Mathematical Structures in Computer Science (2024), **34**, pp. 162–179
doi:10.1017/S096012952300004X

PAPER

Stochastic linearized generalized alternating direction method of multipliers: Expected convergence rates and large deviation properties

Jia Hu[1,2*], Tiande Guo[2] and Congying Han[2]

[1]Networked Supporting Software International S&T Cooperation Base of China, Jiangxi Normal University, Nanchang 330022, P.R. China and [2]School of Mathematical Sciences, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, P.R.China

[*]Corresponding author. Email: hujia17@mails.ucas.ac.cn

(Received 15 November 2022; revised 2 January 2023; accepted 17 January 2023; first published online 14 March 2023)

Abstract

Alternating direction method of multipliers (ADMM) receives much attention in the field of optimization and computer science, etc. The generalized ADMM (G-ADMM) proposed by Eckstein and Bertsekas incorporates an acceleration factor and is more efficient than the original ADMM. However, G-ADMM is not applicable in some models where the objective function value (or its gradient) is computationally costly or even impossible to compute. In this paper, we consider the two-block separable convex optimization problem with linear constraints, where only noisy estimations of the gradient of the objective function are accessible. Under this setting, we propose a stochastic linearized generalized ADMM (called SLG-ADMM) where two subproblems are approximated by some linearization strategies. And in theory, we analyze the expected convergence rates and large deviation properties of SLG-ADMM. In particular, we show that the worst-case expected convergence rates of SLG-ADMM are $\mathscr{O}\left(N^{-1/2}\right)$ and $\mathscr{O}\left(\ln N \cdot N^{-1}\right)$ for solving general convex and strongly convex problems, respectively, where N is the iteration number, similarly hereinafter, and with high probability, SLG-ADMM has $\mathscr{O}\left(\ln N \cdot N^{-1/2}\right)$ and $\mathscr{O}\left((\ln N)^2 \cdot N^{-1}\right)$ constraint violation bounds and objective error bounds for general convex and strongly convex problems, respectively.

Keywords: Alternating direction method of multipliers; stochastic approximation; expected convergence rate; high probability bound; convex optimization; machine learning

1. Introduction

We consider the following two-block separable convex optimization problem with linear equality constraints:

$$\min\left\{\theta_1(x)+\theta_2\left(y\right)\left|Ax+By=b, x\in\mathscr{X}, y\in\mathscr{Y}\right.\right\}, \tag{1}$$

where $A\in\mathbb{R}^{n\times n_1}, B\in\mathbb{R}^{n\times n_2}, b\in\mathbb{R}^n, \mathscr{X}\subseteq\mathbb{R}^{n_1}$, and $\mathscr{Y}\subseteq\mathbb{R}^{n_2}$ are closed convex sets, and $\theta_2:\mathbb{R}^{n_2}\to\mathbb{R}\cup\{+\infty\}$ is a convex function (not necessarily smooth). $\theta_1:\mathbb{R}^{n_1}\to\mathbb{R}$ is a convex function and is smooth on an open set containing $\mathscr{X}$, but has its specific structure; in particular, we assume that there is a stochastic first-order oracle (*SFO*) for θ_1, which returns a stochastic gradient $G(x,\xi)$ at x, where ξ is a random variable whose distribution is supported on $\Xi\subseteq\mathbb{R}^d$, satisfying

$$\text{(a) } \mathbb{E}\left[G\left(x,\xi\right)\right]=\nabla\theta_1\left(x\right), \text{ and}$$
$$\text{(b) } \mathbb{E}\left[\left\|G\left(x,\xi\right)-\nabla\theta_1\left(x\right)\right\|^2\right]\leq\sigma^2,$$

where $\sigma > 0$ is some constant. In addition, we make the following assumptions throughout the paper: (i) The solution set of (1) is assumed to be nonempty, (ii) the gradient of θ_1 is L-Lipschitz continuous for some $L > 0$, i.e., $\left\|\nabla\theta_1\left(x\right)-\nabla\theta_2\left(y\right)\right\|\leq L\left\|x-y\right\|$ for any $x, y \in \mathscr{X}$, (iii) y-subproblem has a minimizer at each iteration. As a linearly constrained convex optimization problem, though the model (1) is special, it is rich enough to characterize many optimization problems arising from various application fields, such as machine learning, image processing, and signal processing. In these fields, a typical scenario is that one of the functions represents a data fidelity term and the other function is a regularization term.

Without considering the specific structure of θ_1, i.e., the function value and gradient information is readily available, a classical method for solving problem (1) is the alternating direction method of multipliers (ADMM). ADMM was originally proposed by Glowinski and Marroco (1975), and Gabay and Mercier (1976), which is a Gauss-Seidel implementation of augmented Lagrangian method (Glowinski, 2014) or an application of Douglas-Rachford splitting method on the dual problem of (1) (Eckstein and Bertsekas, 1992). For both convex and non-convex problems, there are extensive studies on the theoretical properties of ADMM. In particular, for convex optimization problems, theoretical results on convergence behavior are abundant, whether global convergence, sublinear convergence rate, or linear convergence rate, see, e.g., Eckstein and Bertsekas (1992); He and Yuan (2012); Monteiro and Svaiter (2013); He and Yuan (2015); Deng and Yin (2016); Yang and Han (2016); Han *et al.* (2016). Recently, ADMM has been studied on non-convex models satisfying the KL inequality or other similar properties, see, e.g., Li and Pong (2015); Wang *et al.* (2019); Jiang *et al.* (2019); Zhang and Luo (2020). For a thorough understanding on some recent developments of ADMM, one can refer to a survey (Han, 2022). However, when the objective function value (or its gradient) in (1) is computationally costly or even impossible to compute, we can only access some noisy information and deterministic ADMM does not work. Such a setting is exactly what the stochastic programming (SP) model considers. In SP, the objective function is often in the form of expectation. In this case, getting the full function value or gradient information is impractical. To tackle this problem, Robbins and Monro originally introduced the stochastic approximation (SA) approach in 1951 (Robbins and Monro, 1951). Since then, SA has gone through many developments; for more detail, readers are referred to a series of works by Nemirovski, Ghadimi, and Lan, etc, see, e.g., Nemirovski *et al.* (2009); Ghadimi and Lan (2012); Lan (2012); Ghadimi and Lan (2013); Ghadimi *et al.* (2016). As for solving problem (1), motivated by the SA, some stochastic ADMM type algorithms have been proposed recently, see, e.g., Ouyang *et al.* (2013); Suzuki (2013, 2014); Zhao *et al.* (2015); Gao *et al.* (2018). Note that in these works, only the basic iterative scheme of ADMM was considered. It is well-known that incorporating an acceleration factor into the subproblem and the update on the dual variables often improves the algorithmic performance, which is the idea of generalized ADMM (Eckstein and Bertsekas, 1992; Fang *et al.*, 2015). In this paper, we study generalized ADMM in the stochastic setting. In particular, we propose a stochastic linearized generalized ADMM (SLG-ADMM) for solving two-block separable stochastic optimization problem (1) and analyze corresponding worst-case convergence rate by means of the framework of variational inequality. Moreover, we establish the large deviation properties of SLG-ADMM under certain light-tail assumptions.

The rest of this paper is organized as follows. We present the iterative scheme of SLG-ADMM and summarize some preliminaries which will be used in the theoretical analysis in Section 2. In Section 3, we analyze the worst-case convergence rate and the high probability guarantees for objective error and constraint violation for the SLG-ADMM. Finally, we make some conclusions in Section 4.

Notation 1. *For two matrices A and B, the ordering relation $A \succ B$ ($A \succeq B$) means $A - B$ is positive definite (semidefinite). I_m denotes the $m \times m$ identity matrix. For a vector x, $\|x\|$ denotes its*

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

Euclidean norm; for a matrix X, $\|X\|$ denotes its spectral norm. For any symmetric matrix G, define $\|x\|_G^2 := x^T Gx$ and $\|x\|_G := \sqrt{x^T Gx}$ if $G \succeq 0$. $\mathbb{E}[\cdot]$ denotes the mathematical expectation of a random variable. Pr $\{\cdot\}$ *denotes the probability value of an event. ∂ and ∇ denote the subdifferential and gradient operator of a function, respectively. We also sometimes use (x, y) and (x, y, λ) to denote the vectors $\left(x^T, y^T\right)^T$ and $\left(x^T, y^T, \lambda^T\right)^T$, respectively.*

2. Stochastic Linearized Generalized ADMM

In this section, we first present the iterative scheme of SLG-ADMM for solving (1), and then, we introduce some preliminaries that will be frequently used in the later analysis.

```
Algorithm 1: Stochastic Linearized Generalized ADMM (SLG-ADMM)
Initialize $x^0 \in \mathscr{X}, y^0 \in \mathscr{Y}, \lambda^0, \alpha \in (0, 2)$, and two sequences of symmetric
 and positive definite matrices: $\{G_{1,k}\}$ and $\{G_{2,k}\}$.
for $k = 0, 1, \ldots$
    Call the SFO to obtain $G\left(x^k, \xi\right)$.
    $x^{k+1} = \arg\min_{x \in \mathscr{X}} \left\{ G\left(x^k, \xi\right)^T \left(x - x^k\right) - x^T A^T \lambda^k + \frac{\beta}{2} \left\| Ax + By^k - b \right\|^2 + \frac{1}{2} \left\| x - x^k \right\|_{G_{1,k}}^2 \right\}$
    $y^{k+1} = \arg\min_{y \in \mathscr{Y}} \left\{ \theta_2(y) - y^T B^T \lambda^k + \frac{\beta}{2} \left\| \alpha A x^{k+1} + (1-\alpha)\left(b - By^k\right) + By - b \right\|^2 + \frac{1}{2} \left\| y - y^k \right\|_{G_{2,k}}^2 \right\}$
    $\lambda^{k+1} = \lambda^k - \beta \left( \alpha A x^{k+1} + (1-\alpha)\left(b - By^k\right) + By^{k+1} - b \right)$
end
```

We give some remarks on this algorithm. Algorithm 1 is an ADMM type algorithm, which alternates through one x-subproblem, one y-subproblem, and an update on the dual variables (multipliers). The algorithm is stochastic since at each iteration *SFO* is called to obtain a stochastic gradient $G\left(x^k, \xi\right)$ which is an unbiased estimation of $\nabla\theta_1\left(x^k\right)$ and is bounded relative to $\nabla\theta_1\left(x^k\right)$ in expectation. The algorithm is linearized because of the following two aspects: (i) The term $G\left(x^k, \xi\right)^T\left(x - x^k\right)$ in the x-subproblem of SLG-ADMM is a stochastic version of linearization of $\theta_1\left(x^k\right)$. (ii) x-subproblem and y-subproblem are added proximal terms $\frac{1}{2}\left\|x - x^k\right\|_{G_{1,k}}^2$ and $\frac{1}{2}\left\|y - y^k\right\|_{G_{2,k},}^2$ respectively, where $\{G_{1,k}\}$ and $\{G_{2,k}\}$ are two sequences of symmetric and positive definite matrices that can be changed with iteration; with the choice of $G_{2,k} \equiv \tau I_{n_2} - \beta B^T B, \tau > \beta\left\|B^T B\right\|$, the quadratic term in the y-subproblem is linearized. The same fact applies to the x-subproblem. Furthermore, SLG-ADMM incorporates an acceleration factor α; generally, the case with $\alpha \in (1, 2)$ could lead to better numerical results than the special case with $\alpha = 1$. When $\alpha = 1$,

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

$G_{1,k} \equiv I_{n_1}$, and the term $\frac{1}{2}\left\|y - y^k\right\|^2_{G_{2,k}}$ vanishes, SLG-ADMM reduces to the algorithm appeared in earlier literatures (Ouyang *et al.*, 2013; Gao *et al.*, 2018).

Let the Lagrangian function of the problem (1) be

$$L\left(x, y, \lambda\right) = \theta_1\left(x\right) + \theta_2\left(y\right) - \lambda^T\left(Ax + By - b\right),$$

defined on $\mathscr{X} \times \mathscr{Y} \times \mathbb{R}^n$. We call $\left(x^*, y^*, \lambda^*\right)$ a saddle point of $L\left(x, y, \lambda\right) \in \mathscr{X} \times \mathscr{Y} \times \mathbb{R}^n$ if the following inequalities are satisfied:

$$L_{\lambda\in\mathbb{R}^n}\left(x^*, y^*, \lambda\right) \le L\left(x^*, y^*, \lambda^*\right) \le L_{x\in\mathscr{X}, y\in\mathscr{Y}}\left(x, y, \lambda^*\right).$$

Obviously, a saddle point $\left(x^*, y^*, \lambda^*\right)$ can be characterized by the following inequalities

$$\begin{cases} x^* \in \mathscr{X}, L\left(x, y^*, \lambda^*\right) - L\left(x^*, y^*, \lambda^*\right) \ge 0 \;\forall x \in \mathscr{X}, \\ y^* \in \mathscr{Y}, L\left(x^*, y, \lambda^*\right) - L\left(x^*, y^*, \lambda^*\right) \ge 0 \;\forall y \in \mathscr{Y}, \\ \lambda^* \in \mathbb{R}^n, L\left(x^*, y^*, \lambda^*\right) - L\left(x^*, y^*, \lambda\right) \ge 0 \;\forall \lambda \in \mathbb{R}^n. \end{cases}.$$

Below we invoke a proposition which characterize the optimality condition of an optimization model by a variational inequality. The proof can be found in He (2017).

Proposition 2. *Let $\mathscr{X} \subset \mathbb{R}^n$ be a closed convex set and let $\theta\left(x\right) : \mathbb{R}^n \to \mathbb{R}$ and $f\left(x\right) : \mathbb{R}^n \to \mathbb{R}$ be convex functions. In addition, $f\left(x\right)$ is differentiable. Assuming that the solution set of the minimization problem* $\min\left\{\theta\left(x\right) + f\left(x\right) | x \in \mathscr{X}\right\}$ *is nonempty, then we have the assertion that*

$$x^* = \arg\min\left\{\theta\left(x\right) + f\left(x\right) | x \in \mathscr{X}\right\},$$

if and only if

$$x^* \in \mathscr{X}, \theta\left(x\right) - \theta\left(x^*\right) + \left(x - x^*\right)^T \nabla f\left(x^*\right) \ge 0 \;\forall x \in \mathscr{X}.$$

Hence with this proposition, solving (1) is equivalent to solving the following variational inequality problem under the assumption that the solution set of problem (1) is nonempty: Finding $w^* = \left(x^*, y^*, \lambda^*\right) \in \Omega := \mathscr{X} \times \mathscr{Y} \times \mathbb{R}^n$ such that

$$\theta\left(u\right) - \theta\left(u^*\right) + \left(w - w^*\right)^T F\left(w^*\right) \ge 0, \forall w \in \Omega,$$

where

$$u = \begin{pmatrix} x \\ y \end{pmatrix}, w = \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix}, F\left(w\right) = \begin{pmatrix} -A^T\lambda \\ -B^T\lambda \\ Ax + By - b \end{pmatrix}, \text{ and } \theta\left(u\right) = \theta_1\left(x\right) + \theta_2\left(y\right).$$

The variables with superscript or subscript such as u^k, w^k, $\bar{u}_k$, $\bar{w}_k$ are denoted similarly. In addition, we define two auxiliary sequences for the convergence analysis. More specifically, for the sequence $\left\{w^k\right\}$ generated by the SLG-ADMM, let

$$\tilde{w}^k = \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \\ \tilde{\lambda}^k \end{pmatrix} = \begin{pmatrix} x^{k+1} \\ y^{k+1} \\ \lambda^k - \beta\left(Ax^{k+1} + By^k - b\right) \end{pmatrix} \text{ and } \tilde{u}^k = \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \end{pmatrix}. \tag{2}$$

Based on the above notations and the update scheme of λ^k in SLG-ADMM, we have

$$\lambda^{k+1} - \tilde{\lambda}^k = (1 - \alpha)\left(\lambda^k - \tilde{\lambda}^k\right) + \beta B\left(y^k - \tilde{y}^k\right), \tag{3}$$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

and

$$\lambda^k - \lambda^{k+1} = \alpha\left(\lambda^k - \tilde{\lambda}^k\right) + \beta B\left(\tilde{y}^k - y^k\right). \tag{4}$$

Then, we get

$$w^k - w^{k+1} = M\left(w^k - \tilde{w}^k\right), \tag{5}$$

where M is defined as

$$\begin{pmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & -\beta B & \alpha I_n \end{pmatrix}. \tag{6}$$

For notational simplicity, we define two sequences of matrices that will be used later: for $k = 0, 1, \ldots$

$$H_k = \begin{pmatrix} G_{1,k} & 0 & 0 \\ 0 & \frac{\beta}{\alpha}B^TB + G_{2,k} & \frac{1-\alpha}{\alpha}B^T \\ 0 & \frac{1-\alpha}{\alpha}B & \frac{1}{\beta\alpha}I_n \end{pmatrix}, Q_k = \begin{pmatrix} G_{1,k} & 0 & 0 \\ 0 & \beta B^TB + G_{2,k} & (1-\alpha)B^T \\ 0 & -B & \frac{1}{\beta}I_n \end{pmatrix}. \tag{7}$$

Obviously, for any k, the matrices M, H_k, and Q_k satisfy $Q_k = H_kM$.

Throughout the paper, we need the following assumptions:

Assumption.

(i) The primal-dual solution set Ω^* of problem (1) is nonempty.

(ii) $\theta_1(x)$ is differentiable, and its gradient satisfies the L-Lipschitz condition

$$\|\nabla\theta_1(x_1) - \nabla\theta_1(x_2)\| \le L\|x_1 - x_2\|$$

for all $x_1, x_2 \in \mathcal{X}$.

(iii)

$$\text{a) } \mathbb{E}[G(x,\xi)] = \nabla\theta_1(x) \text{ and b) } \quad \mathbb{E}\left[\|G(x,\xi) - \nabla\theta_1(x)\|^2\right] \le \sigma^2,$$

where $\sigma > 0$ is some constant.

Under the second assumption, it holds that for all $x, y \in \mathcal{X}$,

$$\theta_1(x) \le \theta_1(y) + (x-y)^T\nabla\theta_1(y) + \frac{L}{2}\|x-y\|^2.$$

A direct result of combining this property with convexity is shown in the following lemma.

Lemma 3. *Suppose function f is convex and differentiable, and its gradient is L-Lipschitz continuous, then for any x, y, z, we have*

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

In addition, if f is μ-strongly convex, then for any x, y, z we have

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

Proof. Since the gradient of f is L-Lipschitz continuous, then for any y, z we have $f(y) \le f(z) + (y-z)^T\nabla f(z) + \frac{L}{2}\|y-z\|^2$. Also, due to the convexity of f, we have for any x, z, $f(x) \ge$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

$f(z) + (x-z)^T \nabla f(z)$. Adding the above two inequalities, we get the conclusion. If f is μ-strongly convex, then for any x, z, $f(x) \geq f(z) + (x-z)^T \nabla f(z) + \frac{\mu}{2}\|x-z\|^2$. Then, we combine this inequality with $f(y) \leq f(z) + (y-z)^T \nabla f(z) + \frac{L}{2}\|y-z\|^2$, and the proof is completed. □

3. Theoretical Analysis of SLG-ADMM

In this section, we shall establish theoretical properties of SLG-ADMM. More specifically, in Subsection 3.1, we analyze the expected convergence rates of SLG-ADMM. And, we analyze the large deviation properties of SLG-ADMM in Subsection 3.2.

3.1 Expected convergence rate

First, this subsection considers that the function θ_1 is convex. The next several lemmas are to obtain an upper bound of $\theta\left(\tilde{u}^k\right) - \theta(u) + \left(\tilde{w}^k - w\right)^T F\left(\tilde{w}^k\right)$. With such a bound, it is possible to estimate the worst-case convergence rate of SLG-ADMM.

Lemma 4. *Let the sequence $\left\{w^k\right\}$ be generated by the SLG-ADMM and the associated $\left\{\tilde{w}^k\right\}$ be defined in (2). Then, we have*

$$\begin{aligned}\theta(u) - \theta\left(\tilde{u}^k\right) + \left(w - \tilde{w}^k\right)^T F\left(\tilde{w}^k\right) \geq & \left(w - \tilde{w}^k\right)^T Q_k \left(w^k - \tilde{w}^k\right) - \left(x - \tilde{x}^k\right)^T \delta^k \\ & - \frac{L}{2}\left\|x^k - \tilde{x}^k\right\|^2, \forall w \in \Omega,\end{aligned} \tag{8}$$

where Q_k is defined in (7), and $\delta^k = G\left(x^k, \xi\right) - \nabla\theta_1\left(x^k\right)$, similarly hereinafter.

Proof. The optimality condition of the x-subproblem in SLG-ADMM is

$$\begin{aligned}&\left(x - x^{k+1}\right)^T \left(G\left(x^k, \xi\right) - A^T\lambda^k + \beta A^T\left(Ax^{k+1} + By^k - b\right) + G_{1,k}\left(x^{k+1} - x^k\right)\right) \\ &\geq 0, \forall x \in \mathscr{X}.\end{aligned} \tag{9}$$

Using $\tilde{x}^k$ and $\tilde{\lambda}^k$ defined in (2) and notation of δ^k, (9) can be rewritten as

$$\left(x - \tilde{x}^k\right)^T \left(\nabla\theta_1\left(x^k\right) + \delta^k - A^T\tilde{\lambda}^k + G_{1,k}\left(\tilde{x}^k - x^k\right)\right) \geq 0, \forall x \in \mathscr{X}. \tag{10}$$

In lemma 1, letting $y = \tilde{x}^k$, $z = x^k$, and $f = \theta_1$, we get

$$\left(x - \tilde{x}^k\right)^T \nabla\theta_1\left(x^k\right) \leq \theta_1(x) - \theta_1\left(\tilde{x}^k\right) + \frac{L}{2}\left\|x^k - \tilde{x}^k\right\|^2. \tag{11}$$

Combining (10) and (11), we obtain

$$\begin{aligned}&\theta_1(x) - \theta_1\left(\tilde{x}^k\right) + \left(x - \tilde{x}^k\right)^T\left(-A^T\tilde{\lambda}^k\right) \\ &\geq \left(x - \tilde{x}^k\right)^T G_{1,k}\left(x^k - \tilde{x}^k\right) - \left(x - \tilde{x}^k\right)^T \delta^k - \frac{L}{2}\left\|x^k - \tilde{x}^k\right\|^2.\end{aligned} \tag{12}$$

Similarly, the optimality condition of y-subproblem is

$$\theta_2(y) - \theta_2\left(\tilde{y}^k\right) + \left(y - \tilde{y}^k\right)^T\left(-B^T\lambda^{k+1} + G_{2,k}\left(\tilde{y}^k - y^k\right)\right) \geq 0, \forall y \in \mathscr{Y}. \tag{13}$$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

Substituting (3) into (13), we obtain that

$$\begin{aligned}&\theta_2\left(y\right)-\theta_2\left(\tilde{y}^k\right)+\left(y-\tilde{y}^k\right)^T\left(-B^T\tilde{\lambda}^k\right)\\ &\geq(1-\alpha)\left(y-\tilde{y}^k\right)^TB^T\left(\lambda^k-\tilde{\lambda}^k\right)+\left(y-\tilde{y}^k\right)^T\left(\beta B^TB+G_{2,k}\right)\left(y^k-\tilde{y}^k\right),\forall y\in\mathcal{Y}.\end{aligned}\tag{14}$$

At the same time,

$$\begin{aligned}\tilde{\lambda}^k&=\lambda^k-\beta\left(Ax^{k+1}+By^{k+1}-b\right)+\beta B\left(y^{k+1}-y^k\right)\\ &=\lambda^k-\beta\left(A\tilde{x}^k+B\tilde{y}^k-b\right)+\beta B\left(\tilde{y}^k-y^k\right).\end{aligned}$$

That is,

$$\left(\lambda-\tilde{\lambda}^k\right)^T\left(A\tilde{x}^k+B\tilde{y}^k-b\right)=\frac{1}{\beta}\left(\lambda-\tilde{\lambda}^k\right)^T\left(\lambda^k-\tilde{\lambda}^k\right)+\left(\lambda-\tilde{\lambda}^k\right)^TB\left(\tilde{y}^k-y^k\right).\tag{15}$$

Combining (12), (14), and (15), we get

$$\begin{aligned}&\theta\left(u\right)-\theta\left(\tilde{u}^k\right)+\begin{pmatrix}x-\tilde{x}^k\\ y-\tilde{y}^k\\ \lambda-\tilde{\lambda}^k\end{pmatrix}^T\begin{pmatrix}-A^T\tilde{\lambda}^k\\ -B^T\tilde{\lambda}^k\\ A\tilde{x}^k+B\tilde{y}^k-b\end{pmatrix}\\ \geq&\left(x-\tilde{x}^k\right)^TG_{1,k}\left(x^k-\tilde{x}^k\right)-\left(x-\tilde{x}^k\right)^T\delta^k-\frac{L}{2}\left\|x^k-\tilde{x}^k\right\|^2\\ &+(1-\alpha)\left(y-\tilde{y}^k\right)^TB^T\left(\lambda^k-\tilde{\lambda}^k\right)+\left(y-\tilde{y}^k\right)^T\left(\beta B^TB+G_{2,k}\right)\left(y^k-\tilde{y}^k\right)\\ &+\frac{1}{\beta}\left(\lambda-\tilde{\lambda}^k\right)^T\left(\lambda^k-\tilde{\lambda}^k\right)+\left(\lambda-\tilde{\lambda}^k\right)^TB\left(\tilde{y}^k-y^k\right),\forall w\in\Omega.\end{aligned}\tag{16}$$

Finally, by the definition of F and Q_k, we come to the conclusion. □

Next, we need to further explore the terms on the right-hand side of (8).

Lemma 5. *Let the sequence $\left\{w^k\right\}$ be generated by the SLG-ADMM and the associated $\left\{\tilde{w}^k\right\}$ be defined in* (2)*. Then for any $w\in\mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n$, we have*

$$\begin{aligned}&\left(w-\tilde{w}^k\right)^TQ_k\left(w^k-\tilde{w}^k\right)\\ =&\frac{1}{2}\left(\left\|w-w^{k+1}\right\|_{H_k}^2-\left\|w-w^k\right\|_{H_k}^2\right)+\left\|x^k-\tilde{x}^k\right\|_{G_{1,k}}^2+\frac{1}{2}\left\|y^k-\tilde{y}^k\right\|_{G_{2,k}}^2\\ &-\frac{\alpha-2}{2\beta}\left\|\lambda^k-\tilde{\lambda}^k\right\|^2.\end{aligned}\tag{17}$$

Proof. Using $Q_k=H_kM$ and $w^k-w^{k+1}=M\left(w^k-\tilde{w}^k\right)$ in (5), we have

$$\begin{aligned}\left(w-\tilde{w}^k\right)^TQ_k\left(w^k-\tilde{w}^k\right)=&\left(w-\tilde{w}^k\right)^TH_kM\left(w^k-\tilde{w}^k\right)\\ =&\left(w-\tilde{w}^k\right)^TH_k\left(w^k-w^{k+1}\right).\end{aligned}\tag{18}$$

Now applying the identity: for the vectors a, b, c, d and a matrix H with appropriate dimension,

$$(a-b)^TH(c-d)=\frac{1}{2}\left(\|a-d\|_H^2-\|a-c\|_H^2\right)+\frac{1}{2}\left(\|c-b\|_H^2-\|d-b\|_H^2\right).$$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

In this identity, letting $a = w$, $b = \tilde{w}^k$, $c = w^k$, $d = \tilde{w}^k$, and $H = Q_k$, we have

$$\begin{aligned}\left(w-\tilde{w}^k\right)^T H_k\left(w^k-w^{k+1}\right) = & \frac{1}{2}\left(\left\|w-w^{k+1}\right\|_{H_k}^2-\left\|w-w^k\right\|_{H_k}^2\right)\\ & +\frac{1}{2}\left(\left\|w^k-\tilde{w}^k\right\|_{H_k}^2-\left\|w^{k+1}-\tilde{w}^k\right\|_{H_k}^2\right).\end{aligned}$$

Next, we simplify the term $\left\|w^k-\tilde{w}^k\right\|_{H_k}^2-\left\|w^{k+1}-\tilde{w}^k\right\|_{H_k}^2$.

$$\begin{aligned}&\left\|w^k-\tilde{w}^k\right\|_{H_k}^2-\left\|w^{k+1}-\tilde{w}^k\right\|_{H_k}^2\\ &=\left\|w^k-\tilde{w}^k\right\|_{H_k}^2-\left\|w^{k+1}-w^k+w^k-\tilde{w}^k\right\|_{H_k}^2\\ &=\left\|w^k-\tilde{w}^k\right\|_{H_k}^2-\left\|\left(I_{n_1+n_2+n}-M\right)\left(w^k-\tilde{w}^k\right)\right\|_{H_k}^2\\ &=\left(w^k-\tilde{w}^k\right)^T\left(H_k-\left(I_{n_1+n_2+n}-M\right)^T H_k\left(I_{n_1+n_2+n}-M\right)\right)\left(w^k-\tilde{w}^k\right)\\ &=\left(w^k-\tilde{w}^k\right)^T\left(H_kM+M^TH_k-M^TH_kM\right)\left(w^k-\tilde{w}^k\right)\\ &=\left(w^k-\tilde{w}^k\right)^T\left(\left(2I_{n_1+n_2+n}-M^T\right)Q_k\right)\left(w^k-\tilde{w}^k\right),\end{aligned}$$

where the second equality uses $w^k - w^{k+1} = M\left(w^k-\tilde{w}^k\right)$ in (5), and the last equality holds since the transpose of M^TH_k is H_kM, and hence,

$$\begin{aligned}\left(w^k-\tilde{w}^k\right)^T H_kM\left(w^k-\tilde{w}^k\right) = & \left(w^k-\tilde{w}^k\right)^T M^TH_k\left(w^k-\tilde{w}^k\right)\\ = & \left(w^k-\tilde{w}^k\right)^T Q_k\left(w^k-\tilde{w}^k\right).\end{aligned}$$

The remaining task is to prove

$$\begin{aligned}&\left(w^k-\tilde{w}^k\right)^T\left(\left(2I_{n_1+n_2+n}-M^T\right)Q_k\right)\left(w^k-\tilde{w}^k\right)\\ &=\left\|x^k-\tilde{x}^k\right\|_{G_{1,k}}^2+\left\|y^k-\tilde{y}^k\right\|_{G_{2,k}}^2-\frac{\alpha-2}{\beta}\left\|\lambda^k-\tilde{\lambda}^k\right\|^2.\end{aligned}\tag{19}$$

By simple algebraic operation,

$$\left(2I_{n_1+n_2+n}-M^T\right)Q_k=\begin{pmatrix}G_{1,k} & 0 & 0\\ 0 & G_{2,k} & (2-\alpha)B^T\\ 0 & (\alpha-2)B & \frac{2-\alpha}{\beta}I_n\end{pmatrix}.$$

With this result, (19) holds and the proof is completed. □

Now, we are ready to establish the first main theorem for SLG-ADMM. In this theorem, we take $G_{1,k}$ of the form $\tau_kI_{n_1}-\beta A^TA$, $\tau_k>0$, which simplifies the system of linear equation in x-subproblem, and $G_{2,k}\equiv G_2$. Of course, G_2 can also take the similar form as $G_{1,k}$. In particular, if $G_2=\eta I_{n_2}-\beta B^TB$, $\eta\geq\beta\left\|B^TB\right\|$, then y-subproblem reduces to the proximal mapping of g.

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

Theorem. *Let the sequence $\left\{w^k\right\}$ be generated by the SLG-ADMM, the associated $\left\{\tilde{w}^k\right\}$ be defined in* (2)*, and*

$$\bar{w}_N = \frac{1}{N+1}\sum_{t=0}^{N}\tilde{w}^t,$$

for some pre-selected integer N. Choosing $\tau_k \equiv \sqrt{N}+M$, where M is a constant satisfying the ordering relation $MI_{n_1} \succeq LI_{n_1} + \beta A^T A$, then we have

$$\begin{aligned} &\theta\left(\bar{u}_N\right) - \theta\left(u\right) + \left(\bar{w}_N - w\right)^T F\left(w\right) \\ \le& \frac{1}{2\left(N+1\right)}\left\|w^0 - w\right\|_{H_0}^2 + \frac{1}{N+1}\sum_{t=0}^{N}\left(x - x^t\right)^T\delta^t + \frac{1}{N+1}\sum_{t=0}^{N}\frac{1}{2\sqrt{N}}\left\|\delta^t\right\|^2. \end{aligned} \tag{20}$$

Proof. Combining lemma 2 and lemma 3, we get

$$\begin{aligned} &\theta\left(\tilde{u}^t\right) - \theta\left(u\right) + \left(\tilde{w}^t - w\right)^T F\left(\tilde{w}^t\right) \\ \le& \frac{1}{2}\left(\left\|w^t - w\right\|_{H_t}^2 - \left\|w^{t+1} - w\right\|_{H_t}^2\right) - \frac{1}{2}\left\|x^t - \tilde{x}^t\right\|_{G_{1,t}}^2 - \frac{1}{2}\left\|y^t - \tilde{y}^t\right\|_{G_{2,t}}^2 \\ &+ \frac{\alpha-2}{2\beta}\left\|\lambda^t - \tilde{\lambda}^t\right\|^2 + \left(x - \tilde{x}^t\right)^T\delta^t + \frac{L}{2}\left\|x^t - \tilde{x}^t\right\|^2 \\ =& \frac{1}{2}\left(\left\|w^t - w\right\|_{H_t}^2 - \left\|w^{t+1} - w\right\|_{H_t}^2\right) + \left(x - x^t\right)^T\delta^t + \left(x^t - \tilde{x}^t\right)^T\delta^t \\ &+ \frac{1}{2}\left(x^t - \tilde{x}^t\right)^T\left(LI_{n_1} - G_{1,t}\right)\left(x^t - \tilde{x}^t\right) - \frac{1}{2}\left\|y^t - \tilde{y}^t\right\|_{G_{2,t}}^2 + \frac{\alpha-2}{2\beta}\left\|\lambda^t - \tilde{\lambda}^t\right\|^2 \\ \le& \frac{1}{2}\left(\left\|w^t - w\right\|_{H_t}^2 - \left\|w^{t+1} - w\right\|_{H_t}^2\right) + \left(x - x^t\right)^T\delta^t + \frac{1}{2\sqrt{N}}\left\|\delta^t\right\|^2 \\ &+ \frac{1}{2}\left(x^t - \tilde{x}^t\right)^T\left(LI_{n_1} - MI_{n_1} + \beta A^T A\right)\left(x^t - \tilde{x}^t\right) \\ \le& \frac{1}{2}\left(\left\|w^t - w\right\|_{H_t}^2 - \left\|w^{t+1} - w\right\|_{H_t}^2\right) + \left(x - x^t\right)^T\delta^t + \frac{1}{2\sqrt{N}}\left\|\delta^t\right\|^2, \end{aligned} \tag{21}$$

where the second inequality holds owing to the Young's inequality and $\alpha \in (0, 2)$. Meanwhile,

$$\begin{aligned} &\frac{1}{N+1}\sum_{t=0}^{N}\theta\left(\tilde{u}^t\right) - \theta\left(u\right) + \left(\tilde{w}^t - w\right)^T F\left(\tilde{w}^t\right) \\ =& \frac{1}{N+1}\sum_{t=0}^{N}\theta\left(\tilde{u}^t\right) - \theta\left(u\right) + \left(\tilde{w}^t - w\right)^T F\left(w\right) \\ \ge& \theta\left(\bar{u}_N\right) - \theta\left(u\right) + \left(\bar{w}_N - w\right)^T F\left(w\right), \end{aligned} \tag{22}$$

where the equality holds since for any w_1 and w_2,

$$\left(w_1 - w_2\right)^T\left(F\left(w_1\right) - F\left(w_2\right)\right) = 0,$$

and the inequality follows from the convexity of θ. Now summing both sides of (21) from 0 to N and then taking the average, and using (22), the assertion of this theorem follows directly. □

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

Corollary 6. *Let the sequence* $\left\{w^k\right\}$ *be generated by the SLG-ADMM, the associated* $\left\{\tilde{w}^k\right\}$ *be defined in* (2), *and*

$$\bar{w}_N = \frac{1}{N+1}\sum_{t=0}^{N} \tilde{w}^t,$$

for some pre-selected integer N. Choosing $\tau_k \equiv \sqrt{N} + M$*, where M is a constant satisfying the ordering relation* $MI_{n_1} \succeq LI_{n_1} + \beta A^T A$*, then we have*

$$\mathbb{E}\left[\left\|A\bar{x}_N + B\bar{y}_N - b\right\|\right] \le \frac{1}{2\left(N+1\right)}\left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\sigma^2}{2\sqrt{N}}, \tag{23}$$

and

$$\mathbb{E}\left[\theta\left(\bar{u}_N\right) - \theta\left(u^*\right)\right] \le \frac{\|\lambda^*\|+1}{2\left(N+1\right)}\left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\|\lambda^*\|+1}{2\sqrt{N}}\sigma^2, \tag{24}$$

where e is an unit vector satisfying $-e^T\left(A\bar{x}_k + B\bar{y}_k - b\right) = \left\|A\bar{x}_k + B\bar{y}_k - b\right\|$.

Proof. In (20), let $w = \left(x^*, y^*, \lambda\right)$, where $\lambda = \lambda^* + e$. Obviously, $\|e\| = 1$. Then, the left-hand side of (20) is

$$\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) - \left(\lambda^*\right)^T\left(A\bar{x}_N + B\bar{y}_N - b\right) + \left\|A\bar{x}_N + B\bar{y}_N - b\right\|. \tag{25}$$

Such a result is attributed to

$$\begin{aligned}
&\left(\bar{w}_N - w\right)^T F\left(w\right)\\
=&\left(\bar{x}_N - x^*\right)^T\left(-A^T\lambda\right) + \left(\bar{y}_N - y^*\right)^T\left(-B^T\lambda\right) + \left(\bar{\lambda}_N - \lambda\right)^T\left(Ax^* + By^* - b\right)\\
=&\lambda^T\left(Ax^* + By^* - b\right) - \left(\lambda^T\left(A\bar{x}_N + B\bar{y}_N - b\right)\right)\\
=&-\left(\lambda^*\right)^T\left(A\bar{x}_N + B\bar{y}_N - b\right) + \left\|A\bar{x}_N + B\bar{y}_N - b\right\|,
\end{aligned}$$

where the first equality follows from the definition of F, and the second and last equalities hold due to $Ax^* + By^* - b = 0$ and the choice of λ. On the other hand, substituting $w = \bar{w}_N$ into the variational inequality associated with (1), we get

$$\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) - \left(\lambda^*\right)^T\left(A\bar{x}_N + B\bar{y}_N - b\right) \ge 0. \tag{26}$$

Combining (25) and (26), we obtain that the left-hand side of (20) is no less than $\left\|A\bar{x}_N + B\bar{y}_N - b\right\|$ when letting $w = \left(x^*, y^*, \lambda^* + e\right)$. Hence,

$$\mathbb{E}\left[\left\|A\bar{x}_N + B\bar{y}_N - b\right\|\right] \le \frac{1}{2\left(N+1\right)}\left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\sigma^2}{2\sqrt{N}}, \tag{27}$$

where in the first inequality we use $\mathbb{E}\left[\delta^k\right] = 0$ and $\mathbb{E}\left[\left\|\delta^k\right\|^2\right] \le \sigma^2$. The first part of this corollary is proved. Next, we prove the second part. Substituting $w = \bar{w}_N$ into the variational inequality associated with (1), we get

$$\begin{aligned}
&\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) + \left(\bar{w}_N - w^*\right)^T F\left(w^*\right)\\
=&\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) - \left(\lambda^*\right)^T\left(A\bar{x}_N + B\bar{y}_N - b\right)\\
\ge&\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) - \left\|\lambda^*\right\|\left\|A\bar{x}_N + B\bar{y}_N - b\right\|,
\end{aligned}$$

i.e.,

$$\theta\left(\bar{u}_N\right)-\theta\left(u^*\right)\leq\theta\left(\bar{u}_N\right)-\theta\left(u^*\right)+\left(\bar{w}_N-w^*\right)^TF\left(w^*\right)+\left\|\lambda^*\right\|\left\|A\bar{x}_N+B\bar{y}_N-b\right\|. \tag{28}$$

Take expectation on both sides of (28), and hence, (24) is proved. □

Remark 7. (i) In the above theorem or corollary, N needs to be selected in advance, and hence, τ_ks are constant. In fact, τ_k can also vary with the number of iterations. In the case of $\tau_k=\sqrt{k}+M$, if the distance between w^k and w^* is bounded, i.e., $\left\|w^k-w^*\right\|^2\leq R^2$ for any k, we can also obtain a worst-case convergence rate. The main difference with that proof idea in the above theorem or corollary is bounding the term $\sum_{t=0}^k\left(\left\|x^t-x^*\right\|_{G_{1,t}}^2-\left\|x^{t+1}-x^*\right\|_{G_{1,t}}^2\right)$, which is now bounded as follows.

$$\begin{aligned}&\sum_{t=0}^k\left(\left\|x^t-x^*\right\|_{G_{1,t}}^2-\left\|x^{t+1}-x^*\right\|_{G_{1,t}}^2\right)\\&=M\left\|x^0-x^*\right\|^2+\sum_{i=0}^{k-1}\left(\tau_{i+1}-\tau_i\right)\left\|x^{i+1}-x^*\right\|^2-\left\|x^{k+1}-x^*\right\|_{G_{1,k}}^2\\&\leq\left(M+\sum_{i=0}^{k-1}\left(\tau_{i+1}-\tau_i\right)\right)R^2\\&=\left(M+\sqrt{k}\right)R^2.\end{aligned}$$

(ii) The above corollary reveals that the worst-case expected convergence rate of SLG-ADMM for solving general convex problems is $\mathscr{O}\left(\frac{1}{\sqrt{N}}\right)$, where N is the iteration number.

At end of this subsection, we assume that θ_1 is μ-strongly convex, i.e., $\theta_1\left(x\right)\geq\theta_1\left(y\right)+\left\langle\nabla\theta_1\left(y\right),x-y\right\rangle+\frac{\mu}{2}\left\|x-y\right\|^2,\mu>0$ for all $x,y\in\mathscr{X}$. With the strong convexity, we can obtain not only the objective function value gap and constraint violation converge to zero in expectation but also the convergence of ergodic iterates of SLG-ADMM.

Theorem. *Let the sequence* $\left\{w^k\right\}$ *be generated by the SSL-ADMM and the associated* $\left\{\tilde{w}^k\right\}$ *be defined in* (2)*, and*

$$\bar{w}_k=\frac{1}{k}\sum_{t=1}^k\tilde{w}^t.$$

Choosing $\tau_k=\mu\left(k+1\right)+M$*, where* M *is a constant satisfying the ordering relation* $MI_{n_1}\succeq LI_{n_1}+\beta A^TA$*, then SLG-ADMM has the following properties*

$$\begin{aligned}&\mathbb{E}\left[\left\|A\bar{x}_k+B\bar{y}_k-b\right\|\right]\\&\leq\frac{1}{2\left(k+1\right)}\left\|\left(y^0,\lambda^0\right)-\left(y^*,\lambda^*+e\right)\right\|_{H_{1;2\times2}}^2+\frac{1}{2\left(k+1\right)}\left\|x^0-x^*\right\|_{MI_{n_1}-\beta A^TA}^2\\&\quad+\frac{\sigma^2}{2\mu\left(k+1\right)}\left(1+\ln\left(k+1\right)\right),\end{aligned} \tag{29}$$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

and

$$\begin{aligned}&\mathbb{E}\left[\theta\left(\bar{u}_k\right)-\theta\left(u^*\right)\right]\\&\leq\frac{\|\lambda^*\|+1}{2\left(k+1\right)}\left\|\left(y^0,\lambda^0\right)-\left(y^*,\lambda^*+e\right)\right\|^2_{H_{1;2\times2}}+\frac{\|\lambda^*\|+1}{2\left(k+1\right)}\left\|x^0-x^*\right\|^2_{MI_{n_1}-\beta A^TA}\\&\quad+\frac{\sigma^2\left(\|\lambda^*\|+1\right)}{2\mu\left(k+1\right)}\left(1+\ln\left(k+1\right)\right),\end{aligned}\tag{30}$$

where e is an unit vector satisfying $-e^T\left(A\bar{x}_k+B\bar{y}_k-b\right)=\left\|A\bar{x}_k+B\bar{y}_k-b\right\|$, *and*

$$H_{1;2\times2}=\begin{pmatrix}\frac{\beta}{\alpha}\beta B^TB+G_2 & \frac{1-\alpha}{\alpha}B^T\\ \frac{1-\alpha}{\alpha}B & \frac{1}{\beta\alpha}I_n\end{pmatrix}.$$

Proof. First, similar to the proof of lemma 2, using the μ-strong convexity of f, we conclude that for any $w\in\Omega$

$$\begin{aligned}&\theta\left(u\right)-\theta\left(\tilde{u}^k\right)+\left(w-\tilde{w}^k\right)^TF\left(\tilde{w}^k\right)\\&\geq\left(w-\tilde{w}^k\right)^TQ_k\left(w^k-\tilde{w}^k\right)-\left(x-\tilde{x}^k\right)^T\delta^k-\frac{L}{2}\left\|x^k-\tilde{x}^k\right\|^2+\frac{\mu}{2}\left\|x-x^k\right\|^2,\end{aligned}\tag{31}$$

where Q_k is defined in (7). Then using the result in lemma 3,

$$\begin{aligned}&\left(w-\tilde{w}^k\right)^TQ_k\left(w^k-\tilde{w}^k\right)\\&=\frac{1}{2}\left(\left\|w-w^{k+1}\right\|^2_{H_k}-\left\|w-w^k\right\|^2_{H_k}\right)+\frac{1}{2}\left\|x^k-\tilde{x}^k\right\|^2_{G_{1,k}}+\frac{1}{2}\left\|y^k-\tilde{y}^k\right\|^2_{G_2}\\&\quad-\frac{\alpha-2}{2\beta}\left\|\lambda^k-\tilde{\lambda}^k\right\|^2.\end{aligned}\tag{32}$$

Combining (31) and (32), we get

$$\begin{aligned}&\theta\left(\tilde{u}^t\right)-\theta\left(u\right)+\left(\tilde{w}^t-w\right)^TF\left(\tilde{w}^t\right)\\&\leq\frac{1}{2}\left(\left\|w^t-w\right\|^2_{H_t}-\left\|w^{t+1}-w\right\|^2_{H_t}-\mu\left\|x^t-x\right\|^2\right)+\left(x-x^t\right)^T\delta^t+\frac{1}{2\mu\left(t+1\right)}\left\|\delta^t\right\|^2.\end{aligned}\tag{33}$$

Now using (22) and (33), we have

$$\begin{aligned}&\theta\left(\bar{u}_k\right)-\theta\left(u\right)+\left(\bar{w}_k-w\right)^TF\left(w\right)\\&\leq\frac{1}{k+1}\sum_{t=0}^{k}\theta\left(\tilde{u}^t\right)-\theta\left(u\right)+\left(\tilde{w}^t-w\right)^TF\left(\tilde{w}^t\right)\\&\leq\frac{1}{2\left(k+1\right)}\sum_{t=0}^{k}\left(\left\|x^t-x\right\|^2_{MI_{n_1}-\beta A^TA}-\left\|x^{t+1}-x\right\|^2_{MI_{n_1}-\beta A^TA}\right)\\&\quad+\frac{1}{2\left(k+1\right)}\sum_{t=0}^{k}\left\|\left(y^t,\lambda^t\right)-\left(y,\lambda\right)\right\|^2_{H_{1;2\times2}}-\left\|\left(y^{t+1},\lambda^{t+1}\right)-\left(y,\lambda\right)\right\|^2_{H_{1;2\times2}}\end{aligned}$$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

$$+\frac{1}{k+1}\sum_{t=0}^{k}\left(x-x^t\right)^T\delta^t+\frac{1}{2\mu\left(k+1\right)}\sum_{t=0}^{k}\frac{1}{t+1}\left\|\delta^t\right\|^2$$

$$\leq\frac{1}{2\left(k+1\right)}\left\|\left(y^0,\lambda^0\right)-\left(y,\lambda\right)\right\|_{H_{1;2\times 2}}^2+\frac{1}{k+1}\sum_{t=0}^{k}\left(x-x^t\right)^T\delta^t+\frac{1}{2\mu\left(k+1\right)}\sum_{t=0}^{k}\frac{1}{t+1}\left\|\delta^t\right\|^2$$

$$+\frac{1}{2\left(k+1\right)}\left\|x^0-x\right\|_{MI_{n_1}-\beta A^TA}^2. \tag{34}$$

Finally, taking expectation on both sides of (32) and following the proof for getting (27) and (28), we obtain

$$\mathbb{E}\left[\left\|A\bar{x}_k+B\bar{y}_k-b\right\|\right]$$

$$\leq\frac{1}{2\left(k+1\right)}\left(\left\|x^0-x^*\right\|_{MI_{n_1}-\beta A^TA}^2+\left\|\left(y^0,\lambda^0\right)-\left(y^*,\lambda^*+e\right)\right\|_{H_{1;2\times 2}}^2\right)+\frac{\sigma^2}{2\mu\left(k+1\right)}\left(1+\ln\left(k+1\right)\right)$$

and

$$\theta\left(\bar{u}_k\right)-\theta\left(u^*\right)\leq\theta\left(\bar{u}_k\right)-\theta\left(u^*\right)+\left(\bar{w}_k-w^*\right)^TF\left(w^*\right)+\left\|\lambda^*\right\|\left\|A\bar{x}_k+\bar{y}_k-b\right\|.$$

Therefore, (29) and (30) are proved. □

This theorem implies that under the assumption that θ_1 is strongly convex, the worst-case expected convergence rate for the SLG-ADMM can be improved to $\mathcal{O}\left(\ln k/k\right)$ with the choice of diminishing size. The following theorem shows the convergence of ergodic iterates of SLG-ADMM, which is not covered in some earlier literatures (Ouyang *et al.*, 2013; Gao *et al.*, 2018). Furthermore, if θ_2 is also strongly convex, the assumption that B is full column rank can be removed.

Theorem. *Let the sequence $\left\{w^k\right\}$ be generated by the SLG-ADMM, the associated $\left\{\tilde{w}^k\right\}$ be defined in (2), and*

$$\bar{w}_k=\frac{1}{k}\sum_{t=1}^{k}\tilde{w}^t.$$

Choosing $\tau_k=\mu\left(k+1\right)+M$, where M is a constant satisfying the ordering relation $MI_{n_1}\succeq LI_{n_1}+\beta A^TA$, and assuming B is full column rank and $\lambda_{\min}$ denotes the minimum eigenvalue of B^TB, then we have

$$\mathbb{E}\left[\left\|\bar{x}_k-x^*\right\|+\left\|\bar{y}_k-y^*\right\|\right]$$

$$\leq\left(1+\frac{\|A\|}{\sqrt{\lambda_{\min}}}\right)\sqrt{\left[\frac{2}{\mu}\left(\mathbb{E}\left[\theta\left(\bar{u}_k\right)-\theta\left(u^*\right)\right]+\left\|\lambda^*\right\|\mathbb{E}\left[\left\|A\bar{x}_k+B\bar{y}_k-b\right\|\right]\right)\right]} \tag{35}$$

$$+\frac{1}{\sqrt{\lambda_{\min}}}\mathbb{E}\left\|A\bar{x}_k+B\bar{y}_k-b\right\|,$$

where the bounds for $\mathbb{E}\left[\left\|A\bar{x}_k+B\bar{y}_k-b\right\|\right]$ and $\mathbb{E}\left[\theta\left(\bar{u}_k\right)-\theta\left(u^*\right)\right]$ are the same as in (29) and (30), respectively.

Proof. Since $\left(x^*,y^*,\lambda^*\right)$ is a solution of (1), we have $A^T\lambda^*=\nabla\theta_1\left(x^*\right)$ and $B^T\lambda^*\in\partial\theta_2\left(y^*\right)$. Hence, since θ_1 is strongly convex and θ_2 is convex, we have

$$\theta_1\left(\bar{x}_k\right)\geq\theta_1\left(x^*\right)+\left(\lambda^*\right)^T\left(A\bar{x}_k-Ax^*\right)+\frac{\mu}{2}\left\|\bar{x}_k-x^*\right\|^2 \tag{36}$$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

and

$$\theta_2\left(\bar{y}_k\right) \geq \theta_2\left(y^*\right) + \left(\lambda^*\right)^T\left(B\bar{y}_k - By^*\right). \tag{37}$$

Adding up (36) and (37), we get $\theta\left(\bar{u}_k\right) \geq \theta\left(u^*\right) + \left(\lambda^*\right)^T\left(A\bar{x}_k + B\bar{y}_k - b\right) + \frac{\mu}{2}\|\bar{x}_k - x^*\|^2$, that is

$$\begin{aligned}\left\|\bar{x}_k - x^*\right\| &\leq \sqrt{\frac{2}{\mu}\left(\theta\left(\bar{u}_k\right) - \theta\left(u^*\right) - \left(\lambda^*\right)^T\left(A\bar{x}_k + B\bar{y}_k - b\right)\right)} \\ &\leq \sqrt{\frac{2}{\mu}\left(\theta\left(\bar{u}_k\right) - \theta\left(u^*\right) + \|\lambda^*\|\left\|A\bar{x}_k + B\bar{y}_k - b\right\|\right)}.\end{aligned} \tag{38}$$

On the other hand,

$$\begin{aligned}\left\|A\bar{x}_k + B\bar{y}_k - b\right\| &= \left\|A\left(\bar{x}_k - x^*\right) + B\left(\bar{y}_k - y^*\right)\right\| \\ &\geq \left\|B\left(\bar{y}_k - y^*\right)\right\| - \|A\|\left\|\bar{x}_k - x^*\right\|,\end{aligned}$$

this implies $\left\|B\left(\bar{y}_k - y^*\right)\right\| \leq \|A\|\|\bar{x}_k - x^*\| + \left\|A\bar{x}_k + B\bar{y}_k - b\right\|$, and hence,

$$\left\|\bar{y}_k - y^*\right\| \leq \frac{\|A\|}{\sqrt{\lambda_{\min}}}\left\|\bar{x}_k - x^*\right\| + \frac{1}{\sqrt{\lambda_{\min}}}\left\|A\bar{x}_k + B\bar{y}_k - b\right\|. \tag{39}$$

Adding (38) and (39), using Jensen's inequality $\mathbb{E}X^{\frac{1}{2}} \leq (\mathbb{E}X)^{\frac{1}{2}}$ for a random variable X, and taking expectation imply

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{x}_k - x^*\right\| + \left\|\bar{y}_k - y^*\right\|\right] \\ &\leq \left(1 + \frac{\|A\|}{\sqrt{\lambda_{\min}}}\right)\sqrt{\mathbb{E}\left[\frac{2}{\mu}\left(\theta\left(\bar{u}_k\right) - \theta\left(u^*\right) + \|\lambda^*\|\left\|A\bar{x}_k + B\bar{y}_k - b\right\|\right)\right]} \\ &+ \frac{1}{\sqrt{\lambda_{\min}}}\mathbb{E}\left\|A\bar{x}_k + B\bar{y}_k - b\right\|.\end{aligned}$$

The proof is completed. □

3.2 High probability performance analysis

In this subsection, we shall establish the large deviation properties of SLG-ADMM. By (23) and (24), and Markov's inequality, we have for any $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ that

$$\Pr\left\{\left\|A\bar{x}_N + B\bar{y}_N - b\right\| \leq \varepsilon_1\left(\frac{1}{2\left(N+1\right)}\left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\sigma^2}{2\sqrt{N}}\right)\right\} \geq 1 - \frac{1}{\varepsilon_1} \tag{40}$$

and

$$\Pr\left\{\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) \leq \varepsilon_2\left(\frac{\|\lambda^*\| + 1}{2\left(N+1\right)}\left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\|\lambda^*\| + 1}{2\sqrt{N}}\sigma^2\right)\right\} \geq 1 - \frac{1}{\varepsilon_2}. \tag{41}$$

However, these bounds are not strong. In the following, we will show these high probability bounds can be significantly improved when imposing standard "light-tail" assumption, see, e.g., Nemirovski *et al.* (2009); Lan (2020) . Specifically, assume that for any $x \in \mathscr{X}$

$$\mathbb{E}\left[\exp\left\{\|G\left(x, \xi\right) - \nabla\theta_1\left(x\right)\|^2/\sigma^2\right\}\right] \leq \exp\left\{1\right\}.$$

This assumption is a little bit stronger than b) in Assumption (iii), which can be explained by Jensen's inequality. For further analysis, we assume that $\mathscr{X}$ is bounded and its diameter is denoted by D_X, defined as $\max_{x_1, x_2 \in \mathscr{X}}\|x_1 - x_2\|$. The following theorem shows the high probability bound for objective error and constraint violation of SLG-ADMM.

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

Theorem. *Let the sequence* $\left\{w^k\right\}$ *be generated by the SLG-ADMM, the associated* $\left\{\tilde{w}^k\right\}$ *be defined in* (2)*, and*

$$\bar{w}_N = \frac{1}{N+1}\sum_{t=0}^{N} \tilde{w}^t$$

for some pre-selected integer N*. Choosing* $\tau_k \equiv \sqrt{N} + M$*, where* M *is a constant satisfying the ordering relation* $MI_{n_1} \succeq LI_{n_1} + \beta A^T A$*, then SLG-ADMM has the following properties*

(i)

$$\Pr\left\{\left\|A\bar{x}_N + B\bar{y}_N - b\right\| \le \frac{1}{2\left(N+1\right)}\left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} + \frac{1}{2\sqrt{N}}\left(1+\Theta\right)\sigma^2\right\} \ge 1 - \exp\left\{-\Theta^2/3\right\} - \exp\left\{-\Theta\right\}, \tag{42}$$

(ii)

$$\Pr\left\{\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) \le \left(\left\|\lambda^*\right\| + 1\right)\left(\frac{1}{2\left(N+1\right)}\left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} + \frac{1}{2\sqrt{N}}\left(1+\Theta\right)\sigma^2\right)\right\} \ge 1 - \exp\left\{-\Theta^2/3\right\} - \exp\left\{-\Theta\right\}, \tag{43}$$

where e is an unit vector satisfying $-e^T\left(A\bar{x}_N + B\bar{y}_N - b\right) = \left\|A\bar{x}_N + B\bar{y}_N - b\right\|$.

Proof. Let $\zeta^t = \frac{1}{N}\left(x^* - x^t\right)^T \delta^t$. Clearly, $\left\{\zeta^t\right\}_{t\ge 1}$ is a martingale-difference sequence. Moreover, it follows from the definition of D_X and that light-tail assumption that

$$\mathbb{E}\left[\exp\left\{\left(\zeta^t\right)^2 / \left(\frac{1}{N}D_X\sigma\right)^2\right\}\right] \le \mathbb{E}\left[\exp\left\{\left(\frac{1}{N}D_X\left\|\delta^t\right\|\right)^2 / \left(\frac{1}{N}D_X\sigma\right)^2\right\}\right] \le \exp\left\{1\right\}.$$

Now using a well-known result (see Lemma 4.1 in Lan (2020)) for the martingale-difference sequence, we have for any $\Theta \ge 0$

$$\Pr\left\{\sum_{t=1}^{N}\zeta^t > \frac{\Theta D_X \sigma}{\sqrt{N}}\right\} \le \exp\left\{-\Theta^2/3\right\}. \tag{44}$$

Also, observe that by Jensen's inequality

$$\exp\left\{\frac{1}{N}\sum_{t=1}^{N}\left(\left\|\delta^t\right\|^2/\sigma^2\right)\right\} \le \frac{1}{N}\sum_{t=1}^{N}\exp\left\{\left\|\delta^t\right\|^2/\sigma^2\right\},$$

whence, taking expectation,

$$\mathbb{E}\left[\exp\left\{\frac{1}{N}\sum_{t=1}^{N}\left\|\delta^t\right\|^2/\sigma^2\right\}\right] \le \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}\left[\exp\left\{\left\|\delta^t\right\|^2/\sigma^2\right\}\right] \le \exp\left\{1\right\}.$$

It then follows from Markov's inequality that for any $\Theta \ge 0$

$$\Pr\left\{\frac{1}{N}\sum_{t=1}^{N}\left\|\delta^t\right\|^2 \ge \left(1+\Theta\right)\sigma^2\right\} \le \exp\left\{-\Theta\right\}. \tag{45}$$

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

Using (44) and (45) in (20) for $w = \left(x^*, y^*, \lambda^* + e\right)$, we conclude that

$$\Pr\left\{\left\|A\bar{x}_N + B\bar{y}_N - b\right\| > \frac{1}{2\,(N+1)} \left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} + \frac{1}{2\sqrt{N}}\,(1+\Theta)\,\sigma^2\right\} \\ \leq \exp\left\{-\Theta^2/3\right\} + \exp\left\{-\Theta\right\} \tag{46}$$

and

$$\Pr\left\{\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) > \left(\left\|\lambda^*\right\| + 1\right)\left(\frac{1}{2\,(N+1)} \left\|w^0 - \left(x^*, y^*, \lambda^* + e\right)\right\|_{H_0}^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} \right.\right. \\ \left.\left. + \frac{1}{2\sqrt{N}}\,(1+\Theta)\,\sigma^2\right)\right\} \leq \exp\left\{-\Theta^2/3\right\} + \exp\left\{-\Theta\right\}. \tag{47}$$

The result immediately follows from the above inequalities. □

Remark 8. In view of the last Theorem, if we take $\Theta = \ln N$, then we have

$$\Pr\left\{\left\|A\bar{x}_N + B\bar{y}_N - b\right\| \leq \mathcal{O}\left(\frac{\ln N}{\sqrt{N}}\right)\right\} \geq 1 - \frac{1}{N^{2/3}} - \frac{1}{N}$$

and

$$\Pr\left\{\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) \leq \mathcal{O}\left(\frac{\ln N}{\sqrt{N}}\right)\right\} \geq 1 - \frac{1}{N^{2/3}} - \frac{1}{N}.$$

For strongly convex case, using similar derivation, the high probability bound for objective error and constraint violation of SLG-ADMM is

$$\Pr\left\{\left\|A\bar{x}_N + B\bar{y}_N - b\right\| \leq \mathcal{O}\left(\frac{(\ln N)^2}{N}\right)\right\} \geq 1 - \frac{1}{N^{2/3}} - \frac{1}{N},$$

and

$$\Pr\left\{\theta\left(\bar{u}_N\right) - \theta\left(u^*\right) \leq \mathcal{O}\left(\frac{(\ln N)^2}{N}\right)\right\} \geq 1 - \frac{1}{N^{2/3}} - \frac{1}{N}.$$

Observe that the convergence rate of ergodic iterates of SLG-ADMM is obtained in (35). The high probability bound can be also established, which is shown as follows

$$\Pr\left\{\left\|\bar{x}_N - x^*\right\| + \left\|\bar{y}_N - y^*\right\| \leq \mathcal{O}\left(\frac{\ln N}{\sqrt{N}}\right)\right\} \geq 1 - \frac{1}{N^{2/3}} - \frac{1}{N},$$

where N is the iteration number. In contrast to (40) and (41), we can observe that the results in the last theorem are much finer.

4. Conclusion

In this paper, we analyze the expected convergence rates and the large deviation properties of a stochastic variant of generalized ADMM using the variational inequality framework. By means of this framework, the proof is very clear. When the model is deterministic and *SFO* is not needed, our proposed algorithm reduces to a generalized proximal ADMM, and the convergence region of α is the same as that in the corresponding literature.

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

References

Deng, W. and Yin, W. (2016). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing* **66** (3) 889–916.

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

Fang, E. X., He, B., Liu, H. and Yuan, X. (2015). Generalized alternating direction method of multipliers: new theoretical insights and applications. *Mathematical Programming Computation* **7** (2) 149–187.

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

Gao, X., Jiang, B. and Zhang, S. (2018). On the information-adaptive variants of the ADMM: an iteration complexity perspective. *Journal of Scientific Computing* **76** (1) 327–363.

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

Ghadimi, S. and Lan, G. (2012). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming* **156** (1) 59–99.

Ghadimi, S., Lan, G. and Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming* **155** (1) 267–305.

Glowinski, R. (2014). On alternating direction methods of multipliers: a historical perspective. In: *Modeling, Simulation and Optimization for Science and Technology*, pp 59–82. Dordrecht: Springer.

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue Française d'automatique, Informatique, Recherche Opérationnelle. Analyse Numérique* **9** (R2) 41–76.

Han, D. R. (2022). A survey on some recent developments of alternating direction method of multipliers. *Journal of the Operations Research Society of China* **10** (1) 1–52.

Han, D., Sun, D. and Zhang, L. (2018). Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Mathematics of Operations Research* **43** (2) 622–637.

He, B. S. (2017). On the convergence properties of alternating direction method of multipliers. *Numerical Mathematics, a Journal of Chinese Universities(Chinese Series)* **39** 81–96.

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

He, B. and Yuan, X. (2015). On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Numerische Mathematik* **130** (3) 567–577.

Jiang, B., Lin, T., Ma, S. and Zhang, S. (2019). Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications* **72** (1) 115–157.

Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming* **133** (1) 365–397.

Lan, G. (2020). First-order and stochastic optimization methods for machine learning. New York: Springer.

Li, G. and Pong, T. K. (2015). Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization* **25** (4) 2434–2460.

Monteiro, R. D. C. and Svaiter, B. F. (2013). Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization* **23** (1) 475–507.

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

Ouyang, H., He, N., Tran, L. and Gray, A. (2013). Stochastic alternating direction method of multipliers. In: *Proceedings of the 30th International Conference on Machine Learning*, pp. 80–88. Atlanta: PMLR.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics* **22** (3) 400–407.

Suzuki, T. (2013). Dual averaging and proximal gradient descent for online alternating direction multiplier method. In: *Proceedings of the 30th International Conference on Machine Learning*, pp. 392–400. Atlanta: PMLR.

Suzuki, T. (2014). Stochastic dual coordinate ascent with alternating direction method of multipliers. In: *Proceedings of the 31th International Conference on Machine Learning*, pp. 736–744. Beijing: PMLR.

Wang, Y., Yin, W. and Zeng, J. (2019). Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing* **78** (1) 29–63.

Yang, W. H. and Han, D. (2016). Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. *SIAM Journal on Numerical Analysis* **54** (2) 625–640.

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press

Zhang, J., Luo, Z. Q. (2020). A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization* **30** (3) 2272–2302.

Zhao, P., Yang, J., Zhang, T. and Li, P. (2015). Adaptive stochastic alternating direction method of multipliers. In: *Proceedings of the 32th International Conference on Machine Learning*, pp. 69–77. Lille: PMLR.

Cite this article: Hu J, Guo T and Han C (2024). Stochastic linearized generalized alternating direction method of multipliers: Expected convergence rates and large deviation properties. *Mathematical Structures in Computer Science* **34**, 162–179. https://doi.org/10.1017/S096012952300004X

https://doi.org/10.1017/S096012952300004X Published online by Cambridge University Press