

A Symmetric Linearized Alternating Direction Method of Multipliers for a Class of Stochastic Optimization Problems

Jia HU*

*Networked Supporting Software International S&T Cooperation Base of China, Jiangxi Normal
University, Nanchang 330022, China
E-mail: hujia17@mailsucas.ac.cn*

Qimin HU

*Networked Supporting Software International S&T Cooperation Base of China, Jiangxi Normal
University, Nanchang 330022, China
E-mail: qiminhhu@163.com*

Abstract Alternating direction method of multipliers (ADMM) receives much attention in the recent years due to various demands from machine learning and big data related optimization. In 2013, Ouyang et al. extend the ADMM to the stochastic setting for solving some stochastic optimization problems, inspired by the structural risk minimization principle. In this paper, we consider a stochastic variant of symmetric ADMM, named symmetric stochastic linearized ADMM (SSL-ADMM). In particular, using the framework of variational inequality, we analyze the convergence properties of SSL-ADMM. Moreover, we show that, with high probability, SSL-ADMM has $O((\ln N) \cdot N^{-1/2})$ constraint violation bound and objective error bound for convex problems, and has $O((\ln N)^2 \cdot N^{-1})$ constraint violation bound and objective error bound for strongly convex problems, where N is the iteration number. Symmetric ADMM can improve the algorithmic performance compared to classical ADMM, numerical experiments for statistical machine learning show that such an improvement is also present in the stochastic setting.

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

1 Introduction

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

$$\min \{ \theta_1(x) + \theta_2(y) \mid Ax + By = b, x \in \mathcal{X}, y \in \mathcal{Y} \}, \quad (1)$$

where $A \in \mathbb{R}^{n \times n_1}$, $B \in \mathbb{R}^{n \times n_2}$, $b \in \mathbb{R}^n$, $\mathcal{X} \subseteq \mathbb{R}^{n_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{n_2}$ are closed convex sets, and $\theta_1 : \mathbb{R}^{n_1} \rightarrow \mathbb{R}$ and $\theta_2 : \mathbb{R}^{n_2} \rightarrow \mathbb{R}$ are convex functions (not necessarily smooth), while θ_1 has its specific structure. In particular, we assume that there is a stochastic first-order oracle (\mathcal{SFO})

Received October 22, 2022, accepted January 11, 2023

Supported by National Natural Science Foundation of China (61662036)

*The corresponding author

for θ_1 , which returns an unbiased and bounded stochastic gradient $G(x, \xi)$ at x , where ξ is a random variable whose distribution is supported on $\Xi \subseteq \mathbb{R}^d$. Such an assumption is common in the stochastic programming (SP), see, e.g., [1–3] and the references therein. In SP, the objective function is often in the form of expectation, i.e., $\theta_1(x) = \int_{\Xi} \Theta(x, \xi) dP(\xi)$ for some Θ, P , and Ξ , including finite sum as a special case. For both cases (number of terms in the summation is large for the latter case), getting the full function value or gradient information is impractical. Motivated by this, we need to design some stochastic approximation^[4] based algorithms to solve problem (1).

For the problem (1) itself, as a linearly constrained convex optimization problem, 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 where one of the functions represents some data fidelity term, and the other is a regularization term, see, e.g., [5] and the references therein. Without considering the specific structure, i.e., the assumption of \mathcal{SFO} is not needed in the model, a classical method for solving problem (1) is the alternating direction method of multipliers (ADMM). ADMM was originally proposed by Glowinski and Marrocco^[6], and Gabay and Mercier^[7], which is a Gauss-Seidel implementation of augmented Lagrangian method^[8] or an application of Douglas-Rachford splitting method on the dual problem of (1)^[9]. 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., [9–15]. Recently, ADMM has been studied on nonconvex models satisfying the well-known Kurdyka-Lojasiewicz (KL) inequality or other similar properties, see, e.g., [16–19]. For a thorough understanding on some recent developments of ADMM, one can refer to a survey^[20].

However, as we mentioned before, the gradient information of θ_1 in (1) must be obtained by the \mathcal{SFO} due to some computational or other limitation, and hence aforementioned ADMM does not work. To tackle this problem, some stochastic ADMM type algorithms have been proposed recently, see, e.g., [21–24]. Note that in these works, only the basic iterative scheme of ADMM was considered. It is well-known that symmetrically updating the dual variable in a more flexible way often improves the algorithmic performance, which is the idea of symmetric ADMM (or Peaceman-Rachford splitting method applied to the dual of problem (1)), see, e.g., [25–28]. In this paper, we study symmetric ADMM in the stochastic setting. In particular, we propose a symmetric stochastic linearized ADMM (SSL-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 SSL-ADMM under certain light-tail assumptions. Also, numerical experiments on the graph-guided fused lasso problem demonstrate the promising performance compared to non-symmetric ADMM.

The rest of this paper is organized as follows. We introduce some fundamental preliminaries in Section 2. Convergence properties of the proposed algorithm are analyzed in Section 3. The high probability guarantees for objective error and constraint violation of the proposed algorithm are investigated in Section 4. In Section 5, numerical results are presented to indicate

the promising efficiency of symmetrically updating dual variables in the stochastic setting. Finally, a summary is made in Section 6.

Notations 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 Euclidean norm; for a matrix X , $\|X\|$ denotes its spectral norm. For any symmetric matrix G , define $\|x\|_G^2 := x^T G x$ and $\|x\|_G := \sqrt{x^T G x}$ 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 $(x^T, y^T)^T$ and $(x^T, y^T, \lambda^T)^T$, respectively.

2 Preliminaries

In this section, we summarize some preliminaries that will be used in later analysis. Let the Lagrangian function of the problem (1) be

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

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

$$L_{\lambda \in \mathbb{R}^n}(x^*, y^*, \lambda) \leq L(x^*, y^*, \lambda^*) \leq L_{x \in \mathcal{X}, y \in \mathcal{Y}}(x, y, \lambda^*).$$

Obviously, a saddle point (x^*, y^*, λ^*) can be characterized by the following inequalities

$$\begin{cases} x^* \in \mathcal{X}, L(x, y^*, \lambda^*) - L(x^*, y^*, \lambda^*) \geq 0, & \forall x \in \mathcal{X}, \\ y^* \in \mathcal{Y}, L(x^*, y, \lambda^*) - L(x^*, y^*, \lambda^*) \geq 0, & \forall y \in \mathcal{Y}, \\ \lambda^* \in \mathbb{R}^n, L(x^*, y^*, \lambda^*) - L(x^*, y^*, \lambda) \geq 0, & \forall \lambda \in \mathbb{R}^n. \end{cases}$$

Below we invoke two propositions, one of which characterizes the optimality condition of an optimization model by a variational inequality and the other gives a result for the martingale-difference sequence.

Proposition 1 *Let $\mathcal{X} \subset \mathbb{R}^n$ be a closed convex set and let $\theta(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ and $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex functions. In addition, $f(x)$ is differentiable. Assuming that the solution set of the minimization problem $\min\{\theta(x) + f(x) \mid x \in \mathcal{X}\}$ is nonempty, then we have the assertion that*

$$x^* = \arg \min\{\theta(x) + f(x) \mid x \in \mathcal{X}\}$$

if and only if

$$x^* \in \mathcal{X}, \quad \theta(x) - \theta(x^*) + (x - x^*)^T \nabla f(x^*) \geq 0, \quad \forall x \in \mathcal{X}.$$

Proof The proof can be found in [31]. ■

Proposition 2 *Let $\xi_{[t]} \equiv \{\xi_1, \xi_2, \dots, \xi_t\}$ be a sequence of independent identically distributed random variables, and $\zeta_t = \zeta_t(\xi_{[t]})$ be deterministic Borel functions of $\xi_{[t]}$ such that $\mathbb{E}_{|\xi_{[t-1]}}[\zeta_t] = 0$ almost surely and $\mathbb{E}_{|\xi_{[t-1]}}[\exp\{\zeta_t^2/\sigma_t^2\}] \leq \exp\{1\}$ almost surely, where $\sigma_t > 0$*

are deterministic and $\mathbb{E}_{|Y}[X]$ denotes the expectation of random variable X conditional on random variable Y . Then

$$\forall \lambda \geq 0 : \text{Prob} \left\{ \sum_{t=1}^N \zeta_t > \lambda \sqrt{\sum_{t=1}^N \sigma_t^2} \right\} \leq \exp \{-\lambda^2/3\}.$$

Proof The proof can be founded in Lemma 4.1 on page 116–117 of [32]. ■

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

$$\theta(u) - \theta(u^*) + (w - w^*)^T F(w^*) \geq 0, \quad \forall w \in \Omega,$$

where

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

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 $\{w^k\}$ generated by the SSL-ADMM in Section 3, let

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

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

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

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

- (iii)

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

where $\sigma > 0$ is some constant.

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

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

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

Lemma 1 Suppose function f is convex and differentiable, and its gradient is L -Lipschitz continuous, then for any x, y, z we have

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

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

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

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

3 Symmetric Stochastic Linearized ADMM

In this section, we will present and analyze iterative scheme of the proposed symmetric stochastic linearized ADMM, named SSL-ADMM.

Algorithm 1: Symmetric Stochastic Linearized ADMM (SSL-ADMM)

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}, \lambda^0, \beta, (r, s) \in \mathcal{D}$, two sequences of symmetric and positive semidefinite matrices: $\{G_{1,k}\}$ and $\{G_{2,k}\}$, where

$$\mathcal{D} = \{(r, s) \mid r + s > 0, r \leq 1, -r^2 - s^2 - rs + r + s + 1 \geq 0\},$$

for $k = 0, 1, \dots$.

Call the \mathcal{SCO} to obtain $G(x^k, \xi)$;

$$x^{k+1} = \arg \min_{x \in \mathcal{X}} \left\{ G(x^k, \xi)^T (x - x^k) - x^T A^T \lambda^k + \frac{\beta}{2} \|Ax + By^k - b\|^2 + \frac{1}{2} \|x - x^k\|_{G_{1,k}}^2 \right\};$$

$$\lambda^{k+\frac{1}{2}} = \lambda^k - r\beta (Ax^{k+1} + By^k - b);$$

$$y^{k+1} = \arg \min_{y \in \mathcal{Y}} \left\{ \theta_2(y) - y^T B^T \lambda^{k+\frac{1}{2}} + \frac{\beta}{2} \|Ax^{k+1} + By - b\|^2 + \frac{1}{2} \|y - y^k\|_{G_{2,k}}^2 \right\};$$

$$\lambda^{k+1} = \lambda^{k+\frac{1}{2}} - s\beta (Ax^{k+1} + By^{k+1} - b).$$

end

Algorithm 1 is called SSL-ADMM for short. We give some remarks on this algorithm. SSL-ADMM is a ADMM type algorithm, which alternates through one x -subproblem, an update on the multipliers, one y -subproblem, and an update on the multipliers again. The algorithm is symmetric since the dual variable is symmetrically updated twice at each iteration.

The algorithm is stochastic since at each iteration \mathcal{SFO} is called to obtain a stochastic gradient $G(x^k, \xi)$ which is an unbiased estimation of $g(x^k)$, the gradient of $\theta_1(x)$ at x^k , and is bounded relative to $g(x^k)$ in expectation. The algorithm is linearized due to the following two aspects: (i) The term $G(x^k, \xi)^T(x - x^k)$ in the x -subproblem of SSL-ADMM is a stochastic version of linearization of $\theta_1(x^k)$. (ii) x -subproblem and y -subproblem are added proximal terms $\frac{1}{2}\|x - x^k\|_{G_{1,k}}^2$ and $\frac{1}{2}\|y - y^k\|_{G_{2,k}}^2$ respectively, where $\{G_{1,k}\}$ and $\{G_{2,k}\}$ are two sequences of symmetric and positive definite matrices that can be change with iteration; with the choice of $G_{2,k} \equiv \tau I_{n_2} - \beta B^T B$, $\tau > \beta\|B^T B\|$, the quadratic term in the y -subproblem is linearized. The same fact applies to the x -subproblem. Furthermore, when $G_{1,k} \equiv I_{n_1}$ or is of the form $\tau I_{n_1} - \beta A^T A$, $\tau > 0$, the term $\frac{1}{2}\|y - y^k\|_{G_{2,k}}^2$ vanishes, and $r = 0$, SSL-ADMM reduces to the algorithm appeared in earlier literatures [21, 24]. Finally, the convergence region \mathcal{D} of (r, s) is the same as that in [27]. In particular, if $r = 0$, $s \in (0, \frac{1+\sqrt{5}}{2}]$. Recently, Bai, et al.^[29, 30] studied stochastic ADMM algorithms for finite sum optimization problems. The difference between our paper and theirs is that our goal is to consider a more general stochastic optimization problem (1) where the random variable ξ is not necessarily a discrete random variable. For example, in SP, while it is possible to approximate the function θ_1 by the sample average approximation technique, the extra sample approximation error should also be taken into account.

We start to establish the convergence of SSL-ADMM. The next several lemmas are to obtain an upper bound of $\theta(\tilde{u}^k) - \theta(u) + (\tilde{w}^k - w)^T F(\tilde{w}^k)$. With such a bound, it is possible to estimate the worst-case convergence rate of SSL-ADMM.

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

$$\begin{aligned} \theta(u) - \theta(\tilde{u}^k) + (w - \tilde{w}^k)^T F(\tilde{w}^k) &\geq (w - \tilde{w}^k)^T Q_k (w^k - \tilde{w}^k) - (x - \tilde{x}^k)^T \delta^k \\ &\quad - \frac{L}{2} \|x^k - \tilde{x}^k\|^2, \quad \forall w \in \Omega, \end{aligned} \quad (3)$$

where $\delta^k = G(x^k, \xi) - \nabla \theta_1(x^k)$, similarly hereinafter, and

$$Q_k = \begin{pmatrix} G_{1,k} & 0 & 0 \\ 0 & \beta B^T B + G_{2,k} & -r B^T \\ 0 & -B & \frac{1}{\beta} I_n \end{pmatrix}. \quad (4)$$

Proof Due to Proposition 1, the optimality condition of the x -subproblem in SSL-ADMM is

$$(x - x^{k+1})^T (G(x^k, \xi) - A^T (\lambda^k - \beta (Ax^{k+1} + By^k - b))) + G_{1,k} (x^{k+1} - x^k) \geq 0, \quad \forall x \in \mathcal{X}.$$

Using the notation in (2), the above inequality can be rewritten as

$$(x - \tilde{x}^k)^T (\nabla \theta_1(x^k) + \delta^k - A^T \tilde{\lambda}^k + G_{1,k} (\tilde{x}^k - x^k)) \geq 0, \quad \forall x \in \mathcal{X}.$$

And then using Lemma 1, we have

$$\begin{aligned} \theta_1(x) - \theta_1(\tilde{x}^k) + (x - \tilde{x}^k)^T (-A^T \tilde{\lambda}^k) &\geq (x - \tilde{x}^k)^T G_{1,k} (x^k - \tilde{x}^k) - (x - \tilde{x}^k)^T \delta^k \\ &\quad - \frac{L}{2} \|x^k - \tilde{x}^k\|^2, \quad \forall x \in \mathcal{X}. \end{aligned} \quad (5)$$

Similarly, the optimality condition of y -subproblem in SSL-ADMM is

$$\begin{aligned} & \theta_2(y) - \theta_2(y^{k+1}) + (y - y^{k+1})^T \left(-B^T \lambda^{k+\frac{1}{2}} + \beta B^T (Ax^{k+1} + By^{k+1} - b) \right. \\ & \left. + G_{2,k} (y^{k+1} - y^k) \right) \geq 0, \quad \forall y \in \mathcal{Y}. \end{aligned} \quad (6)$$

Using the notation of $\tilde{\lambda}^k$, $\lambda^{k+\frac{1}{2}} = \lambda^k - r(\lambda^k - \tilde{\lambda}^k) = \tilde{\lambda}^k + (1-r)(\lambda^k - \tilde{\lambda}^k)$. Hence

$$\begin{aligned} & -B^T \lambda^{k+\frac{1}{2}} + \beta B^T (Ax^{k+1} + By^{k+1} - b) \\ & = -B^T \left(\tilde{\lambda}^k + (1-r)(\lambda^k - \tilde{\lambda}^k) \right) + B^T (\lambda^k - \tilde{\lambda}^k) + \beta B^T B (\tilde{y}^k - y^k) \\ & = -B^T \tilde{\lambda}^k + r B^T (\lambda^k - \tilde{\lambda}^k) + \beta B^T B (\tilde{y}^k - y^k). \end{aligned}$$

Substituting this equality into (6), we obtain

$$\begin{aligned} \theta_2(y) - \theta_2(\tilde{y}^k) + (y - \tilde{y}^k)^T \left(-B^T \tilde{\lambda}^k \right) & \geq (y - \tilde{y}^k)^T (\beta B^T B + G_{2,k}) (y^k - \tilde{y}^k) \\ & \quad - r(y - \tilde{y}^k)^T B^T (\lambda^k - \tilde{\lambda}^k). \end{aligned} \quad (7)$$

According to the definition of \tilde{w}^k , we have

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k) = 0,$$

and it can be written as

$$(\lambda - \tilde{\lambda}^k)^T \left\{ (A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k) \right\} \geq 0, \quad \forall \lambda \in \mathbb{R}^n. \quad (8)$$

Combining (5), (7), and (8), and using the notation of (4), the proof is completed. \blacksquare

Lemma 3 *Let the sequence $\{w^k\}$ be generated by the SSL-ADMM and the associated $\{\tilde{w}^k\}$ be defined in (2). Then we have*

$$w^{k+1} = w^k - M(w^k - \tilde{w}^k), \quad (9)$$

where

$$M = \begin{pmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & -s\beta B & (r+s)I_n \end{pmatrix}. \quad (10)$$

Proof

$$\begin{aligned} \lambda^{k+1} &= \lambda^{k+\frac{1}{2}} - s\beta (Ax^{k+1} + By^{k+1} - b) \\ &= \lambda^k - r(\lambda^k - \tilde{\lambda}^k) - s(\beta (Ax^{k+1} + By^k - b) - \beta B(y^k - y^{k+1})) \\ &= \lambda^k - (r+s)(\lambda^k - \tilde{\lambda}^k) + s\beta B(y^k - \tilde{y}^k). \end{aligned}$$

Together with $x^{k+1} = \tilde{x}$ and $y^{k+1} = \tilde{y}$, we prove the assertion of this lemma. \blacksquare

Noting that for the matrices Q_k defined in (4) and M defined in (10), there is a matrix H_k such that $Q_k = H_k M$, where

$$H_k = \begin{pmatrix} G_{1,k} & 0 & 0 \\ 0 & \left(1 - \frac{rs}{r+s}\right) \beta B^T B + G_{2,k} & -\frac{r}{r+s} B^T \\ 0 & -\frac{r}{r+s} B & \frac{1}{\beta(r+s)} I_n \end{pmatrix}. \quad (11)$$

It is easy to check for any $(r, s) \in \mathcal{D}$, H_k is positive semidefinite when the matrix B is full column rank. In fact, to make H_k positive semidefinite alone, it is sufficient for $G_{2,k} \succeq (r-1) \beta B^T B$ when other conditions are satisfied.

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

$$\begin{aligned} & (w - \tilde{w}^k)^T Q_k (w^k - \tilde{w}^k) \\ &= \frac{1}{2} \left(\|w - w^{k+1}\|_{H_k}^2 - \|w - w^k\|_{H_k}^2 \right) + \frac{1}{2} (w^k - \tilde{w}^k)^T G (w^k - \tilde{w}^k), \end{aligned} \quad (12)$$

where $G := Q_k + Q_k^T - M^T H_k M$.

Proof Using $Q_k = H_k M$ and $w^{k+1} = w^k - M(w^k - \tilde{w}^k)$, we have $(w - \tilde{w}^k)^T Q_k (w^k - \tilde{w}^k) = (w - \tilde{w}^k)^T H_k (w^k - w^{k+1})$. Applying the identity $(a-b)^T H(c-d) = \frac{1}{2}(\|a-d\|_H^2 - \|a-c\|_H^2) + \frac{1}{2}(\|c-b\|_H^2 - \|d-b\|_H^2)$, we obtain

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

The remaining task is to simplify the last two terms.

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

The proof is completed. ■

From this lemma, $(w - \tilde{w}^k)^T Q_k (w^k - \tilde{w}^k)$ can be written as two terms, one of which is suitable for recursive operation and the other is a quadratic term, but the matrix G is not necessarily semidefinite. Thus we need to analyze this quadratic term in detail. Since

$$G = \begin{pmatrix} G_{1,k} & 0 & 0 \\ 0 & (1-s) \beta B^T B + G_{2,k} & (s-1) B^T \\ 0 & (s-1) B & \frac{2-r-s}{\beta} I_n \end{pmatrix},$$

$$\begin{aligned}
& (w^k - \tilde{w}^k)^T G (w^k - \tilde{w}^k) \\
&= \|x^k - x^{k+1}\|_{G_{1,k}}^2 + \|y^k - y^{k+1}\|_{G_{2,k}}^2 + (1-s)\beta \|B(y^k - y^{k+1})\|^2 \\
&+ \frac{2-r-s}{\beta} \|\lambda^k - \tilde{\lambda}^k\|^2 + 2(s-1) (\lambda^k - \tilde{\lambda}^k)^T B(y^k - y^{k+1}).
\end{aligned}$$

As the following lemma shows, the last two terms of the above equality can be further analyzed.

Lemma 5 Assume that $(r, s) \in \mathcal{D}$. Let the sequence $\{w^k\}$ be generated by the SSL-ADMM and the associated $\{\tilde{w}^k\}$ be defined in (2). Then we have

$$\begin{aligned}
& \frac{2-r-s}{\beta} \|\lambda^k - \tilde{\lambda}^k\|^2 + 2(s-1) (\lambda^k - \tilde{\lambda}^k)^T B(y^k - y^{k+1}) \\
& \geq \frac{2(1-r)(1-s)}{1+r} \beta (Ax^k + By^k - b)^T B(y^k - y^{k+1}) \\
& + \left(s - r - \frac{2r(1-r)}{1+r} \right) \beta \|B(y^k - y^{k+1})\|^2 + (2-r-s)\beta \|Ax^{k+1} + By^{k+1} - b\|^2 \\
& + \frac{2(1-r)\beta}{1+r} \left(\|y^{k+1} - y^k\|_{G_{2,k}}^2 - \frac{1}{2} \|y^k - y^{k-1}\|_{G_{2,k-1}}^2 - \frac{1}{2} \|y^{k+1} - y^k\|_{G_{2,k-1}}^2 \right). \tag{13}
\end{aligned}$$

Proof It follows from the optimality condition of y -subproblem for $(k+1)$ -th iteration that

$$\begin{aligned}
& \theta_2(y^k) - \theta_2(y^{k+1}) + (y^k - y^{k+1})^T \left(-B^T \lambda^{k+\frac{1}{2}} + \beta B^T (Ax^{k+1} + By^{k+1} - b) \right. \\
& \left. + G_{2,k} (y^{k+1} - y^k) \right) \geq 0.
\end{aligned}$$

Similarly, it follows from the optimality condition of y -subproblem for k -th iteration that

$$\begin{aligned}
& \theta_2(y^{k+1}) - \theta_2(y^k) + (y^{k+1} - y^k)^T \left(-B^T \lambda^{k-\frac{1}{2}} + \beta B^T (Ax^k + By^k - b) \right. \\
& \left. + G_{2,k-1} (y^k - y^{k-1}) \right) \geq 0.
\end{aligned}$$

Adding these two inequalities and using

$$\lambda^{k-\frac{1}{2}} - \lambda^{k+\frac{1}{2}} = r\beta (Ax^{k+1} + By^{k+1} - b) + s\beta (Ax^k + By^k - b) + r\beta B(y^k - y^{k+1}),$$

we obtain

$$\begin{aligned}
& \{(r+1)\beta (Ax^{k+1} + By^{k+1} - b) + (s-1)\beta (Ax^k + By^k - b) + r\beta B(y^k - y^{k+1})\}^T \\
& B(y^k - y^{k+1}) \geq (y^k - y^{k+1})^T (G_{2,k-1} (y^k - y^{k-1}) - G_{2,k} (y^{k+1} - y^k)).
\end{aligned}$$

This implies

$$\begin{aligned}
& (Ax^{k+1} + By^{k+1} - b)^T B(y^k - y^{k+1}) \\
& \geq \frac{1-s}{1+r} (Ax^k + By^k - b)^T B(y^k - y^{k+1}) - \frac{r}{1+r} \|B(y^k - y^{k+1})\|^2 \\
& + \frac{\beta}{1+r} \left(\|y^{k+1} - y^k\|_{G_{2,k}}^2 - \frac{1}{2} \|y^k - y^{k-1}\|_{G_{2,k-1}}^2 - \frac{1}{2} \|y^{k+1} - y^k\|_{G_{2,k-1}}^2 \right). \tag{14}
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
& 2(s-1) \left(\lambda^k - \tilde{\lambda}^k \right)^T B (y^k - y^{k+1}) \\
&= 2(s-1) \beta (Ax^{k+1} + By^k - b)^T B (y^k - y^{k+1}) \\
&= 2(s-1) \left\{ \beta (Ax^{k+1} + By^{k+1} - b)^T B (y^k - y^{k+1}) + \beta \|B(y^k - y^{k+1})\|^2 \right\}
\end{aligned} \tag{15}$$

and

$$\begin{aligned}
& \frac{2-r-s}{\beta} \left\| \lambda^k - \tilde{\lambda}^k \right\|^2 \\
&= (2-r-s) \beta \|Ax^{k+1} + By^k - b\|^2 \\
&= (2-r-s) \beta \left\| (Ax^{k+1} + By^{k+1} - b) + B(y^k - y^{k+1}) \right\|^2 \\
&= (2-r-s) \beta \|Ax^{k+1} + By^{k+1} - b\|^2 + (2-r-s) \beta \|B(y^k - y^{k+1})\|^2 \\
&\quad + 2(2-r-s) \beta (Ax^{k+1} + By^{k+1} - b)^T B (y^k - y^{k+1}).
\end{aligned} \tag{16}$$

Combining (14), (15), and (16), we get the assertion of this lemma. \blacksquare

According to this lemma and using Cauchy-Schwarz inequality, the term $(w^k - \tilde{w}^k)^T G(w^k - \tilde{w}^k)$ can be bounded as follows:

$$\begin{aligned}
& (w^k - \tilde{w}^k)^T G(w^k - \tilde{w}^k) \\
&\geq \left(2-r-s - \frac{(1-s)^2}{1+r} \right) \beta \|Ax^{k+1} + By^{k+1} - b\|^2 + \|x^k - x^{k+1}\|_{G_{1,k}}^2 + \|y^k - y^{k+1}\|_{G_{2,k}}^2 \\
&\quad + \frac{2(1-r)\beta}{1+r} \left(\|y^{k+1} - y^k\|_{G_{2,k}}^2 - \frac{1}{2} \|y^k - y^{k-1}\|_{G_{2,k-1}}^2 - \frac{1}{2} \|y^{k+1} - y^k\|_{G_{2,k-1}}^2 \right) \\
&\quad + \frac{(1-s)^2}{1+r} \beta \left(\|Ax^{k+1} + By^{k+1} - b\|^2 - \|Ax^k + By^k - b\|^2 \right).
\end{aligned} \tag{17}$$

Now combining (17), Lemma 2, and Lemma 4, we obtain the following main theorem. In this theorem, we take $G_{1,k}$ of the form $\tau_k I_{n_1} - \beta A^T A$, $\tau_k > 0$, 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 $G_2 = \eta I_{n_2} - \beta B^T B$, $\eta \geq \beta \|B^T B\|$, then y -subproblem reduces to the proximal mapping of g .

Theorem 1 Assume that $(r, s) \in \mathcal{D}$. Let the sequence $\{w^k\}$ be generated by the SSL-ADMM and the associated $\{\tilde{w}^k\}$ be defined in (2), and

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

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

$$\begin{aligned} & \theta(\bar{u}_N) - \theta(u) + (\bar{w}_N - w)^T F(w) \\ & \leq \frac{1}{2N} \|w^1 - w\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 + \frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 \\ & \quad + \frac{1}{N} \sum_{t=1}^N (x - x^t)^T \delta^t + \frac{1}{2N\sqrt{N}} \sum_{t=1}^N \|\delta^t\|^2. \end{aligned} \quad (18)$$

Proof It is sufficient for using convexity of θ , $(x^k - x^{k+1})^T \delta^k \leq \frac{\sqrt{N}}{2} \|x^k - x^{k+1}\|^2 + \frac{1}{2\sqrt{N}} \|\delta^k\|^2$, and $(\tilde{w}^k - w)^T F(w) = (\tilde{w}^k - w)^T F(\tilde{w}^k)$. \blacksquare

Corollary 1 Assume that all the conditions in Theorem 1 hold, then SSL-ADMM has the following properties

(i)

$$\begin{aligned} & \mathbb{E} [\|A\bar{x}_N + B\bar{y}_N - b\|] \\ & \leq \frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 \\ & \quad + \frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\sigma^2}{2\sqrt{N}}, \end{aligned} \quad (19)$$

(ii)

$$\begin{aligned} & \mathbb{E} [\theta(\bar{u}_N) - \theta(u^*)] \\ & \leq (\|\lambda^*\| + 1) \left(\frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 \right) \\ & \quad + (\|\lambda^*\| + 1) \left(\frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\sigma^2}{2\sqrt{N}} \right), \end{aligned} \quad (20)$$

where e is a unit vector satisfying $-e^T (A\bar{x}_N + B\bar{y}_N - b) = \|A\bar{x}_N + B\bar{y}_N - b\|$ and the expectation is taken conditional on w^1 .

Proof Let $w = (x^*, y^*, \lambda)$ in (18), where $\lambda = \lambda^* + e$, then the left hand side of (18) is $\theta(\bar{u}_N) - \theta(u^*) - (\lambda^*)^T (A\bar{x}_N + B\bar{y}_N - b) + \|A\bar{x}_N + B\bar{y}_N - b\|$, which is followed from

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

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 $\theta(\bar{u}_N) - \theta(u^*) - (\lambda^*)^T (A\bar{x}_N + B\bar{y}_N - b) \geq 0$. Hence, the left hand side of (18) is equal or greater than $\|A\bar{x}_N + B\bar{y}_N - b\|$ when letting $w = (x^*, y^*, \lambda^* + e)$ and (19) is obtained by taking expectation. Substituting $w = \bar{w}_N$ into the

variational inequality associated with (1), we can also get

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

i.e.,

$$\begin{aligned} \theta(\bar{u}_N) - \theta(u^*) &\leq \theta(\bar{u}_N) - \theta(u^*) + (\bar{w}_N - w^*)^T F(w^*) \\ &\quad + \|\lambda^*\| \|A\bar{x}_N + B\bar{y}_N - b\|. \end{aligned}$$

Then by taking expectation, (20) is obtained. ■

Remark 1 (i) In Theorem 1 or Corollary 1, τ_k 's are constant, and N needs to be selected in advance. In fact, τ_k can also vary with the number of iterations, e.g., $\tau_k = \sqrt{k} + M$. In this case, if the distance between w^k and w^* is bounded, i.e., $\|w^k - w^*\|^2 \leq R^2$ for any k , we can also obtain a worst-case convergence rate. The difference with the proof idea in Theorem 1 and Corollary 1 is bounding the term $\sum_{t=0}^k (\|x^t - x^*\|_{G_{1,t}}^2 - \|x^{t+1} - x^*\|_{G_{1,t}}^2)$, which is now bounded as follows.

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

(ii) Corollary 1 reveals that the worst-case convergence rate of SSL-ADMM for solving general convex problems is $\mathcal{O}(\frac{1}{\sqrt{N}})$, where N is the iteration number.

At the end of this section, we assume that θ_1 is μ -strongly convex, i.e., $\theta_1(x) \geq \theta_1(y) + \langle \nabla \theta_1(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$, $\mu > 0$ for all $x, y \in \mathcal{X}$. 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 SSL-ADMM.

Theorem 2 Assume that $(r, s) \in \mathcal{D}$. Let the sequence $\{w^k\}$ be generated by the SSL-ADMM and the associated $\{\tilde{w}^k\}$ be defined in (2), and

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

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

(i)

$$\begin{aligned}
& \mathbb{E} [\|A\bar{x}_k + B\bar{y}_k - b\|] \\
& \leq \frac{1}{2k} \|(y^1, \lambda^1) - (y^*, \lambda^* + e)\|_{H_{1;2 \times 2}}^2 + \frac{(1-r)\beta}{2(1+r)k} \|y^1 - y^0\|_{G_2}^2 \\
& \quad + \frac{(1-s)^2\beta}{2(1+r)k} \|Ax^1 + By^1 - b\|^2 + \frac{\mu}{2k} \|x^1 - x^*\|^2 + \frac{(1+\ln k)\sigma^2}{2\mu k},
\end{aligned} \tag{21}$$

(ii)

$$\begin{aligned}
& \mathbb{E} [\theta(\bar{u}_k) - \theta(u^*)] \\
& \leq (\|\lambda^*\| + 1) \left(\frac{1}{2k} \|(y^1, \lambda^1) - (y^*, \lambda^* + e)\|_{H_{1;2 \times 2}}^2 + \frac{(1-r)\beta}{2(1+r)k} \|y^1 - y^0\|_{G_2}^2 \right) \\
& \quad + (\|\lambda^*\| + 1) \left(\frac{(1-s)^2\beta}{2(1+r)k} \|Ax^1 + By^1 - b\|^2 + \frac{\mu}{2k} \|x^1 - x^*\|^2 + \frac{(1+\ln k)\sigma^2}{2\mu k} \right),
\end{aligned} \tag{22}$$

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

$$H_{1;2 \times 2} = \begin{pmatrix} \left(1 - \frac{rs}{r+s}\right) \beta B^T B + G_2 & -\frac{r}{r+s} B^T \\ -\frac{r}{r+s} B & \frac{1}{\beta(r+s)} I_n \end{pmatrix},$$

and the expectation is taken conditional on w^1 .

Proof First, similar to the proof of Lemma 2, using the μ -strong convexity of θ_1 , we conclude that for any Ω

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

Then using $Q_k = H_k M$, $(\tilde{w}^t - w)^T F(\tilde{w}^t) = (\tilde{w}^t - w)^T F(w)$, Lemma 4, and (17), we get

$$\begin{aligned}
& \theta(\tilde{u}^t) - \theta(u) + (\tilde{w}^t - w)^T F(w) \\
& \leq \frac{1}{2} \left(\|(y^t, \lambda^t) - (y, \lambda)\|_{H_{1;2 \times 2}}^2 - \|(y^{t+1}, \lambda^{t+1}) - (y, \lambda)\|_{H_{1;2 \times 2}}^2 \right) + (x - x^t)^T \delta^t \\
& \quad + \frac{(1-s)^2\beta}{2(1+r)} \left(\|Ax^t + By^t - b\|^2 - \|Ax^{t+1} + By^{t+1} - b\|^2 \right) + \frac{1}{2\mu(t+1)} \|\delta^t\|^2 \\
& \quad + \frac{(1-r)\beta}{2(1+r)} \left(\|y^t - y^{t-1}\|_{G_2}^2 - \|y^{t+1} - y^t\|_{G_2}^2 \right) \\
& \quad + \frac{1}{2} \left(\mu t \|x^t - x\|^2 - \mu(t+1) \|x^{t+1} - x\|^2 \right).
\end{aligned}$$

Adding the above inequalities from $t = 1$ to k and divided by k , and then following the proof of Corollary 1, we prove the assertion of this theorem. \blacksquare

This theorem implies that under the assumption that θ_1 is strongly convex, the worst-case convergence rate for the SSL-ADMM can be improved to $\mathcal{O}((\ln k)/k)$ with the choice of diminishing size. The following theorem shows the convergence of ergodic iterates of SSL-ADMM, which is not covered in some earlier literatures [21, 24]. Furthermore, if θ_2 is also strongly convex, the assumption that B is full column rank can be removed.

Theorem 3 Assume that $(r, s) \in \mathcal{D}$. Let the sequence $\{w^k\}$ be generated by the SSL-ADMM, the associated $\{\tilde{w}^k\}$ be defined in (2), and

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

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

$$\begin{aligned} & \mathbb{E} [\|\bar{x}_k - x^*\| + \|\bar{y}_k - y^*\|] \\ & \leq \left(1 + \frac{\|A\|}{\sqrt{\lambda_{\min}}}\right) \sqrt{\left[\frac{2}{\mu} (\mathbb{E} [\theta(\bar{u}_k) - \theta(u^*)] + \|\lambda^*\| \mathbb{E} [\|A\bar{x}_k + B\bar{y}_k - b\|])\right]} \\ & \quad + \frac{1}{\sqrt{\lambda_{\min}}} \mathbb{E} \|A\bar{x}_k + B\bar{y}_k - b\|, \end{aligned} \quad (23)$$

where the bounds for $\mathbb{E} [\|A\bar{x}_k + B\bar{y}_k - b\|]$ and $\mathbb{E} [\theta(\bar{u}_k) - \theta(u^*)]$ are the same as in (21) and (22) respectively, and the expectation is taken conditional on w^1 .

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

$$\theta_1(\bar{x}_k) \geq \theta_1(x^*) + (\lambda^*)^T (A\bar{x}_k - Ax^*) + \frac{\mu}{2} \|\bar{x}_k - x^*\|^2 \quad (24)$$

and

$$\theta_2(\bar{y}_k) \geq \theta_2(y^*) + (\lambda^*)^T (B\bar{y}_k - By^*). \quad (25)$$

Adding up (24) and (25), we get $\theta(\bar{u}_k) \geq \theta(u^*) + (\lambda^*)^T (A\bar{x}_k + B\bar{y}_k - b) + \frac{\mu}{2} \|\bar{x}_k - x^*\|^2$, that is

$$\begin{aligned} \|\bar{x}_k - x^*\| & \leq \sqrt{\frac{2}{\mu} \left(\theta(\bar{u}_k) - \theta(u^*) - (\lambda^*)^T (A\bar{x}_k + B\bar{y}_k - b) \right)} \\ & \leq \sqrt{\frac{2}{\mu} (\theta(\bar{u}_k) - \theta(u^*) + \|\lambda^*\| \|A\bar{x}_k + B\bar{y}_k - b\|)}. \end{aligned} \quad (26)$$

On the other hand,

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

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

$$\|\bar{y}_k - y^*\| \leq \frac{\|A\|}{\sqrt{\lambda_{\min}}} \|\bar{x}_k - x^*\| + \frac{1}{\sqrt{\lambda_{\min}}} \|A\bar{x}_k + B\bar{y}_k - b\|. \quad (27)$$

Adding (26) and (27), 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

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

The proof is completed. ■

4 High Probability Performance Analysis

In this section, we shall establish the large deviation properties of SSL-ADMM. By (19) and (20), and Markov's inequality, we have for any $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ that

$$\Pr \left\{ \|A\bar{x}_N + B\bar{y}_N - b\| \leq \varepsilon_1 \left(\frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 + \frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\sigma^2}{2\sqrt{N}} \right) \right\} \geq 1 - \frac{1}{\varepsilon_1} \quad (28)$$

and

$$\Pr \left\{ \theta(\bar{u}_N) - \theta(u^*) \leq \varepsilon_2 \left((\|\lambda^*\| + 1) \left(\frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 + (\|\lambda^*\| + 1) \left(\frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\sigma^2}{2\sqrt{N}} \right) \right) \right\} \geq 1 - \frac{1}{\varepsilon_2}. \quad (29)$$

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., [1, 32]. Specifically, assume that for any $x \in \mathcal{X}$

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

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 \mathcal{X} is bounded and its diameter is denoted by D_X , defined as $\max_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\|$. The following theorem shows the high probability bound for objective error and constraint violation of SSL-ADMM.

Theorem 4 *Assume that all the conditions in Theorem 1 hold, then SSL-ADMM has the following properties*

(i)

$$\Pr \left\{ \|A\bar{x}_N + B\bar{y}_N - b\| \leq \frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 + \frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} + \frac{1}{2\sqrt{N}} (1 + \Theta) \sigma^2 \right\} \geq 1 - \exp \{-\Theta^2/3\} - \exp \{-\Theta\}, \quad (30)$$

(ii)

$$\Pr \left\{ \theta(\bar{u}_N) - \theta(u^*) \leq (\|\lambda^*\| + 1) \left(\frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 + (\|\lambda^*\| + 1) \left(\frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} + \frac{1}{2\sqrt{N}} (1 + \Theta) \sigma^2 \right) \right\} \geq 1 - \exp \{-\Theta^2/3\} - \exp \{-\Theta\}, \quad (31)$$

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

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

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

Now using Proposition 2 for the martingale-difference sequence, we have for any $\Theta \geq 0$

$$\Pr \left\{ \sum_{t=1}^N \zeta^t > \frac{\Theta D_X \sigma}{\sqrt{N}} \right\} \leq \exp \{-\Theta^2/3\}. \quad (32)$$

Also, observe that by Jensen's inequality for the exponential function

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

whence, taking expectation,

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

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

$$\Pr \left\{ \frac{1}{N} \sum_{t=1}^N \|\delta^t\|^2 \geq (1 + \Theta) \sigma^2 \right\} \leq \exp \{-\Theta\}. \quad (33)$$

Using (32) and (33) in (18) for $w = (x^*, y^*, \lambda^* + e)$, we conclude that

$$\begin{aligned} \Pr \left\{ \|A\bar{x}_N + B\bar{y}_N - b\| > \frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 \right. \\ \left. + \frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} + \frac{1}{2\sqrt{N}} (1 + \Theta) \sigma^2 \right\} \leq \exp \{-\Theta^2/3\} + \exp \{-\Theta\} \end{aligned} \quad (34)$$

and

$$\begin{aligned} \Pr \left\{ \theta(\bar{u}_N) - \theta(u^*) > (\|\lambda^*\| + 1) \left(\frac{1}{2N} \|w^1 - (x^*, y^*, \lambda^* + e)\|_{H_1}^2 + \frac{(1-r)\beta}{2(1+r)N} \|y^1 - y^0\|_{G_2}^2 \right) \right. \\ \left. + (\|\lambda^*\| + 1) \left(\frac{(1-s)^2\beta}{2(1+r)N} \|Ax^1 + By^1 - b\|^2 + \frac{\Theta D_X \sigma}{\sqrt{N}} + \frac{1}{2\sqrt{N}} (1 + \Theta) \sigma^2 \right) \right\} \\ \leq \exp \{-\Theta^2/3\} + \exp \{-\Theta\}. \end{aligned} \quad (35)$$

The result immediately follows from the above inequalities. ■

Remark 2 In view of Theorem 4, if we take $\Theta = \ln N$, then we have

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

and

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

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

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

and

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

Observe that the convergence rate of ergodic iterates of SSL-ADMM is obtained in Theorem 3. The high probability bound can be also established, which is shown as follows

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

where N is the iteration number. In contrast to (28) and (29), we can observe that the results in Theorem 4 are much finer.

5 Preliminary Numerical Experiments

In this section, we report some numerical results on the following graph-guided fused lasso problem in statistical machine learning:

$$\min_x \mathbb{E}_\xi f_\xi(x) + \mu \|Ax\|_1,$$

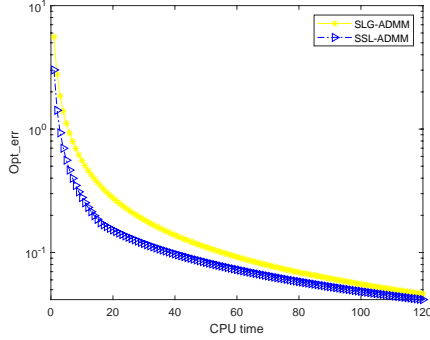
where $f_\xi(x) = \log(1 + \exp(-t \cdot l^T x))$ is the logistic loss function on the feature-label pair $\xi = (l, t) \in \mathbb{R}^d \times \{-1, 1\}$, μ is a given regularization parameter, and $A = [G; I]$, where G is obtained by sparse inverse covariance estimation [33]. By introducing another block variable y , and imposing the constraint $Ax = y$, this problem is reformulated into a form of (1) with $\theta_1(x) = \mathbb{E}_\xi f_\xi(x)$, $\theta_2(y) = \mu \|y\|_1$, $B = -I$, and $b = 0$.

The dataset used in numerical experiments is taken from the LIBSVM website¹, which is summarized in the Table 1. In our experiments, the regularization parameter μ and penalty parameter β are set to be 1×10^{-5} and 1×10^{-3} , respectively; the initial points are set to be uniformly random vectors in the interval $[-1, 1]^d$; other parameters are chosen according to the conditions in corollaries. We plot Opt_err, the maximum of the objective function value error and constraint violation, versus CPU time in seconds, where the approximate optimal objective function value is obtained by running some convergent ADMM-type algorithm for more than 10 minutes. Figure 1 shows the performances of SSL-ADMM and generalized version of [21] (the algorithm in [21] is indeed a stochastic linearized ADMM, hence we call the generalized version of it SLG-ADMM and it is more efficient than the original version), which updates dual variable only once at each iteration, and the results indicate that an improvement from symmetrically updating dual variable also occurs in the stochastic setting. Since our paper is to demonstrate that symmetrically update multipliers are still valid for stochastic optimization problems, we only compare SSL-ADMM with the algorithm in [21] (this paper made comparisons with other algorithms for stochastic optimization), not with the algorithms designed for deterministic finite sum optimization problems, for example, in [29, 30].

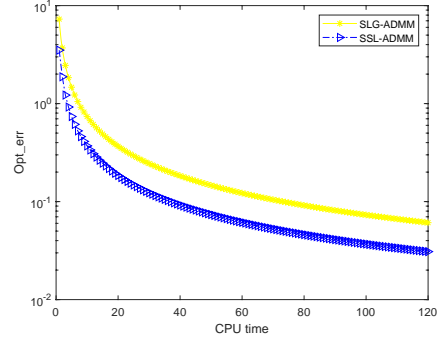
¹<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>.

Table 1 Real-world datasets and regularization parameters

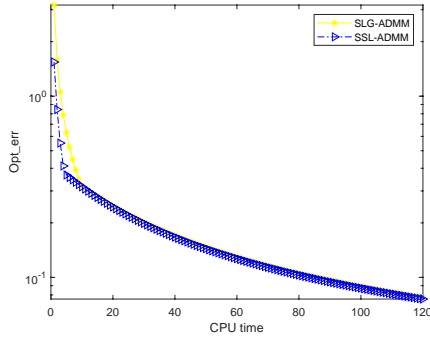
Dataset	Number of samples	Dimensionality
a8a	22696	123
a9a	32561	123
ijcnn1	49990	22
w8a	49749	300



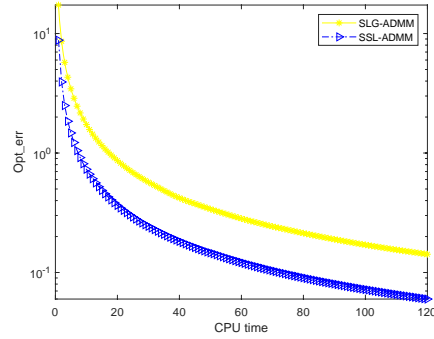
(a) a8a



(b) a9a



(c) ijcnn1



(d) w8a

Figure 1 Opt_err vs (vertical axis) vs the CPU time (s) (horizontal axis) for a8a, a9a, ijcnn1, and w8a respectively

6 Summary

In this paper, we analyze the expected convergence rates and the large deviation properties of a stochastic variant of symmetric ADMM using the variational inequality framework. By means of this framework, the proof is very clear. Numerical experiments on some real-world datasets demonstrate that symmetrically updating the dual variable can lead to an algorithmic improvement in the stochastic setting. When the model is deterministic and \mathcal{SFO} is not needed, our proposed algorithm reduces to a symmetric proximal ADMM, and the convergence region of (r, s) is the same as that in the corresponding literature.

References

- [1] Nemirovski A, Juditsky A, Lan G, et al. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 2009, 19(4): 1574–1609.
- [2] Lan G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2012, 133(1): 365–397.
- [3] Ghadimi S, Lan G, Zhang H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 2016, 155(1): 267–305.
- [4] Robbins H, Monro S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951, 22(3): 400–407.
- [5] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 2011, 3(1): 1–122.
- [6] Glowinski R, Marroco A. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Journal of Equine Veterinary Science*, 1975, 9(2): 41–76.
- [7] Gabay D, Mercier B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 1976, 2(1): 17–40.
- [8] Glowinski R. On alternating direction methods of multipliers: A historical perspective. *Modeling, Simulation and Optimization for Science and Technology*, Springer, Dordrecht, 2014: 59–82.
- [9] Eckstein J, Bertsekas D P. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 1992, 55(1): 293–318.
- [10] He B, Yuan X. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 2012, 50(2): 700–709.
- [11] Monteiro R D C, Svaiter B F. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 2013, 23(1): 475–507.
- [12] He B, Yuan X. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Numerische Mathematik*, 2015, 130(3): 567–577.
- [13] Deng W, Yin W. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 2016, 66(3): 889–916.
- [14] Yang W H, Han D. Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. *SIAM Journal on Numerical Analysis*, 2016, 54(2): 625–640.
- [15] Han D, Sun D, Zhang L. Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Mathematics of Operations Research*, 2018, 43(2): 622–637.
- [16] Li G, Pong T K. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 2015, 25(4): 2434–2460.
- [17] Wang Y, Yin W, Zeng J. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 2019, 78(1): 29–63.
- [18] Jiang B, Lin T, Ma S, et al. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, 2019, 72(1): 115–157.
- [19] Zhang J, Luo Z Q. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 2020, 30(3): 2272–2302.
- [20] Han D R. A survey on some recent developments of alternating direction method of multipliers. *Journal of the Operations Research Society of China*, 2022, 10(1): 1–52.
- [21] Ouyang H, He N, Tran L, et al. Stochastic alternating direction method of multipliers. *International Conference on Machine Learning*, 2013: 80–88.
- [22] Suzuki T. Dual averaging and proximal gradient descent for online alternating direction multiplier method. *International Conference on Machine Learning*, 2013: 392–400.
- [23] Zhao P, Yang J, Zhang T, et al. Adaptive stochastic alternating direction method of multipliers. *International Conference on Machine Learning*, 2015: 69–77.
- [24] Gao X, Jiang B, Zhang S. On the information-adaptive variants of the ADMM: An iteration complexity perspective. *Journal of Scientific Computing*, 2018, 76(1): 327–363.

- [25] He B, Liu H, Wang Z, et al. A strictly contractive Peaceman-Rachford splitting method for convex programming. *SIAM Journal on Optimization*, 2014, 24(3): 1011–1040.
- [26] He B, Ma F, Yuan X. Convergence study on the symmetric version of ADMM with larger step sizes. *SIAM Journal on Imaging Sciences*, 2016, 9(3): 1467–1501.
- [27] Bai J, Li J, Xu F, et al. Generalized symmetric ADMM for separable convex optimization. *Computational Optimization and Applications*, 2018, 70(1): 129–170.
- [28] Lions P L, Mercier B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 1979, 16(6): 964–979.
- [29] Bai J, Hager W W, Zhang H. An inexact accelerated stochastic ADMM for separable convex optimization. *Computational Optimization and Applications*, 2022, 81(2): 479–518.
- [30] Bai J, Han D, Sun H, et al. Convergence analysis of an inexact accelerated stochastic ADMM with larger stepsizes. *CSIAM Transactions on Applied Mathematics*, 2022, 3(3): 448–479.
- [31] He B S. On the convergence properties of alternating direction method of multipliers. *Numerical Mathematics*, 2017, 39: 81–96.
- [32] Lan G. *First-order and stochastic optimization methods for machine learning*, Springer, New York, 2020.
- [33] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008, 9(3): 432–441.