\sqrt{71\kappa_A}\kappa_G d^l L}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ and R^* is a lower bound of function R_t^s. It follows that suppose the smoothing parameter μ and the whole number of iteration $T = mS$ satisfy

$$\frac{1}{\mu^2} \geq \frac{2d^{2+2l}}{\epsilon}\max\big\{\nu_1\nu_4 + \frac{3L^2}{2}, \nu_2\nu_4 + \frac{9L^2}{\sigma_{\min}^A\rho^2}, \nu_3\nu_4\big\},$$
$$T = \frac{4\nu_{\max}(R_0^1 - R^*)}{\epsilon\gamma},$$

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

Proof. First, we define a variable $\theta_t^s = \|x_{t+1}^s - x_t^s\|^2 + \|x_t^s - x_{t-1}^s\|^2 + \frac{d}{b}(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2) + \sum_{j=1}^k\|y_j^{s,t} - y_j^{s,t+1}\|^2$. By the step 8 of Algorithm 1, we have, for all $i \in [k]$

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \partial_{y_j}L(x, y_{[k]}, \lambda))^2\big]_{s,t+1} &= \mathbb{E}\big[\text{dist}(0, \partial\psi_j(y_j^{s,t+1}) - B_j^T\lambda_{t+1}^s)^2\big] \\
&= \|B_j^T\lambda_t^s - \rho B_j^T(Ax_t^s + \sum_{i=1}^{j}B_i y_i^{s,t+1} + \sum_{i=j+1}^{k}B_i y_i^{s,t} - c) - H_j(y_j^{s,t+1} - y_j^{s,t}) - B_j^T\lambda_{t+1}^s\|^2 \\
&= \|\rho B_j^T A(x_{t+1}^s - x_t^s) + \rho B_j^T\sum_{i=j+1}^{k}B_i(y_i^{s,t+1} - y_i^{s,t}) - H_j(y_j^{s,t+1} - y_j^{s,t})\|^2 \\
&\leq k\rho^2\sigma_{\max}^{B_j}\sigma_{\max}^A\|x_{t+1}^s - x_t^s\|^2 + k\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^{k}\sigma_{\max}^{B_i}\|y_i^{s,t+1} - y_i^{s,t}\|^2 \\
&\quad + k\sigma_{\max}^2(H_j)\|y_j^{s,t+1} - y_j^{s,t}\|^2 \\
&\leq k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big)\theta_t^s,
\end{aligned} \tag{42}$$

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 9 of Algorithm 1, we have

$$\begin{aligned}\mathbb{E}[\text{dist}(0,\nabla_x L(x,y_{[k]},\lambda))]_{s,t+1} &= \mathbb{E}\|A^T\lambda_{t+1}^s - \nabla f(x_{t+1}^s)\|^2\\ &= \mathbb{E}\|\hat{g}_t^s - \nabla f(x_{t+1}^s) - \frac{G}{\eta}(x_t^s - x_{t+1}^s)\|^2\\ &= \mathbb{E}\|\hat{g}_t^s - \nabla f(x_t^s) + \nabla f(x_t^s) - \nabla f(x_{t+1}^s) - \frac{G}{\eta}(x_t^s - x_{t+1}^s)\|^2\\ &\le \frac{6L^2d}{b}\|x_t^s - \tilde{x}^s\|^2 + 3(L^2 + \frac{\sigma_{\max}^2(G)}{\eta^2})\|x_t^s - x_{t+1}^s\|^2 + \frac{3L^2d^2\mu^2}{2}\\ &\le \big(6L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}\big)\theta_t^s + \frac{3L^2d^2\mu^2}{2}. \end{aligned} \tag{43}$$

By the step 10 of Algorithm 1, we have

$$\begin{aligned}\mathbb{E}[\text{dist}(0,\nabla_\lambda L(x,y_{[k]},\lambda))]_{s,t+1} &= \mathbb{E}\|Ax_{t+1}^s + By_{t+1}^s - c\|^2\\ &= \frac{1}{\rho^2}\mathbb{E}\|\lambda_{t+1}^s - \lambda_t^s\|^2\\ &\le \frac{18L^2d}{\sigma_{\min}^A\rho^2 b}\big(\|x_t^s - \tilde{x}^s\|^2 + \|x_{t-1}^s - \tilde{x}^s\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\|x_{t+1}^s - x_t^s\|^2\\ &\quad + \frac{3(\sigma_{\max}^2(G) + 3L^2\eta^2)}{\sigma_{\min}^A\eta^2\rho^2}\|x_t^s - x_{t-1}^s\|^2 + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho^2}\\ &\le \big(\frac{18L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\big)\theta_t^s + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho^2}. \end{aligned} \tag{44}$$

Next, using (41), we have

$$\frac{1}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}(\sigma_{\min}^H\sum_{j=1}^{k}\|y_j^{s,t} - y_j^{s,t+1}\|^2 + \frac{Ld}{b}\|x_t^s - \tilde{x}^s\|^2 + \chi_t\|x_{t+1}^s - x_t^s\|^2) \le \frac{R_0^1 - R^*}{T} + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4}, \tag{45}$$

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

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

$$\begin{aligned}c_{t+1} = (\frac{36\kappa_A L^2 d}{\sigma_{\min}^A\rho b} + \frac{2Ld}{b})\frac{(1+\beta)^{m-t} - 1}{\beta} &= \frac{md}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L)\big((1+\frac{1}{m})^{m-t} - 1\big)\\ &\le \frac{md}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L)(e-1) \le \frac{2md}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L). \end{aligned} \tag{46}$$

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

$$\begin{aligned}\chi_t &\ge \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+1/\beta)\frac{2md}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L)\\ &= \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+m)\frac{2md}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L)\\ &\ge \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{4m^2d}{b}(\frac{36\kappa_A L^2}{\sigma_{\min}^A\rho} + 2L)\\ &= \underbrace{\frac{\sigma_{\min}(G)}{\eta} - L - \frac{8m^2dL}{b}}_{Q_1} + \underbrace{\frac{\rho\sigma_{\min}^A}{2} - \frac{6\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{144m^2d\kappa_A L^2}{b\sigma_{\min}^A\rho}}_{Q_2}. \end{aligned} \tag{47}$$

When $1 \leq d < n^{\frac{1}{3}}$, let $m = [n^{\frac{1}{3}}]$, $b = [dn^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{9L}$, we have $Q_1 \geq 0$. Further, let $\eta = \frac{\alpha\sigma_{\min}(G)}{9L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G L}{\sigma^A_{\min}\alpha}$, we have

$$\begin{aligned} Q_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{486\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{144\kappa_A L^2}{\sigma^A_{\min}\rho} \\ &\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{639\kappa_A\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} \\ &= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{639\kappa_A\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\ &\geq \frac{3\sqrt{71\kappa_A}\kappa_G L}{2\alpha}, \end{aligned} \tag{48}$$

where $\kappa_G = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)} \geq 1$, and the third inequality follows $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G L}{\sigma^A_{\min}\alpha}$. Thus, we have $\chi_t \geq \frac{3\sqrt{71\kappa_A}\kappa_G L}{2\alpha} > 0$ for all $t \in \{1, 2, \cdots, m\}$.

By (45), we have

$$\begin{aligned} \min_{s,t} \mathbb{E}\big[\text{dist}(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] &\leq \frac{\nu_{\max}}{T}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\theta_t^s + \max\{\frac{3L^2d^2\mu^2}{2}, \frac{9L^2d^2\mu^2}{\sigma^A_{\min}\rho^2}\} \\ &\leq \frac{2\nu_{\max}(R_0^1 - R^*)}{\gamma T} + \frac{9\nu_{\max}L^2d^2\mu^2}{\gamma\sigma^A_{\min}\rho} + \frac{\nu_{\max}Ld^2\mu^2}{4\gamma} + \max\{\frac{3L^2d^2\mu^2}{2}, \frac{9L^2d^2\mu^2}{\sigma^A_{\min}\rho^2}\} \end{aligned} \tag{49}$$

where $\gamma = \min(\sigma^H_{\min}, \chi_t, L)$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$,

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

Given $\eta = \frac{\alpha\sigma_{\min}(G)}{9L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G L}{\sigma^A_{\min}\alpha}$, it is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(1)$, which are independent on n and d. Thus, we obtain

$$\min_{s,t} \mathbb{E}\big[\text{dist}(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \leq O(\frac{1}{T}) + O(d^2\mu^2). \tag{50}$$

When $n^{\frac{1}{3}} \leq d < n^{\frac{2}{3}}$, let $m = [n^{\frac{1}{3}}]$, $b = [d^{\frac{1}{2}}n^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{9\sqrt{dL}}$, we have $Q_1 \geq 0$. Further, let $\eta = \frac{\alpha\sigma_{\min}(G)}{9\sqrt{dL}}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{71d\kappa_A}\kappa_G L}{\sigma^A_{\min}\alpha}$, we have

$$\begin{aligned} Q_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{486d\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{144\kappa_A L^2}{\sigma^A_{\min}\rho} \\ &\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{639d\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2} \\ &= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{639d\kappa_A\kappa_G^2 L^2}{\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\ &\geq \frac{3\sqrt{71d\kappa_A}\kappa_G L}{2\alpha}, \end{aligned} \tag{51}$$

where the second equality follows by $\rho = \frac{6\sqrt{71d\kappa_A}\kappa_G L}{\sigma^A_{\min}\alpha}$. Thus, we have $\chi_t \geq \frac{3\sqrt{71d\kappa_A}\kappa_G L}{2\alpha} > 0$. Similarly, it is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(d)$. Thus, we obtain

$$\min_{s,t} \mathbb{E}\big[\text{dist}(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \leq O(\frac{d}{T}) + O(d^3\mu^2). \tag{52}$$

When $n^{\frac{2}{3}} \leq d$, let $m = [n^{\frac{1}{3}}]$, $b = [n^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{9dL}$, we have $Q_1 \geq 0$. Further, let $\eta = \frac{\alpha\sigma_{\min}(G)}{9dL}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G dL}{\sigma^A_{\min}\alpha}$, we have

$$\begin{aligned} Q_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{486d^2\kappa_G^2L^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{144\kappa_A L^2}{\sigma^A_{\min}\rho} \\ &\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{639d^2\kappa_G^2L^2}{\sigma^A_{\min}\rho\alpha^2} \\ &= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{639d^2\kappa_A\kappa_G^2L^2}{\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\ &\geq \frac{3\sqrt{71\kappa_A}\kappa_G dL}{2\alpha}, \end{aligned} \tag{53}$$

where the second equality follows by $\rho = \frac{6\sqrt{71\kappa_A}\kappa_G dL}{\sigma^A_{\min}\alpha}$. Thus, we have $\chi_t \geq \frac{3\sqrt{71\kappa_A}\kappa_G dL}{2\alpha} > 0$. Similarly, it is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(d^2)$. Thus, we obtain

$$\min_{s,t} \mathbb{E}\big[\text{dist}(0, \partial L(x_t^s, y_{[k]}^{s,t}, \lambda_t^s))^2\big] \leq O(\frac{d^2}{T}) + O(d^4\mu^2). \tag{54}$$

□

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

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

Lemma 5. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated by Algorithm 2. The following inequality holds*

$$\begin{aligned} \mathbb{E}\|\lambda_{t+1} - \lambda_t\|^2 \leq & \frac{18L^2d}{\sigma^A_{\min}b}\frac{1}{n}\sum_{i=1}^n \big(\|x_t - z_i^t\|^2 + \|x_{t-1} - z_i^{t-1}\|^2\big) + \frac{3\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2}\|x_{t+1} - x_t\|^2 \\ & + \frac{3(\sigma^2_{\max}(G) + 3L^2\eta^2)}{\sigma^A_{\min}\eta^2}\|x_t - x_{t-1}\|^2 + \frac{9L^2d^2\mu^2}{\sigma^A_{\min}}. \end{aligned} \tag{55}$$

Proof. By the optimize condition of the the step 7 in Algorithm 2, we have

$$\hat{g}_t + \frac{1}{\eta}G(x_{t+1} - x_t) - A^T\lambda_t + \rho A^T(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c) = 0. \tag{56}$$

Using the step 8 of Algorithm 2, then we have

$$A^T\lambda_{t+1} = \hat{g}_t + \frac{G}{\eta}(x_{t+1} - x_t). \tag{57}$$

It follows that

$$A^T(\lambda_{t+1} - \lambda_t) = \hat{g}_t - \hat{g}_{t-1} + \frac{G}{\eta}(x_{t+1} - x_t) - \frac{1}{\eta}G(x_t - x_{t-1}). \tag{58}$$

By Assumption 4, we have

$$\|\lambda_{t+1} - \lambda_t\|^2 \leq \frac{1}{\sigma^A_{\min}}\big[3\|\hat{g}_t - \hat{g}_{t-1}\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\|x_{t+1} - x_t\|^2 + \frac{3\sigma^2_{\max}(G)}{\eta^2}\|x_t - x_{t-1}\|^2\big]. \tag{59}$$

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

$$\begin{aligned} \|\hat{g}_t - \hat{g}_{t-1}\|^2 &= \|\hat{g}_t - \nabla f(x_t) + \nabla f(x_t) - \nabla f(x_{t-1}) + \nabla f(x_{t-1}) - \hat{g}_{t-1}\|^2 \\ &\leq 3\|\hat{g}_t - \nabla f(x_t)\|^2 + 3\|\nabla f(x_t) - \nabla f(x_{t-1})\|^2 + 3\|\nabla f(x_{t-1}) - \hat{g}_{t-1}\|^2 \\ &\leq \frac{6L^2d}{b}\frac{1}{n}\sum_{i=1}^n \big(\|x_t - z_i^t\|^2 + \|x_{t-1} - x_i^{t-1}\|^2\big) + 3L^2d\mu^2 + 3\|\nabla f(x_t) - \nabla f(x_{t-1})\|^2 \\ &\leq \frac{6L^2d}{b}\frac{1}{n}\sum_{i=1}^n \big(\|x_t - z_i^t\|^2 + \|x_{t-1} - x_i^{t-1}\|^2\big) + 3L^2\|x_t - x_{t-1}\|^2 + 3L^2d^2\mu^2, \end{aligned} \tag{60}$$

where the second inequality holds by lemma 3 of [Huang *et al.*, 2019], and the third inequality holds by Assumption 1. Finally, combining the inequalities (59) and (60), we can obtain the above result. □

Lemma 6. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2, and define a* Lyapunov *function*

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\kappa_A \sigma_{\max}^2(G)}{\sigma_{\min}^A \rho \eta^2} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A \rho b}\frac{1}{n}\sum_{i=1}^n \|x_{t-1} - z_i^{t-1}\|^2 + c_t \frac{1}{n}\sum_{i=1}^n \|x_t - z_i^t\|^2\big],$$

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

$$c_t = \begin{cases} \dfrac{36\kappa_A L^2 d}{\sigma_{\min}^A \rho b} + \dfrac{2Ld}{b} + (1-p)(1+\beta)c_{t+1},\ 0 \le t \le T-1, \\ 0,\ t \ge T. \end{cases}$$

It follows that

$$\frac{1}{T}\sum_{t=1}^T (\chi_t \|x_t - x_{t+1}\|^2 + \frac{Ld}{b}\frac{1}{n}\sum_{i=1}^n \|x_t - z_i^t\|^2 + \sigma_{\min}^H \sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2) \le \frac{\Omega_0 - \Omega^*}{T} + \frac{9L^2 d^2 \mu^2}{\sigma_{\min}^A \rho} + \frac{Ld^2\mu^2}{4}, \tag{61}$$

where $\chi_t = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho \sigma_{\min}^A}{2} - L - \frac{6\kappa_A \sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho} - \frac{9L^2}{\sigma_{\min}^A \rho} - (1 + \frac{1-p}{\beta})c_{t+1}$ and Ω^* denotes a lower bound of Ω_t.

Proof. By the optimal condition of step 6 in Algorithm 2, we have, for $j \in [k]$

$$\begin{aligned}
0 &= (y_j^t - y_j^{t+1})^T \big(\partial \psi_j(y_j^{t+1}) - B^T \lambda_t + \rho B^T (Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c) + H_j(y_j^{t+1} - y_j^t)\big) \\
&\le \psi_j(y_j^t) - \psi_j(y_j^{t+1}) - (\lambda_t)^T(B_j y_j^t - B_j y_j^{t+1}) + \rho(By_j^t - By_j^{t+1})^T(Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c) \\
&\quad - \|y_j^{t+1} - y_j^t\|_{H_j}^2 \\
&= \psi_j(y_j^t) - \psi_j(y_j^{t+1}) - (\lambda_t)^T(Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^k B_i y_i^t - c) + (\lambda_t)^T(Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c) \\
&\quad + \frac{\rho}{2}\|Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^k B_i y_i^t - c\|^2 - \frac{\rho}{2}\|Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c\|^2 - \|y_j^{t+1} - y_j^t\|_{H_j}^2 \\
&\quad - \frac{\rho}{2}\|B_j y_j^t - B_j y_j^{t+1}\|^2 \\
&= \underbrace{f(x_t) + \sum_{l=1}^j \psi_l(y_l^{t+1}) + \sum_{l=j+1}^k \psi_l(y_l^t) - (\lambda_t)^T(Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^k B_i y_i^t - c) + \frac{\rho}{2}\|Ax_t + \sum_{i=1}^{j-1} B_i y_i^{t+1} + \sum_{i=j}^k B_i y_i^t - c\|^2}_{\mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:k]}^t, \lambda_t)} \\
&\quad - \underbrace{\big(f(x_t) + \sum_{l=1}^{j-1} \psi_l(y_l^{t+1}) + \sum_{l=j}^k \psi_l(y_l^t) - (\lambda_t)^T(Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c) + \frac{\rho}{2}\|Ax_t + \sum_{i=1}^j B_i y_i^{t+1} + \sum_{i=j+1}^k B_i y_i^t - c\|^2\big)}_{\mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:k]}^t, \lambda_t)} \\
&\quad - \|y_j^{t+1} - y_j^t\|_{H_j}^2 - \frac{\rho}{2}\|B_j y_j^t - B_j y_j^{t+1}\|^2 \\
&\le \mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:k]}^t, \lambda_t) - \mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:k]}^t, \lambda_t) - \sigma_{\min}(H_j)\|y_j^t - y_j^{t+1}\|^2,
\end{aligned} \tag{62}$$

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

$$\mathcal{L}_\rho(x_t, y_{[j-1]}^{t+1}, y_{[j:k]}^t, \lambda_t) \le \mathcal{L}_\rho(x_t, y_{[j]}^{t+1}, y_{[j+1:k]}^t, \lambda_t) - \sigma_{\min}(H_j)\|y_j^t - y_j^{t+1}\|^2. \tag{63}$$

Telescoping inequality (63) over j from 1 to k, we obtain

$$\mathcal{L}_\rho(x_t, y_{[k]}^{t+1}, \lambda_t) \le \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) - \sigma_{\min}^H \sum_{j=1}^k \|y_j^t - y_j^{t+1}\|^2, \tag{64}$$

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

By Assumption 1, we have

$$0\leq f(x_t)-f(x_{t+1})+\nabla f(x_t)^T(x_{t+1}-x_t)+\frac{L}{2}\|x_{t+1}-x_t\|^2. \tag{65}$$

Using the step 7 of Algorithm 2, we have

$$0=(x_t-x_{t+1})^T\big(\hat{g}_t-A^T\lambda_t+\rho A^T(Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c)+\frac{G}{\eta}(x_{t+1}-x_t)\big). \tag{66}$$

Combining (65) and (66), 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(\hat{g}_t-A^T\lambda_t+\rho A^T(Ax_{t+1}+\sum_{j=1}^{k}B_jy_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(\hat{g}_t-\nabla f(x_t))\\
&-(\lambda_t)^T(Ax_t-Ax_{t+1})+\rho(Ax_t-Ax_{t+1})^T(Ax_t+\sum_{j=1}^{k}B_jy_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(\hat{g}_t-\nabla f(x_t))-(\lambda_t)^T(Ax_t+\sum_{j=1}^{k}B_jy_j^{t+1}-c)\\
&+(\lambda_t)^T(Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c)+\frac{\rho}{2}\big(\|Ax_t+\sum_{j=1}^{k}B_jy_j^{t+1}-c\|^2-\|Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c\|^2-\|Ax_t-Ax_{t+1}\|^2\big)\\
=\ & \underbrace{f(x_t)+\sum_{j=1}^{k}\psi(y_j^{t+1})-(\lambda_t)^T(Ax_t+\sum_{j=1}^{k}B_jy_j^{t+1}-c)+\frac{\rho}{2}\|Ax_t+\sum_{j=1}^{k}B_jy_j^{t+1}-c\|^2}_{\mathcal{L}_\rho(x_t,y_{[k]}^{t+1},\lambda_t)}\\
&-\underbrace{\big(f(x_{t+1})+\sum_{j=1}^{k}\psi(y_j^{t+1})-(\lambda_t)^T(Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c)+\frac{\rho}{2}\|Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c\|^2\big)}_{\mathcal{L}_\rho(x_{t+1},y_{[k]}^{t+1},\lambda_t)}\\
&+\frac{L}{2}\|x_t-x_{t+1}\|^2+(x_t-x_{t+1})^T(\hat{g}_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_{[k]}^{t+1},\lambda_t)-\mathcal{L}_\rho(x_{t+1},y_{[k]}^{t+1},\lambda_t)-(\frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^{A}}{2}-\frac{L}{2})\|x_t-x_{t+1}\|^2+(x_t-x_{t+1})^T(\hat{g}_t-\nabla f(x_t))\\
\overset{(ii)}{\leq}\ & \mathcal{L}_\rho(x_t,y_{[k]}^{t+1},\lambda_t)-\mathcal{L}_\rho(x_{t+1},y_{[k]}^{t+1},\lambda_t)-(\frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^{A}}{2}-L)\|x_t-x_{t+1}\|^2+\frac{1}{2L}\|\hat{g}_t-\nabla f(x_t)\|^2\\
\overset{(iii)}{\leq}\ & \mathcal{L}_\rho(x_t,y_{[k]}^{t+1},\lambda_t)-\mathcal{L}_\rho(x_{t+1},y_{[k]}^{t+1},\lambda_t)-(\frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^{A}}{2}-L)\|x_t-x_{t+1}\|^2+\frac{Ld}{b}\frac{1}{n}\sum_{i=1}^{n}\|x_t-z_i^t\|^2+\frac{Ld^2\mu^2}{4},
\end{aligned} \tag{67}$$

where the equality (i) holds by applying the equality $(a-b)^Tb=\frac{1}{2}(\|a\|^2-\|b\|^2-\|a-b\|^2)$ on the term $(Ax_t-Ax_{t+1})^T(Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c)$; the inequality (ii) follows by the inequality $a^Tb\leq\frac{L}{2}\|a\|^2+\frac{1}{2L}\|a\|^2$, and the inequality (iii) holds by lemma 3 of [Huang *et al.*, 2019]. Thus, we obtain

$$\begin{aligned}
\mathcal{L}_\rho(x_{t+1},y_{[k]}^{t+1},\lambda_t)\leq&\mathcal{L}_\rho(x_t,y_{[k]}^{t+1},\lambda_t)-(\frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^{A}}{2}-L)\|x_t-x_{t+1}\|^2\\
&+\frac{Ld}{b}\frac{1}{n}\sum_{i=1}^{n}\|x_t-z_i^t\|^2+\frac{Ld^2\mu^2}{4}.
\end{aligned} \tag{68}$$

By the step 8 in Algorithm 2, we have

$$\begin{aligned}\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) - \mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_t) &= \frac{1}{\rho}\|\lambda_{t+1} - \lambda_t\|^2 \\ &\leq \frac{18L^2 d}{\sigma_{\min}^A \rho b}\frac{1}{n}\sum_{i=1}^{n}\big(\|x_t - z_i^t\|^2 + \|x_{t-1} - z_i^{t-1}\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x_{t+1} - x_t\|^2 \\ &\quad + \frac{3(\sigma_{\max}^2(G) + 3L^2\eta^2)}{\sigma_{\min}^A \eta^2 \rho}\|x_t - x_{t-1}\|^2 + \frac{9L^2 d^2 \mu^2}{\sigma_{\min}^A \rho}. \end{aligned} \tag{69}$$

Combining (64), (68) and (69), we have

$$\begin{aligned}\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) \leq \mathcal{L}_\rho(x_t, y_{[k]}^{t}, \lambda_t) &- (\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L)\|x_t - x_{t+1}\|^2 - \sigma_{\min}^H \sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2 \\ &+ \frac{Ld}{b}\frac{1}{n}\sum_{i=1}^{n}\|x_t - z_i^t\|^2 + \frac{18L^2 d}{\sigma_{\min}^A \rho b}\frac{1}{n}\sum_{i=1}^{n}\big(\|x_t - z_i^t\|^2 + \|x_{t-1} - z_i^{t-1}\|^2\big) + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A \eta^2 \rho}\|x_{t+1} - x_t\|^2 \\ &+ \frac{3(\sigma_{\max}^2(G) + 3L^2\eta^2)}{\sigma_{\min}^A \eta^2 \rho}\|x_t - x_{t-1}\|^2 + \frac{9L^2 d^2 \mu^2}{\sigma_{\min}^A \rho} + \frac{Ld^2\mu^2}{4}. \end{aligned} \tag{70}$$

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

$$\Omega_t = \mathbb{E}\big[\mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\kappa_A \sigma_{\max}^2(G)}{\sigma_{\min}^A \rho \eta^2} + \frac{9L^2}{\sigma_{\min}^A \rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A \rho b}\frac{1}{n}\sum_{i=1}^{n}\|x_{t-1} - z_i^{t-1}\|^2 + c_t\frac{1}{n}\sum_{i=1}^{n}\|x_t - z_i^t\|^2\big].$$

By the step 9 of Algorithm 2, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|x_{t+1} - z_i^{t+1}\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\big(p\|x_{t+1} - x_t\|^2 + (1-p)\|x_{t+1} - z_i^t\|^2\big) \\ &= \frac{p}{n}\sum_{i=1}^{n}\|x_{t+1} - x_t\|^2 + \frac{1-p}{n}\sum_{i=1}^{n}\|x_{t+1} - z_i^t\|^2 \\ &= p\|x_{t+1} - x_t\|^2 + \frac{1-p}{n}\sum_{i=1}^{n}\|x_{t+1} - z_i^t\|^2, \end{aligned} \tag{71}$$

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

$$p = 1 - (1 - \frac{1}{n})^b \geq 1 - \frac{1}{1 + b/n} = \frac{b/n}{1 + b/n} \geq \frac{b}{2n}, \tag{72}$$

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

$$\begin{aligned}\|x_{t+1} - z_i^t\|^2 &= \|x_{t+1} - x_t + x_t - z_i^t\|^2 \\ &= \|x_{t+1} - x_t\|^2 + 2(x_{t+1} - x_t)^T(x_t - z_i^t) + \|x_t - z_i^t\|^2 \\ &\leq \|x_{t+1} - x_t\|^2 + 2\big(\frac{1}{2\beta}\|x_{t+1} - x_t\|^2 + \frac{\beta}{2}\|x_t - z_i^t\|^2\big) + \|x_t - z_i^t\|^2 \\ &= (1 + \frac{1}{\beta})\|x_{t+1} - x_t\|^2 + (1+\beta)\|x_t - z_i^t\|^2, \end{aligned} \tag{73}$$

where $\beta > 0$. Combining (71) with (73), we have

$$\frac{1}{n}\sum_{i=1}^{n}\|x_{t+1} - z_i^{t+1}\|^2 \leq (1 + \frac{1-p}{\beta})\|x_{t+1} - x_t\|^2 + \frac{(1-p)(1+\beta)}{n}\sum_{i=1}^{n}\|x_t - z_i^t\|^2. \tag{74}$$

It follows that

$$\begin{aligned}
\Omega_{t+1} &= \mathbb{E}\big[\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_{t+1}-x_t\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A b\rho}\frac{1}{n}\sum_{i=1}^n\|x_t - z_i^t\|^2 + c_{t+1}\frac{1}{n}\sum_{i=1}^n\|x_{t+1}-z_i^{t+1}\|^2\big] \\
&\le \mathcal{L}_\rho(x_t, y_{[k]}^t, \lambda_t) + (\frac{3\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\rho\eta^2} + \frac{9L^2}{\sigma_{\min}^A\rho})\|x_t - x_{t-1}\|^2 + \frac{18\kappa_A L^2 d}{\sigma_{\min}^A\rho b}\frac{1}{n}\sum_{i=1}^n\|x_{t-1}-z_i^{t-1}\|^2 \\
&\quad + \big(\frac{36\kappa_A L^2 d}{\sigma_{\min}^A\rho b} + \frac{2Ld}{b} + (1-p)(1+\beta)c_{t+1}\big)\frac{1}{n}\sum_{i=1}^n\|x_t - z_i^t\|^2 + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4} - \frac{Ld}{b}\frac{1}{n}\sum_{i=1}^n\|x_t-z_i^t\|^2 \\
&\quad - \big(\frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1+\frac{1-p}{\beta})c_{t+1}\big)\|x_t - x_{t+1}\|^2 - \sigma_{\min}^H\sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2 \\
&= \Omega_t - \chi_t\|x_t - x_{t+1}\|^2 - \frac{Ld}{b}\frac{1}{n}\sum_{i=1}^n\|x_t - z_i^t\|^2 - \sigma_{\min}^H\sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2 + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4}, \qquad (75)
\end{aligned}$$

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

By (57), we have

$$\lambda_{t+1} = (A^T)^+(\hat{\nabla}f(x_t) + \frac{G}{\eta}(x_{t+1}-x_t)), \qquad (76)$$

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

Then we have

$$\begin{aligned}
&\mathcal{L}_\rho(x_{t+1}, y_{[k]}^{t+1}, \lambda_{t+1}) = f(x_{t+1}) + \sum_{j=1}^k\psi_j(y_j^{t+1}) - \lambda_{t+1}^T(Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c) + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&= f(x_{t+1}) + \sum_{j=1}^k\psi_j(y_j^{t+1}) - \langle(A^T)^+(\hat{g}_t + \frac{G}{\eta}(x_{t+1}-x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&= f(x_{t+1}) + \sum_{j=1}^k\psi_j(y_j^{t+1}) - \langle(A^T)^+(\hat{g}_t - \nabla f(x_t) + \nabla f(x_t) + \frac{G}{\eta}(x_{t+1}-x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle \\
&\quad + \frac{\rho}{2}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&\ge f(x_{t+1}) + \sum_{j=1}^k\psi_j(y_j^{t+1}) - \frac{5\kappa_A}{2\sigma_{\min}^A\rho}\|\hat{g}_t - \nabla f(x_t)\|^2 - \frac{5\kappa_A}{2\sigma_{\min}^A\rho}\|\nabla f(x_t)\|^2 - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1}-x_t\|^2 \\
&\quad + \frac{\rho}{5}\|Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\|^2 \\
&\ge f(x_{t+1}) + \sum_{j=1}^k\psi_j(y_j^{t+1}) - \frac{5\kappa_A L^2 d}{\sigma_{\min}^A\rho b}\frac{1}{n}\sum_{i=1}^n\|x_t - z_i^t\|_2^2 - \frac{5\kappa_A L^2 d^2\mu^2}{4\sigma_{\min}^A\rho} - \frac{5\kappa_A\delta^2}{2\sigma_{\min}^A\rho} - \frac{5\kappa_A\sigma_{\max}^2(G)}{2\sigma_{\min}^A\eta^2\rho}\|x_{t+1}-x_t\|^2 \qquad (77)
\end{aligned}$$

where the first inequality is obtained by applying $\langle a, b\rangle \le \frac{1}{2\beta}\|a\|^2 + \frac{\beta}{2}\|b\|^2$ to the terms $\langle(A^T)^+(\hat{\nabla}f(x_t) - \nabla f(x_t)), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$, $\langle(A^T)^+\nabla f(x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ and $\langle(A^T)^+\frac{G}{\eta}(x_{t+1}-x_t), Ax_{t+1} + \sum_{j=1}^k B_j y_j^{t+1} - c\rangle$ with $\beta = \frac{\rho}{5}$, respectively; the second inequality follows by Lemma 3 of [Huang *et al.*, 2019] and Assumption 2. Using the definition of Ω_t and Assumption 3, we have

$$\Omega_{t+1} \ge f^* + \sum_{j=1}^k\psi_j^* - \frac{5\kappa_A L^2 d^2\mu^2}{4\sigma_{\min}^A\rho} - \frac{5\kappa_A\delta^2}{2\sigma_{\min}^A\rho}, \text{ for } t = 0, 1, 2, \cdots \qquad (78)$$

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

Finally, telescoping inequality (75) over t from 0 to T, we have

$$\frac{1}{T}\sum_{t=1}^{T}(\chi_t\|x_t - x_{t+1}\|^2 + \frac{Ld}{b}\frac{1}{n}\sum_{i=1}^{n}\|x_t - z_i^t\|^2 + \sigma_{\min}^H\sum_{j=1}^{k}\|y_j^t - y_j^{t+1}\|^2) \leq \frac{\Omega_0 - \Omega^*}{T} + \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho} + \frac{Ld^2\mu^2}{4}, \tag{79}$$

where $\chi_t = \frac{\sigma_{\min}(G)}{\eta} + \frac{\rho\sigma_{\min}^A}{2} - L - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - (1 + \frac{1-p}{\beta})c_{t+1}$.

□

Next, based on the above lemmas, we give the convergence properties of the ZO-SAGA-ADMM. For notational simplicity, let

$$\nu_1 = k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big),\ \nu_2 = 6L^2 + \frac{3\sigma_{\max}^2(G)}{\eta^2}$$
$$\nu_3 = \frac{18L^2}{\sigma_{\min}^A\rho^2} + \frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2},\ \nu_4 = \frac{L}{4} + \frac{9L^2}{\sigma_{\min}^A}.$$

Theorem 4. *Suppose the sequence $\{x_t, y_{[k]}^t, \lambda_t\}_{t=1}^T$ is generated from Algorithm 2. Let $b = n^{\frac{2}{3}}d^{\frac{1-l}{3}}$, $l \in \{0, \frac{1}{2}, 1\}$, $\eta = \frac{\alpha\sigma_{\min}(G)}{33d^lL}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{791\kappa_A}\kappa_G d^lL}{\sigma_{\min}^A\alpha}$ then we have*

$$\min_{1\leq t\leq T}\mathbb{E}\big[dist(0, \partial L(x_t, y_{[k]}^t, \lambda_t))^2\big] \leq O(\frac{d^{2l}}{T}) + O(d^{2+2l}\mu^2),$$

where $\gamma = \min(\sigma_{\min}^H, \chi_t, L)$ with $\chi_t \geq \frac{3\sqrt{791\kappa_A}\kappa_G d^lL}{2\alpha}$, $\nu_{\max} = \max(\nu_1, \nu_2, \nu_3)$ and Ω^* is a lower bound of function Ω_t. It follows that suppose the parameters μ and T satisfy

$$\frac{1}{\mu^2} \geq \frac{2d^{2+2l}}{\epsilon}\max\big\{\nu_1\nu_4 + \frac{3L^2}{2}, \nu_2\nu_4 + \frac{9L^2}{\sigma_{\min}^A\rho^2}, \nu_3\nu_4\big\},$$
$$T = \frac{4\kappa_{\max}}{\epsilon\gamma}(\Omega_0 - \Omega^*),$$

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

Proof. We begin with defining an useful variable $\theta_t = \|x_{t+1} - x_t\|^2 + \|x_t - x_{t-1}\|^2 + \frac{d}{bn}\sum_{i=1}^n(\|x_t - z_i^t\|^2 + \|x_{t-1} - z_i^{t-1}\|^2) + \sum_{j=1}^k\|y_j^t - y_j^{t+1}\|^2)$. By the optimal condition of the step 6 in Algorithm 2, we have, for all $i \in [k]$

$$\begin{aligned}
\mathbb{E}\big[\text{dist}(0, \partial_{y_j}L(x, y_{[k]}, \lambda))^2\big]_{t+1} &= \mathbb{E}\big[\text{dist}(0, \partial\psi_j(y_j^{t+1}) - B_j^T\lambda_{t+1})^2\big] \\
&= \|B_j^T\lambda_t - \rho B_j^T(Ax_t + \sum_{i=1}^{j}B_iy_i^{t+1} + \sum_{i=j+1}^{k}B_iy_i^t - c) - H_j(y_j^{t+1} - y_j^t) - B_j^T\lambda_{t+1}\|^2 \\
&= \|\rho B_j^TA(x_{t+1} - x_t) + \rho B_j^T\sum_{i=j+1}^{k}B_i(y_i^{t+1} - y_i^t) - H_j(y_j^{t+1} - y_j^t)\|^2 \\
&\leq k\rho^2\sigma_{\max}^{B_j}\sigma_{\max}^A\|x_{t+1} - x_t\|^2 + k\rho^2\sigma_{\max}^{B_j}\sum_{i=j+1}^{k}\sigma_{\max}^{B_i}\|y_i^{t+1} - y_i^t\|^2 \\
&\quad + k\sigma_{\max}^2(H_j)\|y_j^{t+1} - y_j^t\|^2 \\
&\leq k\big(\rho^2\sigma_{\max}^B\sigma_{\max}^A + \rho^2(\sigma_{\max}^B)^2 + \sigma_{\max}^2(H)\big)\theta_t,
\end{aligned} \tag{80}$$

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 7 in Algorithm 2, we have

$$\begin{aligned}\mathbb{E}[\text{dist}(0,\nabla_x L(x,y_{[k]},\lambda))]_{t+1} &= \mathbb{E}\|A^T\lambda_{t+1}-\nabla f(x_{t+1})\|^2\\ &= \mathbb{E}\|\hat{g}_t-\nabla f(x_{t+1})-\frac{G}{\eta}(x_t-x_{t+1})\|^2\\ &= \mathbb{E}\|\hat{g}_t-\nabla f(x_t)+\nabla f(x_t)-\nabla f(x_{t+1})-\frac{G}{\eta}(x_t-x_{t+1})\|^2\\ &\le \frac{6L^2d}{bn}\sum_{i=1}^{n}\|x_t-z_i^t\|^2+3(L^2+\frac{\sigma_{\max}^2(G)}{\eta^2})\|x_t-x_{t+1}\|^2+\frac{3L^2d^2\mu^2}{2}\\ &\le \big(6L^2+\frac{3\sigma_{\max}^2(G)}{\eta^2}\big)\theta_t+\frac{3L^2d^2\mu^2}{2},\end{aligned} \tag{81}$$

By the step 8 of Algorithm 2, we have

$$\begin{aligned}\mathbb{E}[\text{dist}(0,\nabla_\lambda L(x,y_{[k]},\lambda))]_{t+1} &= \mathbb{E}\|Ax_{t+1}+\sum_{j=1}^{k}B_jy_j^{t+1}-c\|^2\\ &= \frac{1}{\rho^2}\mathbb{E}\|\lambda_{t+1}-\lambda_t\|^2\\ &\le \frac{18L^2d}{\sigma_{\min}^A\rho^2bn}\sum_{i=1}^{n}\big(\|x_t-z_i^t\|^2+\|x_{t-1}-z_i^{t-1}\|^2\big)+\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\|x_{t+1}-x_t\|^2\\ &\quad+(\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}+\frac{9L^2)}{\sigma_{\min}^A\rho^2})\|x_t-x_{t-1}\|^2+\frac{9L^2d\mu^2}{\sigma_{\min}^A\rho^2}\\ &\le \big(\frac{18L^2}{\sigma_{\min}^A\rho^2}+\frac{3\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho^2}\big)\theta_t+\frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho^2}.\end{aligned} \tag{82}$$

Let $c_T=0$ and $\beta=\frac{b}{4n}$. Since $(1-p)(1+\beta)=1+\beta-p-p\beta\le 1+\beta-p$ and $p\ge\frac{b}{2n}$, it follows that

$$c_t\le c_{t+1}(1-\theta)+\frac{36\kappa_AL^2d}{\sigma_{\min}^Ab\rho}+\frac{2Ld}{b}, \tag{83}$$

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

$$c_t\le\frac{2d}{b}(\frac{18\kappa_AL^2}{\sigma_{\min}^A\rho}+L)\frac{1-\theta^{T-t}}{\theta}\le\frac{2d}{b\theta}(\frac{18\kappa_AL^2}{\sigma_{\min}^A\rho}+L)\le\frac{8nd}{b^2}(\frac{18\kappa_AL^2}{\sigma_{\min}^A\rho}+L). \tag{84}$$

It follows that

$$\begin{aligned}\chi_t &= \frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L-\frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}-\frac{9L^2}{\sigma_{\min}^A\rho}-(1+\frac{1-p}{\beta})c_{t+1}\\ &\ge \frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L-\frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}-\frac{9L^2}{\sigma_{\min}^A\rho}-(1+\frac{4n-2b}{b})\frac{8nd}{b^2}(\frac{18\kappa_AL^2}{\sigma_{\min}^A\rho}+L)\\ &= \frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L-\frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}-\frac{9L^2}{\sigma_{\min}^A\rho}-(\frac{4n}{b}-1)\frac{8nd}{b^2}(\frac{18\kappa_AL^2}{\sigma_{\min}^A\rho}+L)\\ &\ge \frac{\sigma_{\min}(G)}{\eta}+\frac{\rho\sigma_{\min}^A}{2}-L-\frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}-\frac{9L^2}{\sigma_{\min}^A\rho}-\frac{32n^2d}{b^3}(\frac{18\kappa_AL^2}{\sigma_{\min}^A\rho}+L)\\ &= \underbrace{\frac{\sigma_{\min}(G)}{\eta}-L-\frac{32n^2dL}{b^3}}_{Q_1}+\underbrace{\frac{\rho\sigma_{\min}^A}{2}-\frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho}-\frac{9L^2}{\sigma_{\min}^A\rho}-\frac{576n^2d\kappa_AL^2}{\sigma_{\min}^A\rho b^3}}_{Q_2}\end{aligned} \tag{85}$$

When $1\le d<n$, let $b=[d^{\frac{1}{3}}n^{\frac{2}{3}}]$ and $0<\eta\le\frac{\sigma_{\min}(G)}{33L}$, we have $Q_1\ge 0$. Further, let $\eta=\frac{\alpha\sigma_{\min}(G)}{33L}$ $(0<\alpha\le 1)$ and

$\rho = \frac{6\sqrt{791\kappa_A}\kappa_G L}{\sigma_{\min}^A \alpha}$, we have

$$\begin{aligned} Q_2 &= \frac{\rho\sigma_{\min}^A}{2} - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{576n^2d\kappa_A L^2}{\sigma_{\min}^A\rho b^3} \\ &= \frac{\rho\sigma_{\min}^A}{2} - \frac{6534\kappa_A\kappa_G^2 L^2}{\sigma_{\min}^A\rho\alpha^2} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{576\kappa_A L^2}{\sigma_{\min}^A\rho} \\ &\geq \frac{\rho\sigma_{\min}^A}{2} - \frac{7119\kappa_A\kappa_G^2 L^2}{\sigma_{\min}^A\rho\alpha^2} \\ &= \frac{\rho\sigma_{\min}^A}{4} + \underbrace{\frac{\rho\sigma_{\min}^A}{4} - \frac{7119\kappa_A\kappa_G^2 L^2}{\sigma_{\min}^A\rho\alpha^2}}_{\geq 0} \\ &\geq \frac{3\sqrt{791\kappa_A}\kappa_G L}{2\alpha}. \end{aligned} \tag{86}$$

Thus, we have $\chi_t \geq \frac{3\sqrt{791\kappa_A}\kappa_G L}{2\alpha}$.

By (79), we have

$$\begin{aligned} \min_{1\leq t\leq T} \mathbb{E}\big[\text{dist}(0, \partial L(x_t, y_{[k]}^t, \lambda_t))^2\big] &\leq \frac{\nu_{\max}}{T}\sum_{t=1}^{T}\theta_t + \max\{\frac{3L^2d^2\mu^2}{2}, \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho^2}\} \\ &\leq \frac{2\nu_{\max}(\Omega_0 - \Omega^*)}{\gamma T} + \frac{9\nu_{\max}L^2d^2\mu^2}{\gamma\sigma_{\min}^A\rho} + \frac{\nu_{\max}Ld^2\mu^2}{4\gamma} + \max\{\frac{3L^2d^2\mu^2}{2}, \frac{9L^2d^2\mu^2}{\sigma_{\min}^A\rho^2}\} \end{aligned} \tag{87}$$

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

Given $\eta = \frac{\alpha\sigma_{\min}(G)}{33L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{791\kappa_A}\kappa_G L}{\sigma_{\min}^A\alpha}$, since k is relatively small, it is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(1)$, which are independent on n and d. Thus, we obtain

$$\min_{1\leq t\leq T} \mathbb{E}\big[\text{dist}(0, \partial L(x_t, y_{[k]}^t, \lambda_t))^2\big] \leq O(\frac{1}{T}) + O(d^2\mu^2). \tag{88}$$

When $n \leq d < 2n$, let $b = [d^{\frac{1}{6}}n^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{33\sqrt{d}L}$, we have $Q_1 \geq 0$. Further, let $\eta = \frac{\alpha\sigma_{\min}(G)}{33\sqrt{d}L}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{791d\kappa_A}\kappa_G L}{\sigma_{\min}^A\alpha}$, we have

$$\begin{aligned} Q_2 &= \frac{\rho\sigma_{\min}^A}{2} - \frac{6\kappa_A\sigma_{\max}^2(G)}{\sigma_{\min}^A\eta^2\rho} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{576n^2d\kappa_A L^2}{\sigma_{\min}^A\rho b^3} \\ &= \frac{\rho\sigma_{\min}^A}{2} - \frac{6534\kappa_A\kappa_G^2 L^2 d}{\sigma_{\min}^A\rho\alpha^2} - \frac{9L^2}{\sigma_{\min}^A\rho} - \frac{576\kappa_A L^2\sqrt{d}}{\sigma_{\min}^A\rho} \\ &\geq \frac{\rho\sigma_{\min}^A}{2} - \frac{7119\kappa_A\kappa_G^2 L^2 d}{\sigma_{\min}^A\rho\alpha^2} \\ &= \frac{\rho\sigma_{\min}^A}{4} + \underbrace{\frac{\rho\sigma_{\min}^A}{4} - \frac{7119\kappa_A\kappa_G^2 L^2 d}{\sigma_{\min}^A\rho\alpha^2}}_{\geq 0} \\ &\geq \frac{3\sqrt{791d\kappa_A}\kappa_G}{2\alpha}. \end{aligned} \tag{89}$$

Thus, we have $\chi_t \geq \frac{3\sqrt{791d\kappa_A}\kappa_G L}{2\alpha}$. It is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(d)$. Thus, we obtain

$$\min_{1\leq t\leq T} \mathbb{E}\big[\text{dist}(0, \partial L(x_t, y_{[k]}^t, \lambda_t))^2\big] \leq O(\frac{d}{T}) + O(d^3\mu^2). \tag{90}$$

When $2n \leq d$, let $b = [n^{\frac{2}{3}}]$ and $0 < \eta \leq \frac{\sigma_{\min}(G)}{33dL}$, we have $Q_1 \geq 0$. Further, let $\eta = \frac{\alpha\sigma_{\min}(G)}{33dL}$ $(0 < \alpha \leq 1)$ and $\rho = \frac{6\sqrt{791\kappa_A}\kappa_G d}{\sigma^A_{\min}\alpha}$, we have

$$\begin{aligned}
Q_2 &= \frac{\rho\sigma^A_{\min}}{2} - \frac{6\kappa_A\sigma^2_{\max}(G)}{\sigma^A_{\min}\eta^2\rho} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{576n^2d\kappa_AL^2}{\sigma^A_{\min}\rho b^3} \\
&= \frac{\rho\sigma^A_{\min}}{2} - \frac{6534\kappa_A\kappa_G^2L^2d^2}{\sigma^A_{\min}\rho\alpha^2} - \frac{9L^2}{\sigma^A_{\min}\rho} - \frac{576\kappa_AL^2d}{\sigma^A_{\min}\rho} \\
&\geq \frac{\rho\sigma^A_{\min}}{2} - \frac{7119\kappa_A\kappa_G^2L^2d^2}{\sigma^A_{\min}\rho\alpha^2} \\
&= \frac{\rho\sigma^A_{\min}}{4} + \underbrace{\frac{\rho\sigma^A_{\min}}{4} - \frac{7119\kappa_A\kappa_G^2L^2d^2}{\sigma^A_{\min}\rho\alpha^2}}_{\geq 0} \\
&\geq \frac{3\sqrt{791\kappa_A}\kappa_G d}{2\alpha}.
\end{aligned} \tag{91}$$

Thus, we have $\chi_t \geq \frac{3\sqrt{791\kappa_A}\kappa_G d}{2\alpha}$. It is easy verifies that $\gamma = O(1)$ and $\nu_{\max} = O(d^2)$. Thus, we obtain

$$\min_{1\leq t\leq T}\mathbb{E}\big[\text{dist}(0,\partial L(x_t,y^t_{[k]},\lambda_t))^2\big] \leq O(\frac{d^2}{T}) + O(d^4\mu^2). \tag{92}$$

□