On the Information-Adaptive Variants of the ADMM: an Iteration Complexity Perspective

Xiang GAO * Bo JIANG † Shuzhong ZHANG ‡

November 7, 2014

Abstract

Designing algorithms for an optimization model often amounts to maintaining a balance between the degree of information to request from the model on the one hand, and the computational speed to expect on the other hand. Naturally, the more information is available, the faster one can expect the algorithm to converge. The popular algorithm of ADMM demands that objective function is easy to optimize once the coupled constraints are shifted to the objective with multipliers. However, in many applications this assumption does not hold; instead, only some noisy estimations of the gradient of the objective – or even only the objective itself – are available. This paper aims to bridge this gap. We present a suite of variants of the ADMM, where the trade-offs between the required information on the objective and the computational complexity are explicitly given. The new variants allow the method to be applicable on a much broader class of problems where only noisy estimations of the gradient or the function values are accessible, yet the flexibility is achieved without sacrificing the computational complexity bounds.

Keywords: alternating direction method of multipliers (ADMM), iteration complexity, stochastic approximation, first-order method, direct method.

Mathematics Subject Classification: 90C15, 90C25, 68Q25, 62L20

*Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: gaoxx460@umn.edu. Research of this author was supported in part by National Science Foundation (Grant CMMI-1161242).

†Research Center for Management Science and Data Analytics, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China. Email: isyebojiang@gmail.com. Research of this author was supported in part by National Natural Science Foundation of China (Grant 11401364).

‡Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: zhangs@umn.edu. Research of this author was supported in part by National Science Foundation (Grant CMMI-1161242).

1 Introduction

In this paper, we consider the following constrained convex optimization model

$$\begin{array}{ll} \min & f(x)+g(y) \\ \text{s.t.} & Ax+By=b, \\ & x\in\mathcal{X}, y\in\mathcal{Y} \end{array} \tag{1}$$

where $x\in\mathbf{R}^{n_x}$, $y\in\mathbf{R}^{n_y}$, $A\in\mathbf{R}^{m\times n_x}$, $B\in\mathbf{R}^{m\times n_y}$, $b\in\mathbf{R}^m$, and $\mathcal{X}\subseteq\mathbf{R}^{n_x}$, $\mathcal{Y}\subseteq\mathbf{R}^{n_y}$ are closed convex sets; f is a smooth convex function, and g is a convex function and possibly nonsmooth. We further assume that the gradient of f is Lipschitz continuous:

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|,\ \forall\, x,y\in\mathcal{X}, \tag{2}$$

where L is a Lipschitz constant.

An intensive recent research attention for solving problem (1) has been devoted to the so-called Alternating Direction Method of Multipliers (abbreviated as ADMM), which is known to be a form of the operator splitting method (cf. [6, 7] and the references therein). Large-scale optimization problems in the form of (1) can be found in many application domains including compressed sensing, imaging processing, and statistical learning. Due to the large-scale nature, it is often impossible to inquire the second order information (such as the Hessian of the objective function) or invoke any second order operations (such as inverting a full-scale matrix) in the solution process. In this context, the ADMM as a first order method is an attractive approach; see [2]. Specifically, a typical iteration of ADMM for solving (1) runs as follows:

$$\left\{\begin{array}{l} x^{k+1}=\arg\min_{x\in\mathcal{X}}\mathcal{L}_\gamma(x,y^k,\lambda^k) \\ y^{k+1}=\arg\min_{y\in\mathcal{Y}}\mathcal{L}_\gamma(x^{k+1},y,\lambda^k) \\ \lambda^{k+1}=\lambda^k-\gamma(Ax^{k+1}+By^{k+1}-b), \end{array}\right. \tag{3}$$

where $\mathcal{L}_\gamma(x,y,\lambda)$ is the augmented Lagrangian function for problem (1) defined as:

$$\mathcal{L}_\gamma(x,y,\lambda)=f(x)+g(y)-\lambda^\top(Ax+By-b)+\frac{\gamma}{2}\|Ax+By-b\|^2. \tag{4}$$

The convergence of the ADMM for (1) is actually a consequence of the convergence of the so-called Douglas-Rachford operator splitting method (see [15, 7]). However, the rate of convergence for ADMM was established only very recently: [18, 25] show that for problem (1) the ADMM converges at the rate of $O(1/N)$ where N is the number of total iterations. Furthermore, by imposing additional conditions on the objective function or constraints, the ADMM can be shown to converge linearly; see [5, 19, 1, 22]. The ADMM can be naturally extended to solve problems with more than 2 blocks of variables. In spite of its excellent performance in practice, [3] demonstrates that in general the ADMM may diverge even for 3 blocks of variables. However, by imposing various additional conditions or by modifying the original ADMM, [4, 17, 16, 21] show that an $O(1/N)$ convergence rate can still be achieved.

In this paper, we take a different stance towards the applicability of the ADMM, which is dependent on the available informational structure of the problem. Observe that to implement (3), it is necessary that $\arg\min_{x\in\mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k)$ and $\arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^{k+1}, y, \lambda^k)$ can be solved efficiently at each iteration. While this is indeed the case for some classes of the problems (e.g. the lasso problem), it may also fail for many other applications. This triggers a natural question: Given the informational structure of the objective functions in the minimization subroutines, can the multipliers' method be adapted accordingly? In this paper we shall propose some variants of the ADMM to account for this informational structure of the objective functions. To bring out the hierarchy regarding the available information of the functions in question, let us first introduce the following definition.

Definition 1.1 *We call a function $f(x)$ to be* easy to minimize *with respect to x (f is hence said to be* **MinE** *as an abbreviation) if the proximal mapping $\arg\min_x f(x) + \frac{1}{2}\|x - z\|_H^2$ can be computed easily for any fixed z and $H \succeq 0$.*

Some remarks are in order here. If both f and g are **MinE**, then the original ADMM is readily applicable. In case that $\mathcal{L}_\gamma(x, y, \lambda)$ is **MinE** in x but not in y, Ma and Zhang [24] recently proposed an extra-gradient ADMM (EGADM) and showed an $O(1/N)$ iteration bound; in [24], it was posed as an unsolved problem to determine the iteration complexity bound for the following procedure (known as the GADM):

$$\begin{cases} y^{k+1} = \arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^k, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2 \\ x^{k+1} = [x^k - \nabla_x \mathcal{L}_\gamma(x^k, y^{k+1}, \lambda^k)]_{\mathcal{X}} \\ \lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b), \end{cases} \tag{5}$$

where $[x]_{\mathcal{X}}$ denotes the projection of x onto $\mathcal{X}$. In this paper we prove that the GADM also has an iteration bound of $O(1/N)$.

In stochastic programming (SP), the objective function is often in the form of expectation. In this case, even requesting its full gradient information is impractical. In [29], a stochastic version of problem (1) is considered. Historically, Robbins and Monro [32] introduced the so-called stochastic approximation (SA) approach to tackle this problem. Polyak and Juditsky [30, 31] proposed an SA method in which larger step-sizes are adopted and the asymptotical optimal rate of convergence is achieved; cf. [8, 9, 34, 33] for more details. Recently, there has been a renewed interest in SA, in the context of computational complexity analysis for convex optimization [27], which has focussed primarily on bounding the number of iterations required by the SA-type algorithms to ensure the expectation of the objective to be ϵ away from optimality. For instance, Nemirovski et al. [26] proposed a mirror descent SA method for the general nonsmooth convex stochastic programming problem attaining the optimal convergence rate of $O(1/\sqrt{N})$; Lan and his coauthors [11, 12, 10, 13, 20, 14] proposed various first-order methods for SP problems under suitable convex or non-convex settings. In this paper we also consider (5) in the SP framework. We assume that a noisy gradient information of $\nabla\mathcal{L}_\gamma$ via the so called *stochastic first order oracle* ($\mathcal{SFO}$) is available. Specifically,

for a given x, instead of computing $\nabla f(x)$ we actually only get a stochastic gradient $G(x,\xi)$ from the $\mathcal{SFO}$, where ξ is a random variable following a certain distribution. Formally we introduce:

Definition 1.2 *We call a function $f(x)$ to be* easy for gradient estimation – *denoted as* **GraE** – *if there is an $\mathcal{SFO}$ for f, which returns a stochastic gradient estimation $G(x,\xi)$ for ∇f at x, satisfying*

$$\mathsf{E}[G(x,\xi)] = \nabla f(x), \tag{6}$$

and

$$\mathsf{E}[\|G(x,\xi) - \nabla f(x)\|^2] \leq \sigma^2. \tag{7}$$

If the exact gradient information is available then the $\mathcal{SFO}$ is deterministic. When $\mathcal{L}_\gamma(x,y,\lambda)$ is **MinE** with respect to y, and $f(x)$ in (4) is **GraE**, we will then propose a stochastic gradient ADMM (SGADM), which alternates through one exact minimization step ADMM (3), one stochastic approximation iteration, and an update on the dual variables (multipliers). It is clear that the SGADM in the deterministic case is exactly GADM (5), and we will show that the rate of convergence of GADM and SGADM would be $O(1/N)$ and $O(1/\sqrt{N})$ respectively. Moreover, if $f(x)$ and $g(y)$ in (4) are both **GraE**, then we propose a stochastic gradient augmented Lagrangian method (SGALM), and show that it admits a similar iteration complexity bound.

Furthermore, we are also interested in another class of stochastic problems, where even the noisy gradient information is not available; instead we assume that we can only get the noisy estimation of f via the so-called *stochastic zeroth-order oracle* ($\mathcal{SZO}$). Specifically, for any input x, by calling $\mathcal{SZO}$ once it returns a quantity $F(x,\xi)$, which is a noisy approximation of the true function value $f(x)$. More specifically,

Definition 1.3 *We call a function $f(x)$ to be* easy for function evaluation – *denoted as* **ValE** – *if there is an $\mathcal{SZO}$ for f, which returns a stochastic estimation for f at x if $\mathcal{SZO}$ is called, satisfying*

$$\mathsf{E}[F(x,\xi)] = f(x), \tag{8}$$

$$\mathsf{E}[\nabla F(x,\xi)] = \nabla f(x), \tag{9}$$

and

$$\mathsf{E}[\|\nabla F(x,\xi) - \nabla f(x)\|^2] \leq \sigma^2. \tag{10}$$

Inspired by the work of Nesterov [28] for gradient-free minimization, in this paper we will propose a zeroth-order (gradient-free, a.k.a. direct) smoothing method for (1). Instead of using the Gaussian smoothing scheme as in [28], which has an unbounded support set, we apply another smoothing scheme based on the $\mathcal{SZO}$ of f. To be specific, when $\mathcal{L}_\gamma(x,y,\lambda)$ is **MinE** with respect to y, and $f(x)$ in (4) is **ValE**, we will propose a zeroth-order gradient augmented Lagrangian method (zeroth-order GADM) and analyze its complexity. To summarize, according to the available informational structure of the objective functions, in this paper we propose suitable variants of the ADMM to account for the available information. In a nutshell, the details are in the following Table 1.

		Block x		
		MinE	**GraE**	**ValE**
Block y	**MinE**	ADMM	SGADM	zeroth-order GADM
	GraE	SGADM	SGALM	zeroth-order SGADM
	ValE	zeroth-order GADM	zeroth-order SGADM	zeroth-order GALM

Table 1: A summary of informational-hierarchic alternating direction of multiplier methods.

The rest of the paper is organized as follows. In Section 2, we propose the stochastic gradient ADMM (SGADM) algorithm, and analyze its complexity. In Section 3, we present our stochastic gradient augmented Lagrangian method (SGALM) which uses gradient projection in both block variables, and analyze its convergence rate. In Section 4, we propose a zeroth-order GADM through a new smoothing scheme, and present a complexity result. In Section 5, we present the numerical performance of the SGADM for large-scale convex quadratic programming and the fused logistic regression.

2 The Stochastic Gradient Alternating Direction of Multipliers

In this section, we assume $\mathcal{L}_\gamma(x, y, \lambda)$ to be **MinE** with respect to y, and $f(x)$ to be **GraE**. (We will discuss applications of such model in Section 5). That is, for a given x, whenever we need $\nabla f(x)$, we can actually get a stochastic gradient $G(x, \xi)$ from the $\mathcal{SFO}$, where ξ is a random variable following a certain distribution. Moreover, $G(x, \xi)$ satisfies (6) and (7). By the definition of the augmented Lagrangian $\mathcal{L}_\gamma(x, y, \lambda)$, an $\mathcal{SFO}$ for $\mathcal{L}_\gamma(x, y, \lambda)$ can be constructed accordingly:

Definition 2.1 *Denote the $\mathcal{SFO}$ of $\nabla_x \mathcal{L}_\gamma(x, y, \lambda)$ as $G_L(x, y, \lambda, \xi)$, which is defined as:*

$$G_L(x, y, \lambda, \xi) := G(x, \xi) - A^\top \lambda + \gamma A^\top (Ax + By - b). \tag{11}$$

Our first algorithm to be introduced, SGADM, works as follows:

The Stochastic Gradient ADMM (SGADM)

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

for $k = 0, 1, \cdots,$ **do**

$y^{k+1} = \arg\min_{y \in \mathcal{Y}} \mathcal{L}_\gamma(x^k, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2$;

$x^{k+1} = [x^k - \alpha_k G_L(x^k, y^{k+1}, \lambda^k, \xi^{k+1})]_{\mathcal{X}}$;

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

In the above notation, $[x]_{\mathcal{X}}$ denotes the projection of x onto $\mathcal{X}$, H is a pre-specified positive semidefinite matrix, α_k is the stepsize for the k-th iteration. It is easy to see that the deterministic version of SGADM is exactly GADM (5). In the following subsection, we will show that the complexity of SGADM is $O(1/\sqrt{N})$ and the complexity of GADM is $O(1/N)$.

2.1 Convergence Rate Analysis of the SGADM

In this subsection, we shall analyze the convergence rate of SGADM algorithm. First, some notations and preliminaries are introduced to facilitate the discussion.

2.1.1 Preliminaries and Notations

Denote

$$u = \begin{pmatrix} y \\ x \end{pmatrix}, \quad w = \begin{pmatrix} y \\ x \\ \lambda \end{pmatrix}, \quad F(w) = \begin{pmatrix} -B^\top \lambda \\ -A^\top \lambda \\ Ax + By - b \end{pmatrix}, \tag{12}$$

$h(u) = f(x) + g(y)$, and

$$Q_k = \begin{pmatrix} H & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} & 0 \\ 0 & -A & \frac{1}{\gamma} I_m \end{pmatrix}, \quad P = \begin{pmatrix} I_{n_y} & 0 & 0 \\ 0 & I_{n_x} & 0 \\ 0 & -\gamma A & I_m \end{pmatrix}, \quad M_k = \begin{pmatrix} H & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}. \tag{13}$$

Clearly, $Q_k = M_k P$. In addition to the sequence $\{w^k\}$ generated by the SGADM, we introduce an auxiliary sequence:

$$\tilde{w}^k := \begin{pmatrix} \tilde{y}^k \\ \tilde{x}^k \\ \tilde{\lambda}^k \end{pmatrix} = \begin{pmatrix} y^{k+1} \\ x^{k+1} \\ \lambda^k - \gamma(Ax^k + By^{k+1} - b) \end{pmatrix}. \tag{14}$$

Based on (14) and (13), the relationship between the new sequence $\{\tilde{w}^k\}$ and the original $\{w^k\}$ is

$$w^{k+1} = w^k - P(w^k - \tilde{w}^k). \tag{15}$$

We denote $\delta_k \equiv G(x^{k-1}, \xi^k) - \nabla f(x^{k-1})$, which is the error of the noisy gradient generated by $\mathcal{SFO}$. The following lemma is straightforward.

Lemma 2.1 *For any $w^0, w^1, \cdots, w^{N-1}$, let F be defined in (12) and $\bar{w} = \frac{1}{N} \sum_{k=0}^{N-1} w^k$; then it holds*

$$(\bar{w} - w)^\top F(\bar{w}) = \frac{1}{N} \sum_{k=0}^{N-1} (w^k - w)^\top F(w^k).$$

Proof. Since $F(w) = \begin{pmatrix} 0 & 0 & -B \\ 0 & 0 & -A \\ A & B & 0 \end{pmatrix} \begin{pmatrix} y \\ x \\ \lambda \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ b \end{pmatrix}$, for any w_1 and w_2 we have

$$(w_1 - w_2)^\top (F(w_1) - F(w_2)) = 0. \tag{16}$$

Therefore,

$$\begin{aligned}(\bar{w}-w)^\top F(\bar{w}) &\overset{(16)}{=} (\bar{w}-w)^\top F(w)\\ &= \left(\frac{1}{N}\sum_{k=0}^{N-1} w^k - w\right)^\top F(w)\\ &= \frac{1}{N}\sum_{k=0}^{N-1}(w^k-w)^\top F(w)\\ &\overset{(16)}{=} \frac{1}{N}\sum_{k=0}^{N-1}(w^k-w)^\top F(w^k). \end{aligned} \tag{17}$$

□

2.1.2 The Complexity of SGADM

We now present the rate of convergence result for SGADM, which is $O(1/\sqrt{N})$. Denote $\Xi_k = (\xi_1, \xi_2, \ldots, \xi_k)$. In fact, the convergence rate is in the sense of the expectation taken over Ξ_k.

Theorem 2.2 *Suppose that $\mathcal{L}_\gamma(x,y,\lambda)$ is* **MinE** *with respect to y, and $f(x)$ is* **GraE**. *Let w^k be the sequence generated by the SGADM, $\eta_k = \sqrt{k+1}$, $C>0$ be a constant such that $CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0$, and $\alpha_k = \frac{1}{\eta_k + C}$. Let*

$$\bar{w}_n := \frac{1}{n}\sum_{k=0}^{n-1}\tilde{w}^k, \tag{18}$$

where $\tilde{w}^k$ is defined in (14). Then the following holds

$$\mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \le \frac{\sigma^2}{\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right), \tag{19}$$

where $D_x \equiv \sup_{x_a,x_b\in\mathcal{X}}\|x_a - x_b\|$, $D_y \equiv \sup_{y_a,y_b\in\mathcal{Y}}\|y_a - y_b\|_H$, and $D_\lambda \equiv \sup_{\lambda\in\mathcal{B}_\rho}\|\lambda-\lambda^0\|^2, \mathcal{B}_\rho = \{\lambda : \|\lambda\| \le \rho\}$.

As in [18], we first present a bound regarding the sequence $\{\tilde{w}^k\}$ in (14).

Proposition 2.3 *Let $\{\tilde{w}^k\}$ be defined by (14), and the matrices Q_k, M_k, and P be given in (13). For any $w\in\Omega$, we have*

$$\begin{aligned}&h(u) - h(\tilde{u}^k) + (w-\tilde{w}^k)^\top F(\tilde{w}^k)\\ \ge\ & (w-\tilde{w}^k)^\top Q_k(w^k-\tilde{w}^k) - (x-x^k)^\top\delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} - \frac{\eta_k+L}{2}\|x^k-\tilde{x}^k\|^2, \end{aligned} \tag{20}$$

where $\eta_k > 0$ is any constant. Moreover, for any $w \in \Omega$, the term $(w - \tilde{w}^k)^\top Q_k(w^k - \tilde{w}^k)$ on the RHS of (20) can be further bounded below as follows

$$\begin{aligned} & (w - \tilde{w}^k)^\top Q_k(w^k - \tilde{w}^k) \\ \geq & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k). \end{aligned} \tag{21}$$

The proof of Proposition 2.3 involves several steps. In order not to distract the flow of presentation, we delegate its proof to the appendix.

Proof of Theorem 2.2

Proof. Recall that $CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0$ and $\alpha_k = \frac{1}{\eta_k + C}$. By (20) and (21),

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k) \\ & -(x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\|x^k - \tilde{x}^k\|^2 \\ = & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A - (\eta_k + L) I_{n_x}\right)(x^k - \tilde{x}^k) \\ & -(x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k} \\ \geq & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) - (x - x^k)^\top \delta_{k+1} - \frac{\|\delta_{k+1}\|^2}{2\eta_k}. \end{aligned} \tag{22}$$

Using the definition of M_k, from (22) we have

$$\begin{aligned} & h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\ \leq & \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - y^{k+1}\|_H^2\right) + \frac{1}{2\gamma}\left(\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2\right) \\ & + \frac{\|x - x^k\|^2 - \|x - x^{k+1}\|^2}{2\alpha_k} + (x - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}. \end{aligned} \tag{23}$$

Summing up the inequalities (23) for $k = 0, 1, \ldots, N-1$ we have

$$\begin{aligned} & h(\bar{u}_N) - h(u) + (\bar{w}_N - w)^\top F(\bar{w}_N) \\ \leq & \frac{1}{N}\sum_{k=0}^{N-1} h(\tilde{u}^k) - h(u) + \frac{1}{N}\sum_{k=0}^{N-1} (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\ \leq & \frac{1}{2N}\sum_{k=0}^{N-1} \frac{\|x - x^k\|^2 - \|x - x^{k+1}\|^2}{\alpha_k} + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] \\ & + \frac{1}{2N}\left(\|y - y^0\|_H^2 + \frac{1}{\gamma}\|\lambda - \lambda^0\|^2\right), \end{aligned} \tag{24}$$

where the first inequality is due to the convexity of h and Lemma 2.1.

Note the above inequality is true for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $\lambda \in \mathbf{R}^m$, hence it is also true for the optimal solution x^*, y^*, and $\mathcal{B}_\rho = \{\lambda : \|\lambda\| \le \rho\}$. As a result,

$$\begin{aligned}
&\sup_{\lambda\in\mathcal{B}_\rho}\Big[h(\bar{u}_N) - h(u^*) + (\bar{w}_N - w^*)^\top F(\bar{w}_N)\Big] \\
=\ &\sup_{\lambda\in\mathcal{B}_\rho}\Big[h(\bar{u}_N) - h(u^*) + (\bar{x}_N - x^*)^\top(-A^\top\bar{\lambda}_N) + (\bar{y}_N - y^*)^\top(-B^\top\bar{\lambda}_N) + (\bar{\lambda}_N - \lambda)^\top(A\bar{x}_N + B\bar{y}_N - b)\Big] \\
=\ &\sup_{\lambda\in\mathcal{B}_\rho}\Big[h(\bar{u}_N) - h(u^*) + \bar{\lambda}_N^\top(Ax^* + By^* - b) - \lambda^\top(A\bar{x}_N + B\bar{y}_N - b)\Big] \\
=\ &\sup_{\lambda\in\mathcal{B}_\rho}\Big[h(\bar{u}_N) - h(u^*) - \lambda^\top(A\bar{x}_N + B\bar{y}_N - b)\Big] \\
=\ &h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|. \qquad (25)
\end{aligned}$$

Combining (24) and (25), we have

$$\begin{aligned}
&h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\le\ &\frac{1}{2N}\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} + \frac{1}{N}\sum_{k=1}^{N-1}\left[(x^* - x^k)^\top\delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] \\
&+\frac{1}{2N}\left(\|y^* - y^0\|_H^2 + \frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right). \qquad (26)
\end{aligned}$$

Moreover, since $\alpha_k = \frac{1}{\eta_k + C} = \frac{1}{\sqrt{k+1}+C}$, it follows that

$$\begin{aligned}
&\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} \\
=\ &\sum_{k=0}^{N-1}(\sqrt{k+1} + C)(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2) \\
\le\ &C\|x^* - x^0\|^2 + \sum_{k=0}^{N-1}\sqrt{k+1}(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2) \\
\le\ &C\|x^* - x^0\|^2 + \sum_{k=0}^{N-1}\|x^* - x^k\|^2(\sqrt{k+1} - \sqrt{k}) \\
\le\ &C\|x^* - x^0\|^2 + \sum_{k=0}^{N-1}D_x^2(\sqrt{k+1} - \sqrt{k}) \\
=\ &C\|x^* - x^0\|^2 + \sqrt{N}D_x^2. \qquad (27)
\end{aligned}$$

Plugging (27) into (26) we have

$$\begin{aligned}
&h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\le\ &\frac{1}{N}\sum_{k=0}^{N-1}\left[(x - x^k)^\top\delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k}\right] + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right). \qquad (28)
\end{aligned}$$

Recall that $f(x)$ is **GraE**, so (6) and (7) hold. Consequently, $\mathsf{E}[\delta_{k+1}] = \mathsf{E}[G(x^k, \xi^{k+1}) - \nabla f(x^k)] = 0$. In addition, x_k is independent of ξ_{k+1}. Hence,

$$\mathsf{E}_{\Xi_{k+1}}[(x - x^k)^\top \delta_{k+1}] = 0. \tag{29}$$

Now, taking expectation over (28), and applying (7), we have

$$\begin{aligned}
& \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \mathsf{E}_{\Xi_N}\left[\frac{1}{N}\sum_{k=0}^{N-1}((x^* - x^k)^\top \delta_{k+1} + \frac{\|\delta_{k+1}\|^2}{2\eta_k})\right] + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right) \\
\overset{(7)}{\leq}\ & \frac{1}{N}\mathsf{E}_{\Xi_N}\left[\sum_{k=0}^{N-1}(x^* - x^k)^\top \delta_{k+1}\right] + \frac{\sigma^2}{2N}\sum_{k=0}^{N-1}\frac{1}{\eta_k} + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right) \\
\overset{(29)}{=}\ & \frac{\sigma^2}{2N}\sum_{k=0}^{N-1}\frac{1}{\sqrt{k+1}} + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right) \\
\leq\ & \frac{\sigma^2}{2N}\int_1^{N+1}\frac{1}{\sqrt{t}}dt + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right) \\
=\ & \frac{\sigma^2}{\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right).
\end{aligned} \tag{30}$$

This completes the proof. □

Before concluding this section, some comments are in order here. Denote $\hat{u}_N = \mathsf{E}_{\Xi_N}[\bar{u}_N]$. By Jensen's inequality, an immediate consequence is that we have

$$h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \leq \frac{\sigma^2}{\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right).$$

In the above theorem, we see that SGADM has a convergence rate of $O(1/\sqrt{N})$ when $f(x)$ is **GraE**. As we mentioned before, it is easy to slightly modify the proof for (19) to improve the complexity of GADM (i.e. the deterministic SGADM) to $O(1/N)$. In fact, when the exact gradient of f is available, σ in (7) will be 0, and we can let $\eta_k = 1$ and constant stepsize $\alpha_k = \frac{1}{C+1}$. As a result, a sharper bound for the term $\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k}$ would follow:

$$\begin{aligned}
\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} &= \sum_{k=0}^{N-1}(1+C)\left(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2\right) \\
&\leq (C+1)\|x^* - x^0\|^2 \leq (C+1)D_x^2.
\end{aligned} \tag{31}$$

Similar to the proof of Theorem 2.2, but now we can improve the iteration bound to:

$$h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \leq \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + (C+1)D_x^2\right), \tag{32}$$

where $D_x \equiv \sup_{x_a,x_b\in\mathcal{X}} \|x_a - x_b\|$, $D_y \equiv \sup_{y_a,y_b\in\mathcal{Y}} \|y_a - y_b\|_H$, and $D_\lambda \equiv \sup_{\lambda\in\mathcal{B}_\rho} \|\lambda - \lambda^0\|^2, \mathcal{B}_\rho = \{\lambda : \|\lambda\| \le \rho\}$, and this indeed proves the $O(1/N)$ complexity of the SGADM for the deterministic case.

To further assess the feasibility violation of the possibly infeasible solution $\bar{u}_N$ as in (32), similar to Lemma 6 in [20] we introduce the following bound.

Lemma 2.4 *Assume that $\rho > 0$, and $\tilde{x} \in X$ is an approximate solution of the problem $f^* := \inf\{f(x) : Ax - b = 0, x \in X\}$ where f is convex, satisfying*

$$f(\tilde{x}) - f^* + \rho\|A\tilde{x} - b\| \le \delta. \tag{33}$$

Then, we have

$$\|A\tilde{x} - b\| \le \frac{\delta}{\rho - \|y^*\|} \text{ and } f(\tilde{x}) - f^* \le \delta$$

where y^* is an optimal Lagrange multiplier associated with the problem $\inf\{f(x) : Ax - b = 0, x \in X\}$ satisfying $\|y^*\| < \rho$.

Proof. Define $v(u) := \inf\{f(x) : Ax - b = u, x \in X\}$, which is convex. Let y^* be such that $-y^* \in \partial v(0)$. Thus, we have

$$v(u) - v(0) \ge \langle -y^*, u\rangle \quad \forall u \in \mathbf{R}^m. \tag{34}$$

Let $u := A\tilde{x} - b$. Since $v(u) \le f(\tilde{x})$ and $v(0) = f^*$, we have

$$-\|y^*\|\|u\| + \rho\|u\| \le \langle -y^*, u\rangle + \rho\|u\| \le v(u) - v(0) + \rho\|u\| \le f(\tilde{x}) - f^* + \rho\|u\| \le \delta.$$

Thus, $\|A\tilde{x} - b\| = \|u\| \le \frac{\delta}{\rho - \|y^*\|}$, and $f(\tilde{x}) - f^* \le \delta$. □

Lemma 2.4 suggests that, when ρ is sufficiently large, $h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \le \epsilon$ implies

$$|h(\hat{u}_N) - h(u^*)| \le O(\epsilon) \text{ and } \|A\hat{x}_N + B\hat{y}_N - b\| \le O(\epsilon).$$

3 The Stochastic Gradient Augmented Lagrangian Method

SGADM uses gradient projection in one block of variables. It is natural to relax the exact minimization procedure of the other block variables to be replaced by gradient projection too. In this section, we assume both $f(x)$ and $g(y)$ in (4) are **GraE**; that is, we can only get stochastic gradients $S_f(x,\xi)$ and $S_g(y,\zeta)$ from the $\mathcal{SFO}$ for $\nabla f(x)$ and $\nabla g(y)$ respectively, where ξ and ζ are certain random variables. Recall the assumptions on **GraE**:

$$\mathsf{E}[S_f(x,\xi)] = \nabla f(x), \quad \mathsf{E}[S_g(y,\zeta)] = \nabla g(y), \tag{35}$$

and

$$\mathsf{E}[\|S_f(x,\xi)-\nabla f(x)\|^2]\le\sigma_1^2,\quad \mathsf{E}[\|S_g(y,\zeta)-\nabla g(y)\|^2]\le\sigma_2^2. \tag{36}$$

We now propose a stochastic gradient augmented Lagrangian method (SGALM). Given $\mathcal{SFO}$ for f and g, the $\mathcal{SFO}$ for $\nabla_x L_\gamma(x,y,\lambda)$ and $\nabla_y L_\gamma(x,y,\lambda)$ can be constructed as:

$$S_L^f(x,y,\lambda,\xi) := S_f(x,\xi)-A^\top\lambda+\gamma A^\top(Ax+By-b), \tag{37}$$
$$S_L^g(x,y,\lambda,\zeta) := S_g(y,\zeta)-B^\top\lambda+\gamma B^\top(Ax+By-b). \tag{38}$$

Our next algorithm, SGALM, works as follows:

The Stochastic Gradient Augmented Lagrangian Method (SGALM)

Initialize $x^0\in\mathcal{X}, y^0\in\mathcal{Y}$ and λ^0

for $k=0,1,\cdots,$ **do**

$y^{k+1}=[y^k-\beta_k S_L^g(x^k,y^k,\lambda^k,\zeta^{k+1})]_{\mathcal{Y}}$;

$x^{k+1}=[x^k-\alpha_k S_L^f(x^k,y^{k+1},\lambda^k,\xi^{k+1})]_{\mathcal{X}}$;

$\lambda^{k+1}=\lambda^k-\gamma(Ax^{k+1}+By^{k+1}-b)$.

end for

Denote

$$\delta_{k+1}^f := S_f(x^k,\xi^{k+1})-\nabla f(x^k),\ \delta_{k+1}^g := S_g(y^k,\zeta^{k+1})-\nabla g(y^k).$$

Throughout this section, we assume that the gradient ∇g is also Lipschitz continuous. For notational simplicity, we assume its Lipschitz constant is also L. Then, we are able to analyze the convergence rate of SGALM. Denote

$$\hat{Q}_k=\begin{pmatrix} H_k & 0 & 0\\ 0 & \frac{1}{\alpha_k}I_{n_x} & 0\\ 0 & -A & \frac{1}{\gamma}I_m\end{pmatrix},\quad \hat{M}_k=\begin{pmatrix} H_k & 0 & 0\\ 0 & \frac{1}{\alpha_k}I_{n_x} & 0\\ 0 & 0 & \frac{1}{\gamma}I_m\end{pmatrix} \tag{39}$$

where $H_k=\frac{1}{\beta_k}I_{n_y}-\gamma B^\top B$. The identity $\hat{Q}_k=\hat{M}_k P$ still holds where P is given according to (13).

Similar to Proposition 2.3, we have the following bounds regarding the sequence $\{\tilde{w}^k\}$ defined in (14), the proof of which is also delegated to the appendix.

Proposition 3.1 *Suppose that $\{\tilde{w}^k\}$ is given as in (14), and the matrices $\hat{Q}_k$ and $\hat{M}_k$ are given as in (39). For any $w\in\Omega$, we have*

$$\begin{aligned} & h(u)-h(\tilde{u}^k)+(w-\tilde{w}^k)^\top F(\tilde{w}^k)\\ \ge\ & (w-\tilde{w}^k)^\top\hat{Q}_k(w^k-\tilde{w}^k)-(x-x^k)^\top\delta_{k+1}^f-(y-y^k)^\top\delta_{k+1}^g\\ & -\frac{\|\delta_{k+1}^f\|^2+\|\delta_{k+1}^g\|^2}{2\eta_k}-\frac{\eta_k+L}{2}\left(\|x^k-\tilde{x}^k\|^2+\|y^k-\tilde{y}^k\|^2\right), \end{aligned} \tag{40}$$

where $\eta_k > 0$ is any prescribed sequence. Moreover, the term $(w - \tilde{w}^k)^\top \hat{Q}_k(w^k - \tilde{w}^k)$ on the RHS can be further bounded as follows

$$\begin{aligned} & (w - \tilde{w}^k)^\top \hat{M}_k P(w^k - \tilde{w}^k) \\ \geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_{\hat{M}_k} - \|w - w^k\|^2_{\hat{M}_k}\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k) \\ & + \frac{1}{2}(y^k - \tilde{y}^k)^\top \left(\frac{1}{\beta_k} I_{n_y} - \gamma B^\top B\right)(y^k - \tilde{y}^k), \quad \forall w \in \Omega, \end{aligned} \tag{41}$$

where by abusing the notation a bit we denote $\|x\|^2_A := x^\top A x$ with A being a symmetric matrix but not necessarily positive semidefinite.

Now, we are in a position to present our main convergence rate result for the SGALM algorithm. Let us recycle the notation and denote $\Xi_k = (\xi_1, \xi_2, \ldots, \xi_k, \zeta_1, \zeta_2, \ldots, \zeta_k)$; the convergence rate will be in the expectation over Ξ_k.

Theorem 3.2 *Suppose both $f(x)$ and $g(y)$ in (4) are* **GraE**. *Let w^k be the sequence generated by the SGALM, $\eta_k = \sqrt{k+1}$, and C is a constant satisfying*

$$C I_{n_x} - \gamma A^\top A - L I_{n_x} \succeq 0 \text{ and } C I_{n_y} - \gamma B^\top B - L I_{n_y} \succeq 0,$$

and $\beta_k = \alpha_k = \frac{1}{\eta_k + C}$. For any integer $n > 0$, let

$$\bar{w}_n = \frac{1}{n} \sum_{k=0}^{n-1} \tilde{w}^k, \tag{42}$$

where $\tilde{w}^k$ is defined in (14). Then

$$\begin{aligned} & \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\ \leq\ & \frac{\sigma_1^2 + \sigma_2^2}{\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(C D_y^2 + C D_x^2 + \frac{1}{\gamma} D_\lambda^2\right), \end{aligned} \tag{43}$$

where $D_x \equiv \sup_{x_a, x_b \in \mathcal{X}} \|x_a - x_b\|$, $D_y \equiv \sup_{y_a, y_b \in \mathcal{Y}} \|y_a - y_b\|$, and $D_\lambda \equiv \sup_{\lambda \in \mathcal{B}_\rho} \|\lambda - \lambda^0\|^2, \mathcal{B}_\rho = \{\lambda : \|\lambda\| \leq \rho\}$.

Proof. Similar to (22), by (40) and (41) we have

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_{\hat{M}_k} - \|w - w^k\|^2_{\hat{M}_k}\right) - (x - x^k)^\top \delta^f_{k+1} - (y - y^k)^\top \delta^g_{k+1} - \frac{\|\delta^f_{k+1}\|^2 + \|\delta^g_{k+1}\|^2}{2\eta_k}. \end{aligned}$$

Following a similar line of arguments as in Theorem 2.2, we derive that

$$\begin{aligned}
& h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\leq\ & \frac{1}{2N}\sum_{k=0}^{N-1}\frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} + \frac{1}{2N}\sum_{k=0}^{N-1}\left(\|y^* - y^k\|_{H_k}^2 - \|y^* - y^{k+1}\|_{H_k}^2\right) \\
& + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta_{k+1}^f + (y^* - y^k)^\top \delta_{k+1}^g + \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k}\right] \\
& + \frac{1}{2N}\left(\frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right). \qquad (44)
\end{aligned}$$

Compared to (26), the term $\sum_{k=0}^{N-1}(\|y^* - y^k\|_{H_k}^2 - \|y^* - y^{k+1}\|_{H_k}^2)$ is new. Since $\beta_k = \frac{1}{\eta_k + C} = \frac{1}{\sqrt{k+1}+C}$, we have $H_B = CI_{n_y} - \gamma B^\top B \succeq 0$. Thus,

$$\begin{aligned}
& \sum_{k=0}^{N-1}\left(\|y^* - y^k\|_{H_k}^2 - \|y^* - y^{k+1}\|_{H_k}^2\right) \\
=\ & \sum_{k=0}^{N-1}\sqrt{k+1}\left(\|y^* - y^k\|^2 - \|y^* - y^{k+1}\|^2\right) + \sum_{k=0}^{N-1}\left(\|y^* - y^k\|_{H_B}^2 - \|y^* - y^{k+1}\|_{H_B}^2\right) \\
\leq\ & \|y^* - y^0\|_{H_B}^2 + \sum_{k=0}^{N-1}\sqrt{k+1}\left(\|y^* - y^k\|^2 - \|y^* - y^{k+1}\|^2\right) \\
\leq\ & C\|y^* - y^0\|^2 + \sum_{k=0}^{N-1}\|y^* - y^k\|^2\left(\sqrt{k+1} - \sqrt{k}\right) \\
\leq\ & C\|y^* - y^0\|^2 + \sum_{k=0}^{N-1}D_y^2\left(\sqrt{k+1} - \sqrt{k}\right) \\
=\ & C\|y^* - y^0\|^2 + \sqrt{N}D_y^2. \qquad (45)
\end{aligned}$$

Moreover, according to (27), the term $\sum_{k=0}^{N-1}\frac{\|x^*-x^k\|^2 - \|x^*-x^{k+1}\|^2}{\alpha_k}$ is bounded above by $C\|x^* - x^0\|^2 + \sqrt{N}D_x^2$. Consequently, we can further upper bound (44) as follows:

$$\begin{aligned}
& h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\leq\ & \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta_{k+1}^f + (y^* - y^k)^\top \delta_{k+1}^g + \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k}\right] \\
& + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(CD_y^2 + CD_x^2 + \frac{1}{\gamma}D_\lambda^2\right). \qquad (46)
\end{aligned}$$

Recall that $\delta_{k+1}^f = S_f(x^k, \xi^{k+1}) - \nabla f(x^k)$, $\delta_{k+1}^g = S_g(y^k, \zeta^{k+1}) - \nabla g(y^k)$ and (35) holds. Since x_k is independent of ξ_{k+1} and y_k is independent of ζ_{k+1}, we have

$$\mathsf{E}_{\Xi_{k+1}}\left[(x - x^k)^\top \delta_{k+1}^f\right] = 0, \quad \mathsf{E}_{\Xi_{k+1}}\left[(y - y^k)^\top \delta_{k+1}^g\right] = 0. \tag{47}$$

Now, taking the expectation over (46), and applying (36), one has

$$\begin{aligned}
& \mathsf{E}_{\Xi_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \mathsf{E}_{\Xi_N}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left((x^* - x^k)^\top \delta_{k+1}^f + (y^* - y^k)^\top \delta_{k+1}^g + \frac{\|\delta_{k+1}^f\|^2 + \|\delta_{k+1}^g\|^2}{2\eta_k}\right)\right] \\
& + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(CD_y^2 + CD_x^2 + \frac{1}{\gamma}D_\lambda^2\right) \\
\leq\ & \frac{\sigma_1^2 + \sigma_2^2}{\sqrt{N}} + \frac{D_x^2}{2\sqrt{N}} + \frac{D_y^2}{2\sqrt{N}} + \frac{1}{2N}\left(CD_y^2 + CD_x^2 + \frac{1}{\gamma}D_\lambda^2\right).
\end{aligned} \tag{48}$$

□

Before concluding this section, some comments are in order here. The complexity of $O(1/\sqrt{N})$ for SGALM algorithm is the same order of magnitude as that of SGADM. In view of (43), it is easy to see that the complexity of SGALM for the deterministic setting would be $O(1/N)$, since in that case σ_1 and σ_2 in (36) are 0, and we can let $\eta_k = 1$ in Theorem 3.2. Then, following similar argument after Theorem 2.2, the following complexity bound is achieved:

$$h(\hat{u}_N) - h(u^*) + \rho\|A\hat{x}_N + B\hat{y}_N - b\| \leq \frac{1}{2N}\left((C+1)(D_x^2 + D_y^2) + \frac{1}{\gamma}D_\lambda^2\right), \tag{49}$$

where $D_x \equiv \sup_{x_a, x_b \in \mathcal{X}} \|x_a - x_b\|$, $D_y \equiv \sup_{y_a, y_b \in \mathcal{Y}} \|y_a - y_b\|$, and $D_\lambda \equiv \sup_{\lambda \in \mathcal{B}_\rho} \|\lambda - \lambda^0\|^2, \mathcal{B}_\rho = \{\lambda : \|\lambda\| \leq \rho\}$, and $\hat{u}_N = \mathsf{E}_{\Xi_N}[\bar{u}_N]$, which shows an $O(1/N)$ complexity bound for the SGALM in the deterministic case.

4 The Stochastic Zeroth-Order GADM

In this section, we consider another setting, where even the noisy gradient of $f(x)$ is not available. To be specific, we assume that $\mathcal{L}_\gamma(x, y, \lambda)$ is **MinE** with respect to y, and $f(x)$ is **ValE**. In other words, for any given x we can get a noisy approximation of the true function value $f(x)$ by calling an $\mathcal{SZO}$, which returns a quantity $F(x, \xi)$ with ξ being a certain random variable.

Now that we can access the $\mathcal{SZO}$, we shall use the smoothing scheme proposed in [28] to approximate the first order information of a given function f. The smoothing technique is to utilize the integration operator to promote the differentiability. More specifically, suppose that v is a random

vector in $\mathbf{R}^n$ with density function ρ. A smooth approximation of f with the smoothing parameter μ is defined as:

$$f_\mu(x) = \int f(x + \mu v)\rho(v)dv. \tag{50}$$

Theoretically, one can choose to use any pre-specified smoothing distribution $\rho(v)$. For instance, in [28] Nesterov adopted the Gaussian distribution to simplify the computation. However, the Gaussian distribution has a support set of the whole space $\mathbf{R}^n$, which cannot be implemented for problems with constraints. To avoid using the entire space as the sample space, in this paper we shall use the smoothing scheme based on the uniform distribution over a (scalable) ball in $\mathbf{R}^n$ as introduced in [35].

Definition 4.1 *Let U_b be the uniform distribution over the unit Euclidean ball and B be the unit ball. Given $\mu > 0$, the smoothing function f_μ is defined as*

$$f_\mu(x) = \mathsf{E}_{\{v\sim U_b\}}[f(x + \mu v)] = \frac{1}{\alpha(n)} \int_B f(x + \mu v)dv \tag{51}$$

where $\alpha(n)$ is the volume of the unit ball in $\mathbf{R}^n$.

Some properties of the smoothing function are shown in the lemma below, which will be used in our forthcoming discussion; the proof of the lemma can be found in the appendix.

Lemma 4.1 *Suppose that $f \in C^1_L(\mathbf{R}^n)$. Let U_{S_p} be the uniform distribution over the unit Euclidean sphere, and S_p be the unit sphere in $\mathbf{R}^n$. Then we have:*

(a) The smoothing function f_μ is continuously differentiable, and its gradient is Lipschitz continuous with constant $L_\mu \le L$ and

$$\nabla f_\mu(x) = \mathsf{E}_{\{v\sim U_{S_p}\}}\left[\frac{n}{\mu} f(x + \mu v)v\right] = \frac{1}{\beta(n)} \int_{v\in S_p} \frac{n}{\mu} \left[f(x + \mu v) - f(x)\right] dv \tag{52}$$

where $\beta(n)$ is the surface area of the unit sphere in $\mathbf{R}^n$.

(b) For any $x \in \mathbf{R}^n$, we have

$$|f_\mu(x) - f(x)| \le \frac{L\mu^2}{2}, \tag{53}$$

$$\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{\mu n L}{2}, \tag{54}$$

$$\mathsf{E}_v\left[\left\|\frac{n}{\mu}[f(x + \mu v) - f(x)]v\right\|^2\right] \le 2n\|\nabla f(x)\|^2 + \frac{\mu^2}{2}L^2n^2. \tag{55}$$

(c) If f is convex, then f_μ is also convex.

We remark that the bounds in Part (b) are slightly sharper (up to some constant factor) than that of Gaussian smoothing scheme in [28]. Now based on (52) we define the zeroth-order stochastic gradient of f at point x^k:

$$G_\mu(x^k, \xi_{k+1}, v) = \frac{n_x}{\mu}\left[F(x^k + \mu v, \xi_{k+1}) - F(x^k, \xi_{k+1})\right] v, \tag{56}$$

where v is the random vector uniformly distributed over the unit sphere in $\mathbf{R}^{n_x}$. The zeroth-order GADM algorithm is described as follows:

The Zeroth-Order GADM

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

for $k = 0, 1, \cdots,$ **do**

$y^{k+1} = \arg\min_{y\in\mathcal{Y}} \mathcal{L}_\gamma(x^k, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2$.

At the k-th iteration, we call the $\mathcal{SZO}$ m_k times to obtain $G_\mu(x_k, \xi_{k+1,i}, v_{k+1,i}), i = 1, \cdots, m_k$.

Then set $G_{\mu,k} = \frac{1}{m_k}\sum_{i=1}^{m_k} G_\mu(x^k, \xi_{k+1,i}, v_{k+1,i})$, and compute

$x^{k+1} = [x^k - \alpha_k(G_{\mu,k} - A^\top\lambda^k + \gamma A^\top(Ax^k + By^{k+1} - b))]_{\mathcal{X}}$;

$\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.

end for

Before conducting the complexity analysis for the algorithm above, we present some properties of the function $G(x^k, \xi_{k+1}) := \nabla_x F(x^k, \xi_{k+1})$. Note that function f is **ValE**, i.e. (8) and (9) hold. This fact together with Lemma 4.1(a) leads to:

Lemma 4.2 *Suppose that $G_\mu(x^k, \xi_{k+1}, v)$ is defined as in (56), and f is* **ValE**, *i.e. (8), (9) and (10) hold. Then*

$$\mathsf{E}_{v,\xi_{k+1}}[G_\mu(x^k, \xi_{k+1}, v)] = \nabla f_\mu(x^k). \tag{57}$$

If we further assume $\|\nabla f(x)\| \le M$, $\forall x \in \mathcal{X}$, then the following holds

$$\mathsf{E}_{v,\xi_{k+1}}[\|G_\mu(x^k, \xi_{k+1}, v) - \nabla f_\mu(x^k)\|^2] \le \tilde{\sigma}^2, \tag{58}$$

where $\tilde{\sigma}^2 = 2n_x[M^2 + \sigma^2 + \mu^2 L^2 n_x]$.

Proof. The first statement is easy to verify. We shall focus on the second statement. Applying (55) and (10) to $F(x^k, \xi_{k+1})$ and $G(x^k, \xi_{k+1})$, we have

$$\begin{aligned}
& \mathsf{E}_{v,\xi_{k+1}}\left[\|G_\mu(x^k, \xi_{k+1}, v)\|^2\right] \\
= \; & \mathsf{E}_{\xi_{k+1}}\left[\mathsf{E}_v\left[\|G_\mu(x^k, \xi_{k+1}, v)\|^2\right]\right] \\
\le \; & 2n_x\left[\mathsf{E}_{\xi_{k+1}}\left[\|G(x^k, \xi_{k+1})\|^2\right]\right] + \frac{\mu^2}{2}L^2 n_x^2 \\
\le \; & 2n_x\left\{\mathsf{E}_{\xi_{k+1}}[\|\nabla f(x^k)\|^2] + \mathsf{E}_{\xi_{k+1}}\left[\|G(x^k, \xi_{k+1}) - \nabla f(x^k)\|^2\right]\right\} + \mu^2 L^2 n_x^2 \\
\le \; & 2n_x\left\{\|\nabla f(x^k)\|^2 + \sigma^2\right\} + \mu^2 L^2 n_x^2.
\end{aligned} \tag{59}$$

Then from (59), (57), and $\|\nabla f(x^k)\| \le M$, we have

$$\begin{aligned} & \mathsf{E}_{v,\xi_{k+1}}\left[\|G_\mu(x^k,\xi_{k+1},v)-\nabla f_\mu(x^k)\|^2\right] \\ = \; & \mathsf{E}_{v,\xi_{k+1}}\left[\|G_\mu(x^k,\xi_{k+1},v)\|^2\right]-\|\nabla f_\mu(x)\|^2 \\ \le \; & 2n_x\left[M^2+\sigma^2+\mu^2L^2n_x\right]=\tilde{\sigma}^2. \end{aligned} \tag{60}$$

□

4.1 Convergence Rate of Zeroth-Order GADM

To establish the convergence rate, we refer the sequence $\tilde{w}^k$ to be the sequence defined in (14) with the corresponding iterates x^k, y^k, λ^k obtained from the zeroth-order GADM. We let $\delta_{\mu,k} = G_{\mu,k} - \nabla f_\mu(x_k)$, which plays a similar role as δ_k in SGADM. We have the following proposition, whose proof is almost identical to that of (40) in Proposition 3.1 except that δ_{k+1} is now replaced by $\delta_{\mu,k}$.

Proposition 4.3 *Suppose that $\mathcal{L}_\gamma(x,y,\lambda)$ is* **MinE** *with respect to y, and $f(x)$ is* **ValE**. *Let x^k, y^k, λ^k be obtained in the zeroth-order GADM, $\tilde{w}^k$ be specified as in (14), and $h_\mu(u) = f_\mu(x)+g(y)$. Then for any $w \in \Omega$, we have*

$$h_\mu(u)-h_\mu(\tilde{u}^k)+(w-\tilde{w}^k)F(\tilde{w}^k) \ge (w-\tilde{w}^k)^\top Q_k(w^k-\tilde{w}^k)-(x-x^k)^\top\delta_{\mu,k}-\frac{\|\delta_{\mu,k}\|^2}{2\eta_k}-\frac{\eta_k+L}{2}\|x^k-\tilde{x}^k\|^2, \tag{61}$$

where $\eta_k > 0$ can be any positive constant to be specified in the analysis later.

Now, we are ready to present the following theorem which leads to the convergence rate of the zeroth-order GADM. In the rest of this section, we denote $\Omega_n = (\xi_{k,i}, v_{k,i})$ for $k = 1,2,\ldots,n$ and $i = 1,2,\ldots,m_k$, the convergence rate will be considered in the expectation taken on Ω_N.

Theorem 4.4 *Let w^k be the sequence generated by the zeroth-order GADM, and C be a constant such that $CI_{n_x} - \gamma A^\top A - LI_{n_x} \succeq 0$, and $\alpha_k = \frac{1}{\eta_k+C}$. For any integer $n > 0$, let*

$$\bar{w}_n = \frac{1}{n}\sum_{k=0}^{n-1}\tilde{w}^k, \tag{62}$$

where $\tilde{w}^k$ is defined in (14). Then the following holds

$$\begin{aligned} & \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N)-h(u^*)+\rho\|A\bar{x}_N+B\bar{y}_N-b\|\right] \\ \le \; & \frac{1}{2N}\sum_{k=1}^{N}\eta_k(\|x^*-x^k\|^2-\|x^*-x^{k+1}\|^2) \\ & +\frac{\tilde{\sigma}^2}{2N}\sum_{k=1}^{N}\frac{1}{m_k\eta_k}+\frac{1}{2N}(D_y^2+\frac{1}{\gamma}D_\lambda^2+CD_x^2)+L\mu^2, \end{aligned} \tag{63}$$

where $D_x \equiv \sup_{x_a,x_b\in\mathcal{X}} \|x_a - x_b\|$, $D_y \equiv \sup_{y_a,y_b\in\mathcal{Y}} \|y_a - y_b\|_H$, *and* $D_\lambda \equiv \sup_{\lambda\in\mathcal{B}_\rho} \|\lambda - \lambda^1\|^2, \mathcal{B}_\rho = \{\lambda : \|\lambda\| \le \rho\}$, *and* $\{\eta_k > 0\}$ *can be constants.*

Proof. By (61) and (21), it follows that

$$\begin{aligned}
& h_\mu(u) - h_\mu(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\ge\ & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k) \\
& -(x - x^k)^\top \delta_{\mu,k} - \frac{\|\delta_{\mu,k}\|^2}{2\eta_k} - \frac{\eta_k + L}{2}\|x^k - \tilde{x}^k\|^2 \\
=\ & \frac{1}{2}(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2) + \frac{1}{2}(x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A - (\eta_k + L) I_{n_x}\right)(x^k - \tilde{x}^k) \\
& -(x - x^k)^\top \delta_{\mu,k} - \frac{\|\delta_{\mu,k}\|^2}{2\eta_k} \\
\ge\ & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) - (x - x^k)^\top \delta_{\mu,k} - \frac{\|\delta_{\mu,k}\|^2}{2\eta_k}.
\end{aligned}$$

In similar vein as the proof of (26) in Theorem 2.2 (except that δ_{k+1} is replaced by $\delta_{\mu,k}$), we obtain:

$$\begin{aligned}
& h_\mu(\bar{u}_N) - h_\mu(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\| \\
\le\ & \frac{1}{2N}\sum_{k=0}^{N-1} \frac{\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2}{\alpha_k} + \frac{1}{N}\sum_{k=0}^{N-1}\left[(x^* - x^k)^\top \delta_{\mu,k} + \frac{\|\delta_{\mu,k}\|^2}{2\eta_k}\right] \\
& + \frac{1}{2N}\left(\|y^* - y^0\|_H^2 + \frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right).
\end{aligned} \tag{64}$$

Recall that $\delta_{\mu,k} = G_{\mu,k} - \nabla f_\mu(x_k)$, which combined with (57) implies

$$\mathsf{E}_{\xi_{k+1},v_{k+1}}[\delta_{\mu,k}] = \mathsf{E}_{\xi_{k+1},v_{k+1}}[G_{\mu,k} - \nabla f_\mu(x_k)] = 0.$$

In addition, since ξ_{k+1} and v_{k+1} are independent to x_k, we have the following identity

$$\mathsf{E}_{\Omega_{k+1}}[(x^* - x^k)^\top \delta_{\mu,k}] = 0. \tag{65}$$

Now, taking expectation over (64), and applying (58), we have

$$\begin{aligned}
& \mathsf{E}_{\Omega_N}\left[h_\mu(\bar{u}_N) - h_\mu(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\le\ & \mathsf{E}_{\Omega_N}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left((x^* - x^k)^\top \delta_{\mu,k} + \frac{\|\delta_{\mu,k}\|^2}{2\eta_k}\right)\right] \\
& + \frac{1}{2N}\sum_{k=0}^{N-1}\eta_k\left(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2\right) + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right) \\
\overset{(58)}{\le}\ & \frac{\tilde{\sigma}^2}{N}\sum_{k=0}^{N-1}\frac{1}{m_k\eta_k} + \frac{1}{2N}\sum_{k=0}^{N-1}\eta_k\left(\|x^* - x^k\|^2 - \|x^* - x^{k+1}\|^2\right) + \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + CD_x^2\right).
\end{aligned} \tag{66}$$

By (53), we have $|(h_\mu(\bar{u}_N) - h_\mu(u^*)) - (h(\bar{u}_N) - h(u^*))| \leq L\mu^2$, and so

$$\mathsf{E}\left[h(\bar{u}_N) - h(u^*)\right] \leq \mathsf{E}\left[h_\mu(\bar{u}_N) - h_\mu(u^*)\right] + L\mu^2. \tag{67}$$

Finally, combining (66) and (67) yields the desired result. □

In Theorem 4.4, η_k and the batch sizes m_k are generic. It is possible to provide one choice of the parameters so as to yield an overall simpler iteration complexity bound.

Corollary 4.5 *Under the same assumptions as in Theorem 4.4, we let $\eta_k = 1$ for all $k = 1, 2, \ldots, N$, and the batch sizes $m_k = m$ for all $k = 1, 2, \ldots, N$. Then*

$$\mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \leq \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{\mu^2 L^2 n_x^2}{m} + L\mu^2.$$

Proof. It follows from (63), with the specified parameters, that

$$\begin{aligned}
& \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \frac{1}{2N}\left(D_y^2 + \frac{1}{\gamma}D_\lambda^2 + (C+1)D_x^2\right) + \frac{\tilde{\sigma}^2}{2m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{\tilde{\sigma}^2}{2m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{2n_x\left(M^2+\sigma^2+\mu^2L^2n_x\right)}{2m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{n_x\left(M^2+\sigma^2+\mu^2L^2n_x\right)}{m} + L\mu^2 \\
=\ & \frac{D_w^2}{2N} + \frac{n_x\left(M^2+\sigma^2\right)}{m} + \frac{\mu^2L^2n_x^2}{m} + L\mu^2
\end{aligned}$$

where we denote $D_w^2 = D_y^2 + \frac{1}{\gamma}D_\lambda^2 + (C+1)D_x^2$. □

In the corollary above, the complexity bound is dependent on the sample size m, and the smoothing parameter μ. We shall further choose m and μ to obtain an explicit iteration bound.

Corollary 4.6 *Under the same assumptions as in Theorem 4.4 and Corollary 4.5, we have:*

(a) Given a fixed number of iteration N, if the smoothing parameter is chosen to be $\mu \leq \sqrt{\frac{1}{N}}$, and the number of calls to $\mathcal{SZO}$ at each iteration is $m = N$, then we have

$$\mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \leq \frac{1}{N}\left(\frac{D_w^2}{2} + n_x(M^2+\sigma^2) + L\right) + \frac{L^2n_x^2}{N^2}.$$

(b) Given a fixed number of calls to $\mathcal{SZO}$ to be $\bar{N}$, if the smoothing parameter is chosen to be $\mu \leq \sqrt{\frac{1}{N}}$, and the number of calls to the $\mathcal{SZO}$ at each iteration is

$$m = \left\lfloor \min\left\{ \max\left\{ \frac{\sqrt{n_x(M^2+\delta^2)\bar{N}}}{\tilde{D}}, \frac{n_x L}{\tilde{D}} \right\}, \bar{N} \right\} \right\rfloor,$$

for some $\tilde{D} > 0$. Then, $N = \lfloor \frac{\bar{N}}{m} \rfloor$ and

$$\mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \leq \frac{L}{\bar{N}} + \frac{n_x L}{\bar{N}}\left(\tilde{D}\theta_2 + \frac{D_w^2}{\tilde{D}}\right) + \frac{\sqrt{n_x(M^2+\delta^2)}}{\sqrt{\bar{N}}}\left(\tilde{D}\theta_1 + \frac{D_w^2}{\tilde{D}}\right)$$

where

$$\theta_1 = \max\left\{1, \frac{\sqrt{n_x(M^2+\delta^2)}}{\tilde{D}\sqrt{\bar{N}}}\right\} \text{ and } \theta_2 = \max\left\{1, \frac{n_x L}{\tilde{D}\bar{N}}\right\}. \tag{68}$$

Proof. Part (a). Since we have $m = N$, $\mu \leq \sqrt{\frac{1}{N}}$

$$\begin{aligned}
& \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{\mu^2 L^2 n_x^2}{m} + L\mu^2 \\
\leq\ & \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{N} + \frac{L^2 n_x^2}{N^2} + \frac{L}{N} \\
=\ & \frac{1}{N}\left(\frac{D_w^2}{2} + n_x(M^2+\sigma^2) + L\right) + \frac{L^2 n_x^2}{N^2}.
\end{aligned}$$

Part (b). The total number of $\mathcal{SZO}$ calls is now fixed to be $\bar{N}$. Under the assumption that at each iteration m times of $\mathcal{SZO}$ are called, we have $\bar{N}/2m \leq N \leq \bar{N}/m$, and so

$$\begin{aligned}
& \mathsf{E}_{\Omega_N}\left[h(\bar{u}_N) - h(u^*) + \rho\|A\bar{x}_N + B\bar{y}_N - b\|\right] \\
\leq\ & \frac{D_w^2}{2N} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{\mu^2 L^2 n_x^2}{m} + L\mu^2 \\
\leq\ & \frac{D_w^2 m}{\bar{N}} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{L^2 n_x^2}{m\bar{N}} + \frac{L}{\bar{N}} \\
\leq\ & \frac{D_w^2 m}{\bar{N}} + \frac{n_x(M^2+\sigma^2)}{m} + \frac{L^2 n_x^2}{m\bar{N}} + \frac{L}{\bar{N}}.
\end{aligned} \tag{69}$$

Now noting the definitions of θ_1, θ_2 in (68), we equivalently write m as

$$m = \left\lfloor \max\left\{ \frac{\sqrt{n_x(M^2+\delta^2)\bar{N}}}{\tilde{D}\theta_1}, \frac{n_x L}{\tilde{D}\theta_2} \right\} \right\rfloor.$$

Finally,

$$\begin{aligned}
\text{RHS of (69)} &\leq \frac{D_w^2\left(\frac{\sqrt{n_x(M^2+\delta^2)\bar{N}}}{\tilde{D}\theta_1}+\frac{n_xL}{\tilde{D}\theta_2}\right)}{\bar{N}}+\frac{\sqrt{n_x(M^2+\sigma^2)}\tilde{D}\theta_1}{\sqrt{\bar{N}}}+\frac{n_xL\tilde{D}\theta_2}{\bar{N}}+\frac{L}{\bar{N}} \\
&\leq \frac{D_w^2}{\tilde{D}}\frac{\sqrt{n_x(M^2+\delta^2)}}{\sqrt{\bar{N}}}+\frac{D_w^2}{\tilde{D}}\frac{n_xL}{\bar{N}}+\frac{\sqrt{n_x(M^2+\sigma^2)}\tilde{D}\theta_1}{\sqrt{\bar{N}}}+\frac{n_xL\tilde{D}\theta_2}{\bar{N}}+\frac{L}{\bar{N}} \\
&= \frac{L}{\bar{N}}+\frac{n_xL}{\bar{N}}\left(\tilde{D}\theta_2+\frac{D_w^2}{\tilde{D}}\right)+\frac{\sqrt{n_x\left(M^2+\delta^2\right)}}{\sqrt{\bar{N}}}\left(\tilde{D}\theta_1+\frac{D_w^2}{\tilde{D}}\right). \qquad (70)
\end{aligned}$$

□

Remark that the complexity bound of $O(1/N)$ in Part (a) of Corollary 4.6 is in terms of the iteration N. However, in the zeroth-order GADM algorithm we need to call $\mathcal{SZO}$ multiple times at each iteration. The complexity in terms of the total number of calls to $\mathcal{SZO}$ in Part (b) of Corollary 4.6 is denoted as $\bar{N}$, and this gives us a bound on the accuracy of $O(1/\sqrt{\bar{N}})$.

5 Numerical Experiments

In this section, we test the performance of our SGADM (GADM) algorithm by solving two test problems: large-scale convex quadratic program and the fused logistic regression. Specifically, we use GADM, i.e. iteration scheme (5), to solve convex quadratic program, and apply SGADM to the fused logistic regression. More details of those two experiments will be presented in the following subsections separately.

5.1 Convex Quadratic Problem (QP)

The convex quadratic program considered in this subsection is given by

$$\begin{array}{ll}\min\limits_{x\in\mathbf{R}^n} & \frac{1}{2}x^\top Qx+p^\top x \\ \text{s.t.} & Ax=b \\ & x\geq 0,\end{array} \qquad (71)$$

where $Q\in\mathbf{R}^{n\times n}$ is positive semidefinite, and $A\in\mathbf{R}^{m\times n}$, $b\in\mathbf{R}^m$. To fit (71) into the framework, we reformulate it as

$$\begin{array}{ll}\min\limits_{x,y\in\mathbf{R}^n} & \frac{1}{2}x^\top Qx+p^\top x \\ \text{s.t.} & Ax=b \\ & x-y=0 \\ & y\geq 0.\end{array} \qquad (72)$$

Indeed, (72) is a special case of (1), with $f(x) = \frac{1}{2}x^\top Qx + p^\top x$, $g(y) = 0$; the equality constraints $Ax - b = 0$ and $x - y = 0$, and finally $\mathcal{X} = \mathbf{R}^n$ and $\mathcal{Y} = \mathbf{R}^n_+$. In this case, the augmented Lagrangian function $\mathcal{L}_\gamma(x, y, \lambda, \mu)$ can be specified as

$$\mathcal{L}_\gamma(x, y, \lambda, \mu) = \frac{1}{2}x^\top Qx + p^\top x - \lambda^\top(Ax - b) - \mu^\top(x - y) + \frac{\gamma}{2}\|Ax - b\|^2 + \frac{\gamma}{2}\|x - y\|^2. \tag{73}$$

Notice that $\mathcal{L}_\gamma(x, y, \lambda, \mu)$ is **MinE** with respect to both x and y, thus the standard ADMM scheme (3) should work. However, to minimize the convex quadratic function involves inverting a matrix which is not exactly a first-order operation. Instead, we shall apply the GADM iteration scheme (5). In this particular case, we choose the predetermined matrix H in SGADM to be βI, and the iterative process runs as follows:

$$\begin{cases} y^{k+1} = \frac{1}{\gamma+\beta}\left[-\mu^k + \gamma x^k + \beta y^k\right]_+ \\ x^{k+1} = x^k - \alpha_k\left(Qx^k + p - A^\top\lambda^k - \mu^k + \gamma A^\top(Ax^k - b) + \gamma(x^k - y^{k+1})\right) \\ \lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} - b) \\ \mu^{k+1} = \mu^k - \gamma(x^{k+1} - y^{k+1}). \end{cases} \tag{74}$$

For this convex quadratic program, we perform some preliminary experiments where we set dimension n to be 50 and 100. Recall we also shown how to choose stepsize for deterministic problem in previous section, we set the α_k to be $1/C$, where C is a constant that is predetermined by problem itself. As a result, we will report the objective value $\frac{1}{2}\hat{x}^\top Q\hat{x} + p^\top\hat{x}$ where $\hat{x}$ solution given by (74). Since the size of the problem is small, this allows us to compare our solution with the solution obtained from CVX. The test results can be found in Table 2, where "GADM" represents the objective value we discussed above with "time" being the CPU time (in seconds) of GADM, and "CVX" represents the objective value returned by CVX. The results suggest that our algorithm indeed returns with compatible good solutions in fairly quick computational time. For the size $n = 2000$, CVX stops working on our PC while the GADM still returns a solution in time roughly proportional to the size.

Inst.	time	GADM	CVX
Dimension $n = 50$			
1	0.33	2.916e+003	2.928e+003
2	0.20	3.082e+003	3.118e+003
3	1.21	2.388e+003	2.399e+003
4	0.80	3.226e+003	3.242e+003
5	0.38	2.851e+003	2.862e+003
Dimension $n = 100$			
1	2.86	2.145e+004	2.147e+004
2	3.57	1.759e+004	1.761e+004
3	1.54	2.005e+004	2.006e+004
4	1.10	1.610e+004	1.614e+004
5	1.64	2.235e+004	2.238e+004

Table 2: GADM for Convex QP

5.2 Fused Logistic Regression

In this subsection, we show how to use SGADM to solve the fused logistic regression problem in a stochastic setting. As suggested in [24], fused logistic regression, which incorporates a certain ordering information, is derived from the fused lasso problem and sparse logistic regression. Specifically, the sparse logistic regression problem (see [23]) is given by:

$$\min_{x\in\mathbf{R}^n, c\in\mathbf{R}} \quad l(x,c)+\beta\|x\|_1, \tag{75}$$

where $l(x,c)=\frac{1}{m}\sum\limits_{i=1}^{m}\log(1+\exp(-b_i(a_i^\top x+c)))$, and $\{(a_i,b_i),\ i=1,\ldots,m\}$ is a given training set with m samples $a_1,a_2,\ldots,a_m$ and $b_i\in\{\pm 1\}, i=1,\ldots,m$ as the binary class labels. Combining requirements from the fused lasso [36] and the sparse logistic regression (75), the fused logistic regression that incorporates certain existed natural ordering features can be formulated as:

$$\min_{x\in\mathbf{R}^n, c\in\mathbf{R}} \quad l(x,c)+\beta\|x\|_1+\rho\sum_{j=2}^{n}|x_j-x_{j-1}|. \tag{76}$$

Note that the function $l(x,c)$ in the above formula involves the sample points $\{a_i,b_i\}, i=1,\ldots,m$. Without specifying samples, the corresponding stochastic version is

$$\min_{x\in\mathbf{R}^n, c\in\mathbf{R}} \quad \mathsf{E}_{u,v}[\log(1+\exp(-v(u^\top x+c)))]+\beta\|x\|_1+\rho\sum_{j=2}^{n}|x_j-x_{j-1}|. \tag{77}$$

Similar to the above generalization, the stochastic version of the lasso problem was studied in [29]. Moreover, if we denote $f(x,c)=\mathsf{E}_{u,v}[\log(1+\exp(-v(u^\top x+c)))]$, by adding two additional variables y,z, problem (77) can be rewritten as follows, for which our SGADM algorithm can be applied:

$$\begin{array}{ll}\min\limits_{x\in\mathbf{R}^n, z\in\mathbf{R}^{n-1}, y\in\mathbf{R}^n, c\in\mathbf{R}} & f(y,c)+\beta\|x\|_1+\rho\|z\|_1\\ \text{s.t.} & x=y\\ & z=My,\end{array} \tag{78}$$

where M is a $(n-1)\times n$ dimensional matrix with all ones in the diagonal and negative ones in the super-diagonal and zeros elsewhere. The augmented Lagrangian function of (78) is

$$\begin{aligned}&\mathcal{L}_\gamma(x,z,y,c,\lambda_1,\lambda_2)\\ =\ & f(y,c)+\beta\|x\|_1+\rho\|z\|_1-\lambda_1^\top(x-y)-\lambda_2^\top(z-My)+\frac{\gamma}{2}\|x-y\|^2+\frac{\gamma}{2}\|z-My\|^2.\end{aligned}$$

Based on the definition of $f(y,c)$, we can easily define the $\mathcal{SFO}$s: $G_1(y,u,v)$ and $G_2(c,u,v)$ of $f(y,c)$ as

$$\begin{cases} G_1(y,u,v):=\nabla_y\log(1+\exp(-v(u^\top x+c)))=-(1-d)vu,\\ G_2(c,u,v):=\nabla_c\log(1+\exp(-v(u^\top x+c)))=-(1-d)v,\end{cases} \tag{79}$$

where u, v are the underlying random variables, and $d = 1/(1+\exp(-v(u^\top x + c)))$. Consequently, the SGADM iteration scheme of (78) can be specified as

$$\begin{cases} (x^{k+1}, z^{k+1}) = \arg\min_{x,z} \mathcal{L}_\gamma(x, z, y^k, c^k, \lambda_1^k, \lambda_2^k) \\ d^k = 1/(1+\exp(-v^k((u^k)^\top x + c))) \\ y^{k+1} = y^k - \alpha_k(-(1-d^k)v^k(u^k) + \lambda_1^k + M^\top \lambda_2^k + \gamma(y^k - x^{k+1}) + \gamma M^\top(My^k - z^{k+1})) \\ c^{k+1} = c^k - \alpha_k(-(1-d^k)v^k) \\ \lambda_1^{k+1} = \lambda_1^k - \gamma(x^{k+1} - y^{k+1}) \\ \lambda_2^{k+1} = \lambda_1^k - \gamma(z^{k+1} - My^{k+1}). \end{cases} \tag{80}$$

The first operation in (80) has closed form solutions:

$$x^{k+1} = \text{Shrink}(y^k + \lambda_1^k/\gamma, \beta/\gamma)$$

and

$$z^{k+1} = \text{Shrink}(My^k + \lambda_2^k/\gamma, \rho/\gamma),$$

where the shrinkage operator $\text{Shrink}(x, \tau)$ is defined as

$$\text{Shrink}(x, \tau) := \text{sign}(x) \circ \max\{|x| - \tau, 0\}.$$

In the tests, we assume that u and v are drawn from normal distribution $\mathcal{N}(0,1)$ and $\text{sign}(\mathcal{N}(0,1))$ respectively. Following the rule stated in Theorem 2.2, the stepsize α_k is set to be $1/(\sqrt{k+1}+C)$, where $C > 0$ is a constant.

For each instance, we assess the expected performance via 10 independent trials. In each trial, we run the algorithm 10 times, and take the average to approximate the expectation. After we obtain the solution $\hat{x}, \hat{c}$ from (80), 50 random samples of $(a^i, b^i), i = 1, 2 \ldots, 50$, are generated from $\mathcal{N}(0,1)$ and $\text{sign}(\mathcal{N}(0,1))$ respectively, and we use $\frac{1}{50}\sum_{i=1}^{50} \log(1+\exp(-b^i((a^i)^\top \hat{x} + \hat{c})))$ to approximate the true objective. Besides, we also report the $\|\hat{x}\|_0$ and $\|M\hat{x}\|_0$ values which reflect the sparsity and the ordering of the solutions. The results can be found in Table 3, where problem dimensions are chosen from 50 to 1000, "obj" represents the approximate objective value, and "time" refers to the CPU time (in seconds) of SGADM. As we can see, SGADM indeed returns sparse solutions in terms of $\|x\|_0$ and $\|Mx\|_0$ fairly quickly.

6 Appendix

6.1 Proof of Proposition 2.3

Here we will prove Proposition 2.3. Before that, we present some technical lemmas as preparation.

Lemma 6.1 *Suppose function f is smooth and its gradient is Lipschitz continuous, i.e. (2) holds, then we have*

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{L}{2}\|x - y\|^2. \tag{81}$$

As a result, for a gradient Lipschitz continuous function, we also have the following result.

Lemma 6.2 *Suppose function f is smooth and convex, and its gradient is Lipschitz continuous with the constant L, i.e. (2) holds, then we have*

$$(x - y)^\top \nabla f(z) \leq f(x) - f(y) + \frac{L}{2}\|z - y\|^2. \tag{82}$$

Proof. Since the f is convex, we have

$$\begin{aligned} & (x - y)^\top \nabla f(z) = (x - z)^\top \nabla f(z) + (z - y)^\top \nabla f(z) \\ \leq \;& f(x) - f(z) - (y - z)^\top \nabla f(z). \end{aligned} \tag{83}$$

Based on (81), we have

$$f(y) - f(z) - \frac{L}{2}\|z - y\|^2 \leq (y - z)^\top \nabla f(z). \tag{84}$$

Combining (83) and (84), we have

$$\begin{aligned} (x - y)^\top \nabla f(z) &\leq f(x) - f(z) - (y - z)^\top \nabla f(z) \\ &\leq f(x) - f(z) - (f(y) - f(z) - \frac{L}{2}\|z - y\|^2) \\ &= f(x) - f(y) + \frac{L}{2}\|z - y\|^2. \end{aligned} \tag{85}$$

□

Proof of (20) in Proposition 2.3

Proof. First, by the optimality condition of the two subproblems in SGADM, we have

$$(y - y^{k+1})^\top \left(\partial g(y^{k+1}) - B^\top \left(\lambda^k - \gamma(Ax^k + By^{k+1} - b) \right) - H(y^k - y^{k+1}) \right) \geq 0, \quad \forall y \in \mathcal{Y},$$

and

$$(x-x^{k+1})^\top\left(x^{k+1}-\left(x^k-\alpha_k\left(G(x^k,\xi^{k+1})-A^\top(\lambda^k-\gamma(Ax^k+By^{k+1}-b))\right)\right)\right)\geq 0,\quad \forall x\in\mathcal{X},$$

where $\partial g(y)$ is a subgradient of g at y. Using $\tilde{\lambda}^k=\lambda^k-\gamma(Ax^k+By^{k+1}-b)$ and the definition of $\tilde{w}^k$ in (14), the above two inequalities are equivalent to

$$(y-\tilde{y}^k)^\top\left(\partial g(\tilde{y}^k)-B^\top\tilde{\lambda}^k-H(y^k-y^{k+1})\right)\geq 0,\quad \forall y\in\mathcal{Y}, \tag{86}$$

and

$$(x-\tilde{x}^k)^\top\left(\alpha_k\left(G(x^k,\xi^{k+1})-A^\top\tilde{\lambda}^k\right)-(x^k-\tilde{x}^k)\right)\geq 0,\quad \forall x\in\mathcal{X}. \tag{87}$$

Moreover,

$$(A\tilde{x}^k+B\tilde{y}^k-b)-\left(-A(x^k-\tilde{x}^k)+\frac{1}{\gamma}\left(\lambda^k-\tilde{\lambda}^k\right)\right)=0.$$

Thus

$$(\lambda-\tilde{\lambda}^k)^\top(A\tilde{x}^k+B\tilde{y}^k-b)=\left(\lambda-\tilde{\lambda}^k\right)^\top\left(-A\left(x^k-\tilde{x}^k\right)+\frac{1}{\gamma}\left(\lambda^k-\tilde{\lambda}^k\right)\right). \tag{88}$$

By the convexity of $g(y)$ and (86),

$$g(y)-g(\tilde{y}^k)+(y-\tilde{y}^k)^\top\left(-B^\top\tilde{\lambda}^k\right)\geq(y-\tilde{y}^k)^\top H(y^k-\tilde{y}^k),\quad \forall y\in\mathcal{Y}. \tag{89}$$

Since $\delta_{k+1}=G(x^k,\xi^{k+1})-\nabla f(x^k)$, and by (87) we have

$$(x-\tilde{x}^k)^\top\left(\alpha_k(\nabla f(x^k)-A^\top\tilde{\lambda}^k)+\alpha_k\delta_{k+1}-(x^k-\tilde{x}^k)\right)\geq 0,\quad \forall x\in\mathcal{X}$$

which leads to

$$(x-\tilde{x}^k)^\top\left(\alpha_k(\nabla f(x^k)-A^\top\tilde{\lambda}^k)\right)\geq(x-\tilde{x}^k)^\top\left(x^k-\tilde{x}^k\right)-\alpha_k\left(x-\tilde{x}^k\right)^\top\delta_{k+1},\quad \forall x\in\mathcal{X}.$$

Using (82), the above further leads to

$$\begin{aligned}&\alpha_k(f(x)-f(\tilde{x}^k))+(x-\tilde{x}^k)^\top(-\alpha_kA^\top\tilde{\lambda}^k)\\ \geq\ &(x-\tilde{x}^k)^\top\left(x^k-\tilde{x}^k\right)-\alpha_k\left(x-\tilde{x}^k\right)^\top\delta_{k+1}-\frac{\alpha_kL}{2}\|x^k-\tilde{x}^k\|^2,\quad \forall x\in\mathcal{X}.\end{aligned} \tag{90}$$

Furthermore,

$$\begin{aligned}(x-\tilde{x}^k)^\top\delta_{k+1}&=(x-x^k)^\top\delta_{k+1}+(x^k-\tilde{x}^k)^\top\delta_{k+1}\\ &\leq(x-x^k)^\top\delta_{k+1}+\frac{\eta_k}{2}\|x^k-\tilde{x}^k\|^2+\frac{\|\delta_{k+1}\|^2}{2\eta_k}.\end{aligned} \tag{91}$$

Substituting (91) in (90), and dividing both sides by α_k, we get

$$\begin{aligned}&f(x)-f(\tilde{x}^k)+(x-\tilde{x}^k)^\top(-A^\top\tilde{\lambda}^k)\\ \geq\ &\frac{(x-\tilde{x}^k)^\top(x^k-\tilde{x}^k)}{\alpha_k}-(x-x^k)^\top\delta_{k+1}-\frac{\|\delta_{k+1}\|^2}{2\eta_k}-\frac{\eta_k+L}{2}\|x^k-\tilde{x}^k\|^2.\end{aligned} \tag{92}$$

Finally, (20) follows by summing (92), (89), and (88). □

Now we show the second statement in Proposition 2.3.

Proof of (21) in Proposition 2.3

Proof. First, by (15), we have $P(w^k - \tilde{w}^k) = (w^k - w^{k+1})$, and so

$$(w - \tilde{w}^k)^\top Q_k(w^k - \tilde{w}^k) = (w - \tilde{w}^k)^\top M_k P(w^k - \tilde{w}^k) = (w - \tilde{w}^k)^\top M_k(w^k - w^{k+1}).$$

Applying the identity

$$(a-b)^\top M_k(c-d) = \frac{1}{2}\left(\|a-d\|_{M_k}^2 - \|a-c\|_{M_k}^2\right) + \frac{1}{2}\left(\|c-b\|_{M_k}^2 - \|d-b\|_{M_k}^2\right)$$

to the term $(w - \tilde{w}^k)^\top M(w^k - w^{k+1})$, we obtain

$$\begin{aligned} & (w - \tilde{w}^k)^\top M_k(w^k - w^{k+1}) \\ = \; & \frac{1}{2}\left(\|w - w^{k+1}\|_{M_k}^2 - \|w - w^k\|_{M_k}^2\right) + \frac{1}{2}\left(\|w^k - \tilde{w}^k\|_{M_k}^2 - \|w^{k+1} - \tilde{w}^k\|_{M_k}^2\right). \end{aligned} \tag{93}$$

Using (15) again, we have

$$\begin{aligned} & \|w^k - \tilde{w}^k\|_{M_k}^2 - \|w^{k+1} - \tilde{w}^k\|_{M_k}^2 \\ = \; & \|w^k - \tilde{w}^k\|_{M_k}^2 - \|(w^k - \tilde{w}^k) - (w^k - w^{k+1})\|_{M_k}^2 \\ = \; & \|w^k - \tilde{w}^k\|_{M_k}^2 - \|(w^k - \tilde{w}^k) - P(w^k - \tilde{w}^k)\|_{M_k}^2 \\ = \; & (w^k - \tilde{w}^k)^\top (2M_kP - P^\top M_k P)(w^k - \tilde{w}^k). \end{aligned} \tag{94}$$

Note that $Q_k = M_kP$ and the definition of those matrices (see (13)), we have

$$2M_kP - P^\top M_k P = 2Q_k - P^\top Q_k = \begin{pmatrix} H & 0 & 0 \\ 0 & \frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A & A^\top \\ 0 & -A & \frac{1}{\gamma} I_m \end{pmatrix}.$$

As a result,

$$\begin{aligned} & (w^k - \tilde{w}^k)^\top (2M_kP - P^\top M_k P)(w^k - \tilde{w}^k) \\ = \; & \|y^k - \tilde{y}^k\|_H^2 + \frac{1}{\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 + (x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k) \\ \geq \; & (x^k - \tilde{x}^k)^\top \left(\frac{1}{\alpha_k} I_{n_x} - \gamma A^\top A\right)(x^k - \tilde{x}^k). \end{aligned} \tag{95}$$

Combining (95), (94), and (93), the desired inequality (21) follows. □

6.2 Proof of Proposition 3.1

We first show the first part of Proposition 3.1.

Proof of (40) in Proposition 3.1

Proof. by the optimality condition of the two subproblems in SGALM, we have

$$(y-y^{k+1})^\top\left(y^{k+1}-y^k+\beta_k\left(S_g(y^k,\zeta^{k+1})-B^\top(\lambda^k-\gamma(Ax^k+By^k-b))\right)\right)\geq 0,\quad \forall y\in\mathcal{Y},$$

and also

$$(x-x^{k+1})^\top\left(x^{k+1}-x^k+\alpha_k\left(S_f(x^k,\xi^{k+1})-A^\top(\lambda^k-\gamma(Ax^k+By^{k+1}-b))\right)\right)\geq 0,\quad \forall x\in\mathcal{X}.$$

Using $\tilde{\lambda}^k=\lambda^k-\gamma(Ax^k+By^{k+1}-b)$ and the definition of $\tilde{w}^k$, the above two inequalities are equivalent to

$$(y-\tilde{y}^k)^\top\left(\beta_k\left(S_g(y^k,\zeta^{k+1})-B^\top\tilde{\lambda}^k\right)-(I_{n_y}-\beta_k\gamma B^\top B)(y^k-\tilde{y}^k)\right)\geq 0,\quad \forall y\in\mathcal{Y}, \tag{96}$$

and

$$(x-\tilde{x}^k)^\top\left(\alpha_k(S_f(x^k,\xi^{k+1})-A^\top\tilde{\lambda}^k)-(x^k-\tilde{x}^k)\right)\geq 0,\quad \forall x\in\mathcal{X}. \tag{97}$$

Also,

$$(\lambda-\tilde{\lambda}^k)^\top(A\tilde{x}^k+B\tilde{y}^k-b)=(\lambda-\tilde{\lambda}^k)^\top\left(-A(x^k-\tilde{x}^k)+\frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\right). \tag{98}$$

Since $\delta_{k+1}^f=S_f(x^k,\xi^{k+1})-\nabla f(x^k)$ and using (97), similar to (90) and (91) we have

$$\begin{aligned}&f(x)-f(\tilde{x}^k)+(x-\tilde{x}^k)^\top(-A^\top\tilde{\lambda}^k)\\ \geq\ &\frac{(x-\tilde{x}^k)^\top(x^k-\tilde{x}^k)}{\alpha_k}-(x-x^k)^\top\delta_{k+1}^f-\frac{\|\delta_{k+1}^f\|^2}{2\eta_k}-\frac{\eta_k+L}{2}\|x^k-\tilde{x}^k\|^2.\end{aligned} \tag{99}$$

Similarly, since $\delta_{k+1}^g=S_g(y^k,\zeta^{k+1})-\nabla g(y^k)$ and using (96), we also have

$$\begin{aligned}&g(y)-g(\tilde{y}^k)+(y-\tilde{y}^k)^\top(-B^\top\tilde{\lambda}^k)\\ \geq\ &(y-\tilde{y}^k)^\top\left(\frac{1}{\beta_k}I_{n_y}-\gamma B^\top B\right)(y^k-\tilde{y}^k)\\ &-(y-y^k)^\top\delta_{k+1}^g-\frac{\|\delta_{k+1}^g\|^2}{2\eta_k}-\frac{\eta_k+L}{2}\|y^k-\tilde{y}^k\|^2.\end{aligned} \tag{100}$$

Finally, (40) follows by summing (100), (99), and (98). □

Notice that $\hat{Q}_k=\hat{M}_kP$ and

$$2M_kP-P^\top M_kP=\begin{pmatrix}H_k & 0 & 0\\ 0 & \frac{1}{\alpha_k}I_{n_x}-\gamma A^\top A & A^\top\\ 0 & -A & \frac{1}{\gamma}I_m\end{pmatrix}=\begin{pmatrix}\frac{1}{\alpha_k}I_{n_x}-\gamma B^\top B & 0 & 0\\ 0 & \frac{1}{\alpha_k}I_{n_x}-\gamma A^\top A & A^\top\\ 0 & -A & \frac{1}{\gamma}I_m\end{pmatrix}.$$

Inequality (41) in Proposition 3.1 follows similarly as the derivation of (21) in Proposition 2.3.

6.3 Properties of the Smoothing Function

In this subsection, we will prove Lemma 4.1. Before that, we need some technical preparations which are summarized in the following lemma.

Lemma 6.3 *Let $\alpha(n)$ be the volume of the unit ball in $\mathbf{R}^n$, and $\beta(n)$ be the surface area of the unit sphere in $\mathbf{R}^n$. We also denote B, and S_p, to be the unit ball and unit sphere respectively.*

(a) If M_p is defined as $M_p = \frac{1}{\alpha(n)} \int_{v \in B} \|v\|^p dv$, we have

$$M_p = \frac{n}{n+p}. \tag{101}$$

(b) Let I be the identity matrix in $\mathbf{R}^{n \times n}$, then

$$\int_{S_p} vv^\top dv = \frac{\beta(n)}{n} I. \tag{102}$$

Proof. For (a), we can directly compute M_p by using the polar coordinates,

$$M_p = \frac{1}{\alpha(n)} \int_B \|v\|^p dv = \frac{1}{\alpha(n)} \int_0^1 \int_{S_p} r^p r^{n-1} dr d\theta = \frac{1}{n+p} \frac{\beta(n)}{\alpha(n)} = \frac{n}{n+p}.$$

For (b), Let $V = vv^\top$, then we know that $V_{ij} = v_i v_j$. Therefore, if $i \neq j$, by the symmetry of the unit sphere S_p (i.e. if $v \in S_p, v = (v_1, v_2, \ldots, v_n)$, then $w \in S_p$ for all $w = (\pm v_1, \pm v_2, \ldots, \pm v_n)$), we have

$$\int_{S_p} V_{ij} dv = \int_{S_p} v_i v_j dv = \int_{S_p} -v_i v_j dv = \int_{S_p} -V_{ij} dv.$$

Thus, we obtain $\int_{S_p} V_{ij} dv = 0$.

If $i = j$, we know that $V_{ii} = v_i^2$. Since we already know that

$$\int_{S_p} (v_1^2 + v_2^2 + \cdots + v_n^2) dv = \int_{S_p} \|v\|^2 dv = \beta(n).$$

Then, by symmetry, we have

$$\int_{S_p} v_1^2 dv = \int_{S_p} v_2^2 dv = \cdots = \int_{S_p} v_n^2 dv = \frac{\beta(n)}{n}.$$

Thus we also have $\int_{S_p} V_{ii}^2 dv = \frac{\beta(n)}{n}$, for $i = 1, 2, \ldots, n$. Therefore, $\int_{S_p} vv^\top dv = \frac{\beta(n)}{n} I$. □

By the next three propositions, the part (b) of Lemma 4.1 is shown; for part (a) and (c) of Lemma 4.1, the proof can be found in [35].

Proposition 6.4 *If $f \in C_L^1(\mathbf{R}^n)$, then*

$$|f_\mu(x) - f(x)| \leq \frac{L\mu^2}{2}. \tag{103}$$

Proof. Since $f \in C_L^1(\mathbf{R}^n)$, we have

$$\begin{aligned}
|f_\mu(x) - f(x)| &= \left|\frac{1}{\alpha(n)}\int_B f(x+\mu v)dv - f(x)\right| \\
&= \left|\frac{1}{\alpha(n)}\int_B (f(x+\mu v) - f(x) - \nabla f(x)^\top \mu v)dv\right| \\
&\leq \int_B \left|(f(x+\mu v) - f(x) - \nabla f(x)^\top \mu v)\right| dv \\
&\leq \int_B \frac{L\mu^2}{2}\|v\|^2 dv \\
&\overset{(101)}{=} \frac{L\mu^2}{2}\frac{n}{n+2} \leq \frac{L\mu^2}{2}.
\end{aligned}$$

□

Proposition 6.5 *If $f \in C_L^1(\mathbf{R}^n)$, then*

$$\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu n L}{2}. \tag{104}$$

Proof.

$$\begin{aligned}
&\|\nabla f_\mu(x) - \nabla f(x)\| \\
= &\left\|\frac{1}{\beta(n)}\left[\frac{n}{\mu}\int_{S_p} f(x+\mu v)v dv\right] - \nabla f(x)\right\| \\
\overset{(102)}{=} &\left\|\frac{1}{\beta(n)}\left[\frac{n}{\mu}\int_{S_p} f(x+\mu v)v dv - \int_{S_p}\frac{n}{\mu}f(x)v dv - \int_{S_p}\frac{n}{\mu}\langle\nabla f(x), \mu v\rangle v dv\right]\right\| \\
\leq &\frac{n}{\beta(n)\mu}\int_{S_p} |f(x+\mu v) - f(x) - \langle\nabla f(x), \mu v\rangle|\|v\| dv \\
\leq &\frac{n}{\beta(n)\mu}\frac{L\mu^2}{2}\int_{S_p}\|v\|^3 dv = \frac{\mu n L}{2}.
\end{aligned}$$

□

Proposition 6.6 *If $f \in C_L^1(\mathbf{R}^n)$, and the $\mathcal{SZO}$ defined as $g_\mu(x) = \frac{n}{\mu}[f(x+\mu v) - f(x)]v$, then we have*

$$\mathsf{E}_v\left[\|g_\mu(x)\|^2\right] \leq 2n\|\nabla f(x)\|^2 + \frac{\mu^2}{2}L^2 n^2. \tag{105}$$

Proof.

$$\begin{aligned}
\mathsf{E}_v[\|g_\mu(x)\|^2] &= \frac{1}{\beta(n)}\int_{S_p}\frac{n^2}{\mu^2}|f(x+\mu v)-f(x)|^2\|v\|^2 dv \\
&= \frac{n^2}{\beta(n)\mu^2}\int_{S_p}\left[f(x+\mu v)-f(x)-\langle\nabla f(x),\mu v\rangle+\langle\nabla f(x),\mu v\rangle\right]^2 dv \\
&\leq \frac{n^2}{\beta(n)\mu^2}\int_{S_p}\left[2\left(f(x+\mu v)-f(x)-\langle\nabla f(x),\mu v\rangle\right)^2+2\left(\langle\nabla f(x),\mu v\rangle\right)^2\right] dv \\
&\leq \frac{n^2}{\beta(n)\mu^2}\left[\int_{S_p}2\left(\frac{L\mu^2}{2}\|v\|^2\right)^2 dv+2\mu^2\int_{S_p}\nabla f(x)^\top vv^\top\nabla f(x)dv\right] \\
&\overset{(102)}{=} \frac{n^2}{\beta(n)\mu^2}\left[\frac{L^2\mu^4}{2}\beta(n)+2\mu^2\frac{\beta(n)}{n}\|\nabla f(x)\|^2\right] \\
&= 2n\|\nabla f(x)\|^2+\frac{\mu^2}{2}L^2n^2.
\end{aligned}$$

□

References

[1] D. Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM Journal on Optimization*, 23(4):2183–2207, 2013. 2

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. 2

[3] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Optimization Online*, 2013. 2

[4] W. Deng, M. Lai, and W. Yin. On the $o(1/k)$ convergence and parallelization of the alternating direction method of multipliers. *arXiv preprint arXiv:1312.3040*, 2013. 2

[5] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. 2012. 2

[6] J. Douglas and H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956. 2

[7] J. Eckstein and D. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992. 2

[8] Y. Ermoliev. Stochastic quasigradient methods and their application to system optimization. *An International Journal of Probability and Stochastic Processes*, 9(1-2):1–36, 1983. 3

[9] A. Gaivoronskii. Nonstationary stochastic programming problems. *Cybernetics and Systems Analysis*, 14(4):575–579, 1978. 3

[10] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012. 3

[11] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *arXiv preprint arXiv:1310.3787*, 2013. 3

[12] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013. 3

[13] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. 3

[14] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *arXiv preprint arXiv:1308.6594*, 2013. 3

[15] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*, volume 9. SIAM, 1989. 2

[16] B. He, L. Hou, and X. Yuan. On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming. 2013. 2

[17] B. He, M. Tao, and X. Yuan. Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming. *Math. Oper. Res., under revision*, 2012. 2

[18] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012. 2, 7

[19] M. Hong and Z. Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012. 2

[20] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012. 3, 11

[21] T. Lin, S. Ma, and S. Zhang. On the convergence rate of multi-block ADMM. *Optimization Online, 2014-08-4503*, 2014. 2

[22] T. Lin, S. Ma, and S. Zhang. On the global linear convergence of the ADMM with multi-block variables. *arXiv preprint arXiv:1408.4266*, 2014. 2

[23] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009. 24

[24] S. Ma and S. Zhang. An extragradient-based alternating direction method for convex minimization. *arXiv preprint arXiv:1301.6308*, 2013. 3, 24

[25] R. Monteiro and B. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *Optimization-online preprint*, 2713:1, 2010. 2

[26] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. 3

[27] A. Nemirovski and D. Yudin. Problem complexity and method efficiency in optimization. 1983. 3

[28] Y. Nesterov et al. Random gradient-free minimization of convex functions. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011. 4, 15, 16, 17

[29] H. Ouyang, N. He, L. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, pages 80–88, 2013. 3, 24

[30] B. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990. 3

[31] B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. 3

[32] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 3

[33] A. Ruszczyński and W. Syski. A method of aggregate stochastic subgradients with on-line stepsize rules for convex stochastic programming problems. In *Stochastic Programming 84 Part II*, pages 113–131. Springer, 1986. 3

[34] J. Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, pages 373–405, 1958. 3

[35] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011. 16, 30

[36] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005. 24

Trial Num.	time	$\|x\|_0$	$\|Mx\|_0$	obj
Dimension $n = 50$				
1	0.93	4	5	7.117e-001
2	0.91	5	6	7.143e-001
3	0.94	5	7	7.123e-001
4	0.91	4	6	7.118e-001
5	0.90	5	5	7.149e-001
6	0.87	4	5	7.097e-001
7	0.90	5	5	7.116e-001
8	0.90	4	5	7.128e-001
9	0.90	4	6	7.128e-001
10	0.93	4	5	7.107e-001
Dimension $n = 100$				
1	1.43	15	19	7.025e-001
2	1.34	15	18	7.058e-001
3	1.39	14	18	7.036e-001
4	1.25	14	18	7.072e-001
5	1.52	16	19	7.051e-001
6	1.63	15	19	7.035e-001
7	2.46	13	16	7.066e-001
8	1.52	15	18	7.048e-001
9	1.51	14	18	7.049e-001
10	1.53	16	21	7.052e-001
Dimension $n = 200$				
1	1.96	28	33	6.676e-001
2	1.96	30	35	6.677e-001
3	1.86	33	39	6.705e-001
4	1.91	28	34	6.666e-001
5	2.09	30	36	6.696e-001
6	4.91	30	35	6.715e-001
7	2.31	29	35	6.662e-001
8	1.84	26	31	6.688e-001
9	1.85	29	35	6.666e-001
10	2.39	26	32	6.673e-001
Dimension $n = 500$				
1	10.60	33	39	8.251e-001
2	9.29	33	38	8.316e-001
3	9.17	34	41	8.181e-001
4	10.00	32	38	8.385e-001
5	9.58	29	34	8.259e-001
6	10.09	31	37	8.247e-001
7	12.05	37	41	8.252e-001
8	10.41	33	37	8.282e-001
9	13.24	31	35	8.260e-001
10	11.00	32	39	8.341e-001
Dimension $n = 1000$				
1	62.85	50	57	8.034e-001
2	63.75	53	59	8.010e-001
3	64.20	48	53	8.268e-001
4	64.70	54	63	8.239e-001
5	73.68	54	60	8.129e-001
6	66.00	52	56	8.228e-001
7	74.10	55	61	8.126e-001
8	66.56	57	64	8.055e-001
9	65.94	55	64	8.130e-001
10	61.35	49	54	8.122e-001

Table 3: SGADM for Fused Logistic Regression