CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2021-0021

Vol. **x**, No. x, pp. 1-32
xx 2022

Convergence on a Symmetric Accelerated Stochastic ADMM with Larger Stepsizes

Jianchao Bai[1], Deren Han[2,*], Hao Sun[3] and Hongchao Zhang[4]

[1] *School of Mathematics and Statistics, the MIIT Key Laboratory of Dynamics and Control of Complex Systems, Northwestern Polytechnical University, Xi'an, 710129, P.R. China.*
[2] *LMIB of the Ministry of Education, School of Mathematical Sciences, Beihang University, Beijing, 100191, P.R. China.*
[3] *School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an, 710129, P.R. China.*
[4] *Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA.*

Received 30 March 2021; Accepted 19 December 2021

Abstract. In this paper, we develop a symmetric accelerated stochastic Alternating Direction Method of Multipliers (SAS-ADMM) for solving separable convex optimization problems with linear constraints. The objective function is the sum of a possibly nonsmooth convex function and an average function of many smooth convex functions. Our proposed algorithm combines both ideas of ADMM and the techniques of accelerated stochastic gradient methods possibly with variance reduction to solve the smooth subproblem. One main feature of SAS-ADMM is that its dual variable is symmetrically updated after each update of the separated primal variable, which would allow a more flexible and larger convergence region of the dual variable compared with that of standard deterministic or stochastic ADMM. This new stochastic optimization algorithm is shown to have ergodic converge in expectation with $\mathcal{O}(1/T)$ convergence rate, where T denotes the number of outer iterations. Our preliminary experiments indicate the proposed algorithm is very effective for solving separable optimization problems from big-data applications. Finally, 3-block extensions of the algorithm and its variant of an accelerated stochastic augmented Lagrangian method are discussed in the appendix.

AMS subject classifications: 65K10, 65Y20, 68W40, 90C25

Key words: Convex optimization, stochastic ADMM, symmetric ADMM, larger stepsize, proximal mapping, complexity.

*Corresponding author. *Email addresses:* jianchaobai@nwpu.edu.cn (J. Bai), handr@buaa.edu.cn (D. Han), hsun@nwpu.edu.cn (H. Sun), hozhang@math.lsu.edu (H. Zhang)

1 Introduction

We consider the following structured composite convex optimization problem with linear equality constraints:

$$\min\{f(\boldsymbol{x})+g(\boldsymbol{y})\,|\,\boldsymbol{x}\in\mathcal{X},\ \boldsymbol{y}\in\mathcal{Y},\ A\boldsymbol{x}+B\boldsymbol{y}=\boldsymbol{b}\},\tag{1.1}$$

where $\mathcal{X}\subset\mathbb{R}^{n_1}$, $\mathcal{Y}\subset\mathbb{R}^{n_2}$ are closed convex subsets, $A\in\mathbb{R}^{n\times n_1}$, $B\in\mathbb{R}^{n\times n_2}$, $\boldsymbol{b}\in\mathbb{R}^n$ are given, $g:\mathcal{Y}\to\mathbb{R}\cup\{+\infty\}$ is a convex but possibly nonsmooth function, and f is an average of N real-valued convex functions:

$$f(\boldsymbol{x})=\frac{1}{N}\sum_{j=1}^{N}f_j(\boldsymbol{x}).$$

We assume that each f_j defined on an open set containing $\mathcal{X}$ is Lipschitz continuously differentiable on $\mathcal{X}$. Problem (1.1) is also referred as the regularized empirical risk minimization in big-data applications [26,35], including classification and regression models in machine learning, where N denotes the sample size and f_j corresponds to the empirical loss. A major difficulty for solving (1.1) is that the sample size N can be very large such that it is often computationally prohibitive to evaluate either the full function value or the gradient of f at each iteration of an algorithm. Hence, it is essential for an effective algorithm, e.g., a stochastic gradient method, to explore the summation structure of f in the objective function.

The augmented Lagrangian function of (1.1) is

$$\mathcal{L}_\beta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda})=\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda})+\frac{\beta}{2}\|A\boldsymbol{x}+B\boldsymbol{y}-\boldsymbol{b}\|^2,\tag{1.2}$$

where $\beta>0$ is a penalty parameter, $\boldsymbol{\lambda}$ is the Lagrange multiplier and the Lagrangian of (1.1) is defined as

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda})=f(\boldsymbol{x})+g(\boldsymbol{y})-\boldsymbol{\lambda}^{\mathsf{T}}(A\boldsymbol{x}+B\boldsymbol{y}-\boldsymbol{b}).\tag{1.3}$$

Although the Augmented Lagrangian Method (ALM) can be applied to solve (1.1), it does not take full advantage of the separable structure of (1.1). As a splitting version of ALM, the standard Alternating Direction Method of Multipliers (ADMM, [11,12]) exploits the separable structure of the objective function and performs the following iterations:

$$\begin{cases}\boldsymbol{x}^{k+1}\in\arg\min\limits_{\boldsymbol{x}\in\mathcal{X}}\mathcal{L}_\beta(\boldsymbol{x},\boldsymbol{y}^k,\boldsymbol{\lambda}^k),\\ \boldsymbol{y}^{k+1}\in\arg\min\limits_{\boldsymbol{y}\in\mathcal{Y}}\mathcal{L}_\beta(\boldsymbol{x}^{k+1},\boldsymbol{y},\boldsymbol{\lambda}^k),\\ \boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^k-s\beta\Big(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\Big),\end{cases}$$

where $s\in(0,\frac{1+\sqrt{5}}{2})$ is the stepsize for updating the dual variable $\boldsymbol{\lambda}$.

If the Peaceman-Rachford Splitting Method (PRSM, [27]) is applied to the dual of (1.1), then we obtain a variation of ADMM, whose iteration reads

$$\begin{cases} \boldsymbol{x}^{k+1} \in \arg\min\limits_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}_\beta(\boldsymbol{x},\boldsymbol{y}^k,\boldsymbol{\lambda}^k), \\ \boldsymbol{\lambda}^{k+\frac{1}{2}} = \boldsymbol{\lambda}^k - \beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}\right), \\ \boldsymbol{y}^{k+1} \in \arg\min\limits_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}_\beta(\boldsymbol{x}^{k+1},\boldsymbol{y},\boldsymbol{\lambda}^{k+\frac{1}{2}}), \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^{k+\frac{1}{2}} - \beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right). \end{cases}$$

PRSM is also called Symmetric ADMM (S-ADMM) since the Lagrange multipliers are symmetrically updated twice in each loop. Note that both updates of dual variable in PRSM use the same constant stepsize 1. Recently, Luo-Yang [25] proposed a fast S-ADMM for solving (1.1) with only equality constraints, where the Nesterov's acceleration technique was applied for an additional update of $\boldsymbol{\lambda}^{k+1}$ and the $\boldsymbol{y}$ variable was updated again by solving

$$\min_{\boldsymbol{y}} g(\boldsymbol{y}) - (\widehat{\boldsymbol{\lambda}}^{k+1})^\top B\boldsymbol{y} \quad \text{with} \quad \widehat{\boldsymbol{\lambda}}^{k+1} = \boldsymbol{\lambda}^{k+1} + \frac{\theta^{k+1}(1-\theta^k)}{\theta^k}(\boldsymbol{\lambda}^{k+1}-\boldsymbol{\lambda}^k)$$

and $\theta^{k+1} = 2/(k+1)$. Motivated from the ideas of enlarging the dual stepsize in [18], Gu, et al. [14] proposed a symmetric proximal ADMM whose dual variable is updated twice with different stepsizes. Meanwhile, the following extension of S-ADMM was developed by He, et al. [19]:

$$\begin{cases} \boldsymbol{x}^{k+1} \in \arg\min\limits_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}_\beta(\boldsymbol{x},\boldsymbol{y}^k,\boldsymbol{\lambda}^k), \\ \boldsymbol{\lambda}^{k+\frac{1}{2}} = \boldsymbol{\lambda}^k - \tau\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}\right), \\ \boldsymbol{y}^{k+1} \in \arg\min\limits_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}_\beta(\boldsymbol{x}^{k+1},\boldsymbol{y},\boldsymbol{\lambda}^{k+\frac{1}{2}}), \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^{k+\frac{1}{2}} - s\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right), \end{cases} \tag{1.4}$$

where, for the sake of convergence, the stepsize pair (τ,s) is required to belong to the following region:

$$\Delta_0 = \left\{(\tau,s) \mid s\in(0,(1+\sqrt{5})/2),\ \tau+s>0,\ \tau\in(-1,1),\ |\tau|<1+s-s^2\right\}.$$

Bai et al. [1] further designed a Generalized Symmetric ADMM (GS-ADMM) for solving a multi-block separable convex optimization and enlarged the above region Δ_0 to Δ defined in (3.1). Numerical experiments show that symmetrically updating the dual variable in a more flexible way often improves the algorithm performance [1, 14, 19]. The sublinear convergence rate of GS-ADMM in the nonergodic sense and its linear convergence rate were shown in [2]. To our knowledge, Δ is currently the largest convergence region

of the dual stepsizes for symmetric ADMM-type algorithms and has been used in the logarithmic-quadratic proximal based ADMM for solving the 2-block problems [17, 29] and the grouped multi-block problems [4].

Another line of developing ADMM is to apply relaxation techniques such as using

$$\chi^{k+1}=\alpha A\boldsymbol{x}^{k+1}+(1-\alpha)(\boldsymbol{b}-B\boldsymbol{y}^k)$$

to replace $A\boldsymbol{x}^{k+1}$ during the updates of $\boldsymbol{y}^{k+1}$ and $\boldsymbol{\lambda}^{k+1}$, where $\alpha\in(0,2)$ is a relaxation factor. This leads to the classical Generalized ADMM (G-ADMM, [8]):

$$\begin{cases}\boldsymbol{x}^{k+1}\in\arg\min\limits_{\boldsymbol{x}\in\mathcal{X}}\mathcal{L}_\beta(\boldsymbol{x},\boldsymbol{y}^k,\boldsymbol{\lambda}^k),\\ \boldsymbol{y}^{k+1}\in\arg\min\limits_{\boldsymbol{y}\in\mathcal{Y}}g(\boldsymbol{y})-(\boldsymbol{\lambda}^k)^{\mathsf{T}}B\boldsymbol{y}+\dfrac{\beta}{2}\left\|\chi^{k+1}+B\boldsymbol{y}-\boldsymbol{b}\right\|^2,\\ \boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^k-\beta\left(\chi^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right).\end{cases}$$

Clearly, G-ADMM with $\alpha=1$ would reduce to the standard ADMM with unit dual stepsize. ADMM using some additional proximal terms in their subproblems is also called G-ADMM. For instance, the $\boldsymbol{x}$-subproblem in [9] was proposed as $\min\limits_{\boldsymbol{x}\in\mathcal{X}}\mathcal{L}_\beta(\boldsymbol{x},\boldsymbol{y}^k,\boldsymbol{\lambda}^k)+\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}^k\|_G^2$, where G is a symmetric positive definite matrix. For more recent G-ADMMs using possibly indefinite proximal terms, one may refer to the references [21, 30].

For convergence rate of ADMM, it is well-known that most of deterministic ADMM algorithms [1, 6, 7, 9, 15, 16, 18, 19, 21, 28, 31, 32] enjoy a global $\mathcal{O}(1/T)$ ergodic convergence rate for separable convex optimization, where T is the iteration number. Under the assumption that the subdifferential of each component objective function is piecewise linear, Yang-Han [34] established linear convergence rate of ADMM for two-block separable convex optimization. Assuming that an error bound condition holds, the dual stepsize is sufficiently small and the coefficient matrices in the equality constraint have full column ranks, Hong-Luo [20] showed a linear convergence rate of their multi-block ADMM. Zhang et al. [36] developed a majorized ADMM with indefinite proximal terms (iPADMM) for a class of composite convex optimization problems, and the authors analyzed the convergence of this iPADMM with a linear convergence rate under a local error bound condition. Moreover, Chang et al. [6] proposed a linearized symmetric ADMM with indefinite proximal regularization and optimal proximal parameter for solving the multi-block separable convex optimization. More recently, Yuan-Zeng-Zhang [33] showed that the local linear convergence of ADMM can be guaranteed by a partial error bound condition. For more details about linear convergence rate under strongly convexity assumption, we refer interested readers to [3, 5, 13, 23, 24] and the references therein.

2 Preliminaries

2.1 Notations and assumptions

Let $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times l}$ be the sets of real numbers, n dimensional real column vectors, and $n\times l$ dimensional real matrices, respectively. The $\boldsymbol{I}$ and $\mathbf{0}$ denote the identity matrix and the zero matrix/vector, respectively. For any symmetric matrices A and B of the same dimension, $A\succ B$ ($A\succeq B$) means $A-B$ is a positive definite (semidefinite) matrix. For any symmetric matrix G, define $\|\boldsymbol{x}\|_G^2:=\boldsymbol{x}^\top G\boldsymbol{x}$ and $\|\boldsymbol{x}\|_G:=\sqrt{\boldsymbol{x}^\top G\boldsymbol{x}}$ if $G\succeq\mathbf{0}$. We use $\|\cdot\|$ to denote the standard Euclidean norm equipped with inner product $\langle\cdot,\cdot\rangle$, $\nabla f(\boldsymbol{x})$ to represent the gradient of f at $\boldsymbol{x}$, and $\mathbb{E}[\cdot]$ to denote mathematical expectation of a random variable. We also define

$$\boldsymbol{w}=\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{\lambda}\end{pmatrix},\quad \mathcal{J}(\boldsymbol{w})=\begin{pmatrix}-A^\top\boldsymbol{\lambda}\\ -B^\top\boldsymbol{\lambda}\\ A\boldsymbol{x}+B\boldsymbol{y}-\boldsymbol{b}\end{pmatrix}, \tag{2.1}$$

and

$$\boldsymbol{w}^k=\begin{pmatrix}\boldsymbol{x}^k\\ \boldsymbol{y}^k\\ \boldsymbol{\lambda}^k\end{pmatrix},\quad \mathcal{J}(\boldsymbol{w}^k)=\begin{pmatrix}-A^\top\boldsymbol{\lambda}^k\\ -B^\top\boldsymbol{\lambda}^k\\ A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\end{pmatrix}. \tag{2.2}$$

For convenience of analysis, we simply denote $F(\boldsymbol{w})=f(\boldsymbol{x})+g(\boldsymbol{y})$.

We make the following two assumptions:

Assumption 2.1. *The primal-dual solution set Ω^* of problem (1.1) is nonempty, and the problem $\min\limits_{\boldsymbol{y}\in\mathcal{Y}}\{g(\boldsymbol{y})+\frac{1}{2}\boldsymbol{y}^\top B^\top B\boldsymbol{y}+\boldsymbol{z}^\top\boldsymbol{y}\}$ has a minimizer for any $\boldsymbol{z}\in\mathbb{R}^{n_2}$.*

Assumption 2.2. *For any $\mathcal{H}\succ\mathbf{0}$, there exists a constant $\nu>0$ such that the gradients ∇f_j satisfy the Lipschitz condition*

$$\|\nabla f_j(\boldsymbol{x}_1)-\nabla f_j(\boldsymbol{x}_2)\|_{\mathcal{H}^{-1}}\le\nu\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_{\mathcal{H}} \tag{2.3}$$

for every $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}$ and $j=1,2,\cdots,N$.

The first assumption is a basic assumption to ensure the solvability of the problem. Under Assumption 2.2, it holds that for every $\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}$, we have

$$f(\boldsymbol{x}_1)\le f(\boldsymbol{x}_2)+\langle\nabla f(\boldsymbol{x}_2),\boldsymbol{x}_1-\boldsymbol{x}_2\rangle+\frac{\nu}{2}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_{\mathcal{H}}^2.$$

2.2 Variational characterization of (1.1)

Denote $\Omega=\mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n$. It's well-known in convex optimization that any saddle-point of the Lagrangian (1.3) corresponds to a primal-dual solution of problem (1.1). A point $\boldsymbol{w}^*:=(\boldsymbol{x}^*;\boldsymbol{y}^*;\boldsymbol{\lambda}^*)\in\Omega$ is called a saddle-point of $\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda})$ if it satisfies

$$\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^*,\boldsymbol{\lambda})\le\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^*,\boldsymbol{\lambda}^*)\le\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda}^*),\quad\forall\boldsymbol{w}\in\Omega, \tag{2.4}$$

which is equivalent to

$$\begin{cases} f(\boldsymbol{x})-f(\boldsymbol{x}^*)+(\boldsymbol{x}-\boldsymbol{x}^*)^{\mathsf{T}}\left(-A^{\mathsf{T}}\boldsymbol{\lambda}^*\right)\geq 0, \\ g(\boldsymbol{y})-g(\boldsymbol{y}^*)+(\boldsymbol{y}-\boldsymbol{y}^*)^{\mathsf{T}}\left(-B^{\mathsf{T}}\boldsymbol{\lambda}^*\right)\geq 0, \\ A\boldsymbol{x}^*+B\boldsymbol{y}^*-\boldsymbol{b}=\boldsymbol{0}. \end{cases}$$

Rewriting these inequalities as a more compact form, it gives

$$F(\boldsymbol{w})-F(\boldsymbol{w}^*)+(\boldsymbol{w}-\boldsymbol{w}^*)^{\mathsf{T}}\mathcal{J}(\boldsymbol{w}^*)\geq 0,\quad \forall \boldsymbol{w}\in\Omega. \tag{2.5}$$

Notice that the affine mapping $\mathcal{J}(\cdot)$ is skew-symmetric. So, we have

$$(\boldsymbol{w}-\overline{\boldsymbol{w}})^{\mathsf{T}}[\mathcal{J}(\boldsymbol{w})-\mathcal{J}(\overline{\boldsymbol{w}})]\equiv 0,\quad \forall \boldsymbol{w},\overline{\boldsymbol{w}}\in\Omega. \tag{2.6}$$

Hence, (2.5) is also equivalent to

$$F(\boldsymbol{w})-F(\boldsymbol{w}^*)+(\boldsymbol{w}-\boldsymbol{w}^*)^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\geq 0,\quad \forall \boldsymbol{w}\in\Omega. \tag{2.7}$$

The above discussion shows that the saddle-point $\boldsymbol{w}^*$ can be also characterized by the variational inequality (2.7).

3 The proposed algorithm

Motivated from the stochastic AS-ADMM developed in [3] and the deterministic GS-ADMM proposed in [1], we now propose a Symmetric Accelerated Stochastic ADMM (SAS-ADMM, i.e., Algorithm 1), which has the similarly dual stepsize region Δ to that of GS-ADMM defined as

$$\Delta=\left\{(\tau,s)\mid \tau+s>0,\ \tau\leq 1,\ -\tau^2-s^2-\tau s+\tau+s+1\geq 0\right\}. \tag{3.1}$$

The main features of SAS-ADMM are summarized as follows:

(i) SAS-ADMM has many analogous advantages to AS-ADMM developed in [3]. Specifically, SAS-ADMM has low memory requirement since there is no need to save previous stochastic gradients and iterates. The subroutine **xsub** is a variant of deterministic accelerated gradient method where the full gradient is replaced by a stochastic gradient. In addition, users have the flexibility of choosing a zero mean random vector $\boldsymbol{e}_t$ to reduce the variance of $\widehat{\boldsymbol{g}}_t$. A simple choice is $\boldsymbol{e}_t=\boldsymbol{0}$, while faster convergence is observed in the numerical experiments when a variance reduction technique is employed (see (5.3) in Section 5). Under our blanket assumption that g is a proper convex function, the proximal $\boldsymbol{y}$-subproblem is solvable. Also, under the assumption that the projection onto the constraint set $\mathcal{X}$ is simple, the iterations in subroutine **xsub** can be performed efficiently when both $\mathcal{M}_k$ and $\mathcal{H}$ are multiples of identity matrix[†].

[†]As explained in Remark 4.1 and experiments, both $\mathcal{M}_k$ and $\mathcal{H}$ could be chosen as multiples of identity matrix. So the $\check{\boldsymbol{x}}_{t+1}$-subproblem is equivalent to a projection onto $\mathcal{X}$.

Algorithm 1 Symmetric accelerated stochastic ADMM (SAS-ADMM)

Parameters: $\beta>0$, $\mathcal{H}\succ\mathbf{0}$, $L\succeq\mathbf{0}$ and $(\tau,s)\in\Delta$ given by (3.1).
Initialization: $(\boldsymbol{x}^0,\boldsymbol{y}^0,\boldsymbol{\lambda}^0)\in\mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n$, $\check{\boldsymbol{x}}^0=\boldsymbol{x}^0$.
For $k=0,1,\cdots$
Choose $m_k>0$, $\eta_k>0$ and $\mathcal{M}_k$ such that $\mathcal{M}_k-\beta A^{\mathsf{T}}A\succeq\mathbf{0}$.
$\boldsymbol{h}^k:=-A^{\mathsf{T}}\left[\boldsymbol{\lambda}^k-\beta(A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b})\right]$.
$(\boldsymbol{x}^{k+1},\check{\boldsymbol{x}}^{k+1})=$ **xsub** $(\boldsymbol{x}^k,\check{\boldsymbol{x}}^k,\boldsymbol{h}^k)$.
$\boldsymbol{\lambda}^{k+\frac{1}{2}}=\boldsymbol{\lambda}^k-\tau\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}\right)$.
$\boldsymbol{y}^{k+1}\in\underset{\boldsymbol{y}\in\mathcal{Y}}{\arg\min}\mathcal{L}_\beta\left(\boldsymbol{x}^{k+1},\boldsymbol{y},\boldsymbol{\lambda}^{k+\frac{1}{2}}\right)+\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^k\right\|_L^2$.
$\boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^{k+\frac{1}{2}}-s\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right)$.
end

$(\boldsymbol{x}^+,\check{\boldsymbol{x}}^+)=$ **xsub** $(\boldsymbol{x}_1,\ \check{\boldsymbol{x}}_1,\ \boldsymbol{h})$.
For $t=1,2,\cdots,m_k$
Randomly select $\xi_t\in\{1,2,\cdots,N\}$ with uniform probability.
$\beta_t=2/(t+1)$, $\gamma_t=2/(t\eta_k)$, $\widehat{\boldsymbol{x}}_t=\beta_t\check{\boldsymbol{x}}_t+(1-\beta_t)\boldsymbol{x}_t$.
$\boldsymbol{d}_t=\widehat{\boldsymbol{g}}_t+\boldsymbol{e}_t$, where $\widehat{\boldsymbol{g}}_t=\nabla f_{\xi_t}(\widehat{\boldsymbol{x}}_t)$ and $\boldsymbol{e}_t$ is a random vector satisfying $\mathbb{E}[\boldsymbol{e}_t]=\mathbf{0}$.
$\check{\boldsymbol{x}}_{t+1}=\arg\min\left\{\langle\boldsymbol{d}_t+\boldsymbol{h},\boldsymbol{x}\rangle+\frac{\gamma_t}{2}\|\boldsymbol{x}-\check{\boldsymbol{x}}_t\|_{\mathcal{H}}^2+\frac{1}{2}\left\|\boldsymbol{x}-\boldsymbol{x}^k\right\|_{\mathcal{M}_k}^2:\boldsymbol{x}\in\mathcal{X}\right\}$.
$\boldsymbol{x}_{t+1}=\beta_t\check{\boldsymbol{x}}_{t+1}+(1-\beta_t)\boldsymbol{x}_t$.
end
Return $(\boldsymbol{x}^+,\check{\boldsymbol{x}}^+)=(\boldsymbol{x}_{m_k+1},\check{\boldsymbol{x}}_{m_k+1})$.

(ii) Unlike the classical ADMM and AS-ADMM [3], the dual variable of SAS-ADMM is symmetrically updated twice and allowed to use the large stepsize region (3.1). SAS-ADMM will reduce to the aforementioned PRSM if the $\boldsymbol{x}$-subproblem is solved deterministically as in S-ADMM, $L=\mathbf{0}$ and $(\tau,s)=(1,1)\in\Delta$. When the stepsize $\tau=0$ and $L=\mathbf{0}$, SAS-ADMM reduces to AS-ADMM with stepsize $s\in(0,\frac{1+\sqrt{5}}{2}]$ (the half-open interval). Compared with the standard stepsize region $(0,\frac{1+\sqrt{5}}{2})$ for the dual variable of ADMM, this symmetric updates of dual variable is more balanced, flexible and often lead better numerical performance.

(iii) If $m_k=1,N=1$, then SAS-ADMM degrades to a linearized symmetric ADMM. When $m_k>1,N=1$, SAS-ADMM is a multi-step deterministic inexact symmetric ADMM. Hence, the convergence properties developed in this paper also apply to these deterministic algorithms as special cases. Moreover, by taking $L=\gamma\boldsymbol{I}-\beta B^{\mathsf{T}}B$ for some

$\gamma>0$, the $\boldsymbol{y}$-subproblem would become the following proximal mapping problem:

$$\mathrm{prox}_{\gamma}^{g}(\boldsymbol{y}_c^k):=\arg\min_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{y})+\frac{\gamma}{2}\left\|\boldsymbol{y}-\boldsymbol{y}_c^k\right\|^2, \tag{3.2}$$

where $\boldsymbol{y}_c^k=\boldsymbol{y}^k-\beta B^{\mathsf{T}}(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}-\boldsymbol{\lambda}^{k+\frac{1}{2}}/\beta)/\gamma$. In this case, the Assumption 2.1 is not required since strong convexity of the $\boldsymbol{y}$-subproblem implies a unique global solution and a closed-form solution may exist when g has certain structure.

(iv) With the aid of variational analysis, we show that SAS-ADMM has the worst-case $\mathcal{O}(1/T)$ ergodic convergence rate in terms of the expectation of both the objective value gap and the constraint violation, where T is the number of the outer iterations. Preliminary experiments and results show that SAS-ADMM performs competitively well and often slightly better than AS-ADMM [3] for solving a family of separable convex optimization problems arising from big-data applications.

4 Convergence analysis

To establish the convergence of Algorithm 1, we first need the following lemma about the iterates generated by the **xsub** routine in Algorithm 1. The lemma was given in [3] and thus we omit its proof.

Lemma 4.1. [3, Lemma 3.2] *Let $\delta_t=\nabla f(\widehat{\boldsymbol{x}}_t)-\boldsymbol{d}_t$. Suppose $\eta_k\in(0,1/\nu)$ and Assumption 2.2 holds. Then, the iterates generated by Algorithm 1 satisfy*

$$f(\boldsymbol{x})-f(\boldsymbol{x}^{k+1})+\left\langle \boldsymbol{x}-\boldsymbol{x}^{k+1},-A^{\mathsf{T}}\widetilde{\boldsymbol{\lambda}}^k\right\rangle \geq \left\langle \boldsymbol{x}^{k+1}-\boldsymbol{x},\mathcal{D}_k(\boldsymbol{x}^{k+1}-\boldsymbol{x}^k)\right\rangle+\zeta^k \tag{4.1}$$

for all $\boldsymbol{x}\in\mathcal{X}$, where

$$\widetilde{\boldsymbol{\lambda}}^k=\boldsymbol{\lambda}^k-\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}\right),\quad \mathcal{D}_k=\mathcal{M}_k-\beta A^{\mathsf{T}}A \tag{4.2}$$

and

$$\begin{aligned}\zeta^k=\frac{2}{m_k(m_k+1)}\Bigg[&\frac{1}{\eta_k}\left(\left\|\boldsymbol{x}-\check{\boldsymbol{x}}^{k+1}\right\|_{\mathcal{H}}^2-\left\|\boldsymbol{x}-\check{\boldsymbol{x}}^{k}\right\|_{\mathcal{H}}^2\right)\\&-\sum_{t=1}^{m_k}t\langle\delta_t,\check{\boldsymbol{x}}_t-\boldsymbol{x}\rangle-\frac{\eta_k}{4(1-\eta_k\nu)}\sum_{t=1}^{m_k}t^2\left\|\delta_t\right\|_{\mathcal{H}^{-1}}^2\Bigg].\end{aligned} \tag{4.3}$$

Based on the above lemma, we can immediately establish the following result.

Lemma 4.2. *Suppose $\eta_k\in(0,1/\nu)$. Then, the iterates generated by Algorithm 1 satisfy*

$$F(\boldsymbol{w})-F(\widetilde{\boldsymbol{w}}^k)+\left\langle \boldsymbol{w}-\widetilde{\boldsymbol{w}}^k,\mathcal{J}(\boldsymbol{w})\right\rangle \geq(\boldsymbol{w}-\widetilde{\boldsymbol{w}}^k)^{\mathsf{T}}Q_k(\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k)+\zeta^k \tag{4.4}$$

for all $\boldsymbol{w}\in\Omega$, *where* ζ^k *and* $\widetilde{\boldsymbol{\lambda}}^k$ *are defined in* (4.3) *and* (4.2),

$$\widetilde{\boldsymbol{w}}^k=\begin{pmatrix}\widetilde{\boldsymbol{x}}^k\\ \widetilde{\boldsymbol{y}}^k\\ \widetilde{\boldsymbol{\lambda}}^k\end{pmatrix}=\begin{pmatrix}\boldsymbol{x}^{k+1}\\ \boldsymbol{y}^{k+1}\\ \widetilde{\boldsymbol{\lambda}}^k\end{pmatrix}\quad\text{and}\quad Q_k=\begin{bmatrix}\mathcal{D}_k & & \\ & L+\beta B^\top B & -\tau B^\top\\ & -B & \frac{1}{\beta}\boldsymbol{I}\end{bmatrix}. \tag{4.5}$$

Proof. By the first-order optimality condition of the $\boldsymbol{y}$-subproblem, we have

$$g(\boldsymbol{y})-g(\boldsymbol{y}^{k+1})+\left\langle \boldsymbol{y}-\boldsymbol{y}^{k+1},\boldsymbol{p}_k\right\rangle\geq 0,\quad\forall\boldsymbol{y}\in\mathcal{Y}, \tag{4.6}$$

where $\boldsymbol{p}_k$ is the gradient of the smooth terms in the objective function of the $\boldsymbol{y}$-subproblem:

$$\begin{aligned}\boldsymbol{p}_k&=-B^\top\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B^\top\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right)+L(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k)\\ &=-B^\top\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B^\top\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k}-\boldsymbol{b}\right)+\left[L+\beta B^\top B\right]\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right)\\ &=-B^\top\boldsymbol{\lambda}^{k+\frac{1}{2}}+B^\top\left(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k\right)+\left[L+\beta B^\top B\right]\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right)\\ &=-B^\top\widetilde{\boldsymbol{\lambda}}^k+\tau B^\top(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k)+\left[L+\beta B^\top B\right]\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right).\end{aligned}$$

The above last equality uses the following relation

$$\boldsymbol{\lambda}^{k+\frac{1}{2}}=\boldsymbol{\lambda}^k-\tau(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k). \tag{4.7}$$

By the definition of $\widetilde{\boldsymbol{\lambda}}^k$, we have

$$\left(A\widetilde{\boldsymbol{x}}^k+B\widetilde{\boldsymbol{y}}^k-\boldsymbol{b}\right)-B\left(\widetilde{\boldsymbol{y}}^k-\boldsymbol{y}^k\right)+\frac{1}{\beta}\left(\widetilde{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k\right)=\boldsymbol{0}. \tag{4.8}$$

Taking inner product of the above equality with $\boldsymbol{\lambda}-\widetilde{\boldsymbol{\lambda}}^k$, we get

$$\left\langle\boldsymbol{\lambda}-\widetilde{\boldsymbol{\lambda}}^k,A\widetilde{\boldsymbol{x}}^k+B\widetilde{\boldsymbol{y}}^k-\boldsymbol{b}\right\rangle=\left\langle\boldsymbol{\lambda}-\widetilde{\boldsymbol{\lambda}}^k,-B\left(\boldsymbol{y}^k-\widetilde{\boldsymbol{y}}^k\right)+\frac{1}{\beta}\left(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k\right)\right\rangle. \tag{4.9}$$

Then, the inequality (4.4) is achieved by combining (4.1), (4.6), (4.9) together with the property in (2.6). □

4.1 More technical results

We show the following corollary and lemmas for establishing the main convergence theorem of our Algorithm 1.

Corollary 4.1. *Suppose $\eta_k \in (0, 1/\nu)$. Then, the iterates generated by Algorithm 1 satisfy*

$$\begin{aligned} &F(\boldsymbol{w}) - F(\widetilde{\boldsymbol{w}}^k) + (\boldsymbol{w} - \widetilde{\boldsymbol{w}}^k)^{\mathsf{T}} \mathcal{J}(\boldsymbol{w}) \\ \geq &\frac{1}{2}\left\{ \left\| \boldsymbol{w} - \boldsymbol{w}^{k+1} \right\|_{\widetilde{Q}_k}^2 - \left\| \boldsymbol{w} - \boldsymbol{w}^k \right\|_{\widetilde{Q}_k}^2 + \left\| \boldsymbol{w}^k - \widetilde{\boldsymbol{w}}^k \right\|_{\widetilde{G}_k}^2 \right\} + \zeta^k \end{aligned} \tag{4.10}$$

for all $\boldsymbol{w} \in \Omega$, where ζ^k is defined in (4.3) and

$$\begin{aligned} \widetilde{Q}_k &= \begin{bmatrix} \mathcal{D}_k & & \\ & L + \left(1 - \frac{\tau s}{\tau + s}\right) \beta B^{\mathsf{T}} B & -\frac{\tau}{\tau + s} B^{\mathsf{T}} \\ & -\frac{\tau}{\tau + s} B & \frac{1}{\beta(\tau + s)} \boldsymbol{I} \end{bmatrix}, \\ \widetilde{G}_k &= \begin{bmatrix} \mathcal{D}_k & & \\ & L + (1 - s)\beta B^{\mathsf{T}} B & (s-1) B^{\mathsf{T}} \\ & (s-1) B & \frac{2 - \tau - s}{\beta} \boldsymbol{I} \end{bmatrix}. \end{aligned} \tag{4.11}$$

Proof. By (4.7) and the way of generating $\boldsymbol{\lambda}^{k+1}$, we have

$$-s\beta B\left(\boldsymbol{y}^k - \widetilde{\boldsymbol{y}}^k\right) + (\tau + s)\left(\boldsymbol{\lambda}^k - \widetilde{\boldsymbol{\lambda}}^k\right) = \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^{k+1}, \tag{4.12}$$

which, by the definition of $\widetilde{\boldsymbol{w}}^k$ in (4.5), further shows

$$\boldsymbol{w}^k - \boldsymbol{w}^{k+1} = P\left(\boldsymbol{w}^k - \widetilde{\boldsymbol{w}}^k\right) \quad \text{with} \quad P = \begin{bmatrix} \boldsymbol{I} & & \\ & \boldsymbol{I} & \\ & -s\beta B & (\tau + s)\boldsymbol{I} \end{bmatrix}. \tag{4.13}$$

Hence, the relation $Q_k(\boldsymbol{w}^k - \widetilde{\boldsymbol{w}}^k) = Q_k P^{-1}(\boldsymbol{w}^k - \boldsymbol{w}^{k+1})$ holds and

$$Q_k P^{-1} = \begin{bmatrix} \mathcal{D}_k & & \\ & L + \left(1 - \frac{\tau s}{\tau + s}\right) \beta B^{\mathsf{T}} B & -\frac{\tau}{\tau + s} B^{\mathsf{T}} \\ & -\frac{\tau}{\tau + s} B & \frac{1}{\beta(\tau + s)} \boldsymbol{I} \end{bmatrix} = \widetilde{Q}_k.$$

For any $\boldsymbol{w} \in \Omega$, it follows from (4.4) and the above relation that

$$\begin{aligned} &F(\boldsymbol{w}) - F(\widetilde{\boldsymbol{w}}^k) + (\boldsymbol{w} - \widetilde{\boldsymbol{w}}^k)^{\mathsf{T}} \mathcal{J}(\boldsymbol{w}) \\ \geq &\zeta^k + (\boldsymbol{w} - \widetilde{\boldsymbol{w}}^k)^{\mathsf{T}} \widetilde{Q}_k (\boldsymbol{w}^k - \boldsymbol{w}^{k+1}) \\ = &\zeta^k + \frac{1}{2}\left\{ \left\| \boldsymbol{w} - \boldsymbol{w}^{k+1} \right\|_{\widetilde{Q}_k}^2 - \left\| \boldsymbol{w} - \boldsymbol{w}^k \right\|_{\widetilde{Q}_k}^2 + \left\| \boldsymbol{w}^k - \widetilde{\boldsymbol{w}}^k \right\|_{\widetilde{Q}_k}^2 - \left\| \boldsymbol{w}^{k+1} - \widetilde{\boldsymbol{w}}^k \right\|_{\widetilde{Q}_k}^2 \right\}, \end{aligned} \tag{4.14}$$

where the equality uses the identity

$$2(a-b)^{\mathsf{T}} \widetilde{Q}_k (c-d) = \|a-d\|_{\widetilde{Q}_k}^2 - \|a-c\|_{\widetilde{Q}_k}^2 + \|c-b\|_{\widetilde{Q}_k}^2 - \|b-d\|_{\widetilde{Q}_k}^2$$

with specifications $a:=\boldsymbol{w}$, $b:=\widetilde{\boldsymbol{w}}^k$, $c:=\boldsymbol{w}^k$, $d:=\boldsymbol{w}^{k+1}$.

Now, by (4.13) again, we deduce

$$\begin{aligned}\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{Q}_k}^2-\left\|\boldsymbol{w}^{k+1}-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{Q}_k}^2&=\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{Q}_k}^2-\left\|\boldsymbol{w}^{k+1}-\boldsymbol{w}^k+\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{Q}_k}^2\\&=\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{Q}_k}^2-\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k-P(\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k)\right\|_{\widetilde{Q}_k}^2\\&=\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{G}_k}^2,\end{aligned}$$

where we uses the relation $\widetilde{G}_k=P^{\mathsf{T}}\widetilde{Q}_k+\widetilde{Q}_kP-P^{\mathsf{T}}\widetilde{Q}_kP$ to obtain the last equality. Then, (4.10) follows from (4.14). □

In the above Corollary 4.1 and its proof, since $\widetilde{Q}_k$ is not necessarily positive semidefinite for any parameter τ, we abuse the notation $\|\boldsymbol{w}^k\|_{\widetilde{Q}_k}^2:=(\boldsymbol{w}^k)^{\mathsf{T}}\widetilde{Q}_k\boldsymbol{w}^k$. Next, we provide a sufficient condition to ensure the positive semidefiniteness of $\widetilde{Q}_k$.

Lemma 4.3. *Let $L\succeq(\tau-1)\beta B^{\mathsf{T}}B$. Then, the matrix $\widetilde{Q}_k$ given by* (4.11) *is symmetric positive semidefinite for any $(\tau,s)\in\Delta$.*

Proof. Clearly, we just need to check the lower-upper 2-by-2 block of $\widetilde{Q}_k$, i.e.,

$$\begin{aligned}\widetilde{Q}_k^L&=\begin{bmatrix}L+\left(1-\frac{\tau s}{\tau+s}\right)\beta B^{\mathsf{T}}B & -\frac{\tau}{\tau+s}B^{\mathsf{T}}\\ -\frac{\tau}{\tau+s}B & \frac{1}{\beta(\tau+s)}\boldsymbol{I}\end{bmatrix}\\&\succeq\begin{bmatrix}\left(\tau-\frac{\tau s}{\tau+s}\right)\beta B^{\mathsf{T}}B & -\frac{\tau}{\tau+s}B^{\mathsf{T}}\\ -\frac{\tau}{\tau+s}B & \frac{1}{\beta(\tau+s)}\boldsymbol{I}\end{bmatrix}\\&=\begin{bmatrix}\beta^{\frac{1}{2}}B & \\ & \beta^{-\frac{1}{2}}\boldsymbol{I}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\frac{\tau^2}{\tau+s}\boldsymbol{I} & -\frac{\tau}{\tau+s}\boldsymbol{I}\\ -\frac{\tau}{\tau+s}\boldsymbol{I} & \frac{1}{\tau+s}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\beta^{\frac{1}{2}}B & \\ & \beta^{-\frac{1}{2}}\boldsymbol{I}\end{bmatrix}\end{aligned}$$

is positive semidefinite. Notice that

$$\begin{bmatrix}\boldsymbol{I} & \\ \tau\boldsymbol{I} & \boldsymbol{I}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\frac{\tau^2}{\tau+s}\boldsymbol{I} & -\frac{\tau}{\tau+s}\boldsymbol{I}\\ -\frac{\tau}{\tau+s}\boldsymbol{I} & \frac{1}{\tau+s}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \\ \tau\boldsymbol{I} & \boldsymbol{I}\end{bmatrix}=\begin{bmatrix}\boldsymbol{0} & \\ & \frac{1}{\tau+s}\boldsymbol{I}\end{bmatrix}.$$

So, $\widetilde{Q}_k^L$ is positive semidefinite since $\tau+s>0$ for any $(\tau,s)\in\Delta$. □

To show the global convergence of Algorithm 1, we need to further establish a useful lower bound on the term $\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{G}_k}^2$, since $\widetilde{G}_k$ is not necessarily positive definite for any $(\tau,s)\in\Delta$. In the following lemma, we assume $L\succeq\boldsymbol{0}$, which implies $L\succeq(\tau-1)\beta B^{\mathsf{T}}B$ since $\tau\leq1$, and as a consequence, Lemma 4.3 holds.

Lemma 4.4. *Let $L\succeq\mathbf{0}$. Then, for any $(\tau,s)\in\Delta$ defined in (3.1), we have*

$$\begin{aligned}\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|^2_{\widetilde{G}_k}\geq&\left\|\boldsymbol{x}^k-\boldsymbol{x}^{k+1}\right\|^2_{\mathcal{D}_k}+\omega_0\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^2\\&+\omega_1\left(\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^2-\left\|A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right\|^2\right)\\&+\omega_2\left(\left\|\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right\|^2_L-\left\|\boldsymbol{y}^{k-1}-\boldsymbol{y}^k\right\|^2_L\right),\end{aligned}\tag{4.15}$$

where $\omega_i\geq0$, $i=0,1,2$, are given as

$$\omega_0=\left(2-\tau-s-\frac{(1-s)^2}{1+\tau}\right)\beta,\quad\omega_1=\frac{(1-s)^2}{1+\tau}\beta\quad\textit{and}\quad\omega_2=\frac{1-\tau}{1+\tau}.\tag{4.16}$$

Proof. For any $(\tau,s)\in\Omega$, we have $1+\tau>0$. By the structure of $\widetilde{G}_k$ in (4.11), $L\succeq\mathbf{0}$ and (4.8), we have

$$\begin{aligned}&\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|^2_{\widetilde{G}_k}\\=&\left\|\boldsymbol{x}^k-\boldsymbol{x}^{k+1}\right\|^2_{\mathcal{D}_k}+\left\|\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right\|^2_{L+(1-s)\beta B^\mathsf{T}B}\\&+2(s-1)\left(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k\right)^\mathsf{T}B\left(\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right)+\frac{2-\tau-s}{\beta}\left\|\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k\right\|^2\\\geq&\left\|\boldsymbol{x}^k-\boldsymbol{x}^{k+1}\right\|^2_{\mathcal{D}_k}+(2-\tau-s)\beta\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^2\\&+(1-\tau)\beta\left\|B\left(\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right)\right\|^2+2(1-\tau)\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-b\right)^\mathsf{T}B\left(\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right).\end{aligned}\tag{4.17}$$

We now estimate the last crossing term in (4.17). Taking $\boldsymbol{y}=\boldsymbol{y}^k$ in the first-order optimality condition (4.6) yields

$$g(\boldsymbol{y}^k)-g(\boldsymbol{y}^{k+1})+\left\langle\boldsymbol{y}^k-\boldsymbol{y}^{k+1},-B^\mathsf{T}\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B^\mathsf{T}(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b})+L(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k)\right\rangle\geq0.$$

Similarly, letting $\boldsymbol{y}=\boldsymbol{y}^{k+1}$ in the first-order optimality condition of the $\boldsymbol{y}$-subproblem at the $(k-1)$-th iteration gives

$$g(\boldsymbol{y}^{k+1})-g(\boldsymbol{y}^k)+\left\langle\boldsymbol{y}^{k+1}-\boldsymbol{y}^k,-B^\mathsf{T}\boldsymbol{\lambda}^{k-\frac{1}{2}}+\beta B^\mathsf{T}\left(A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right)+L(\boldsymbol{y}^k-\boldsymbol{y}^{k-1})\right\rangle\geq0.$$

Summing up the above two inequalities together with the relation

$$\boldsymbol{\lambda}^{k-\frac{1}{2}}-\boldsymbol{\lambda}^{k+\frac{1}{2}}=\tau\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right)+s\beta\left(A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right)+\tau\beta B(\boldsymbol{y}^k-\boldsymbol{y}^{k+1}),$$

and noticing that $1+\tau>0$, we obtain

$$\begin{aligned}
&\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right)^{\mathsf{T}}B\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)\\
\geq&\frac{1-s}{1+\tau}\left(A\boldsymbol{x}^{k}+B\boldsymbol{y}^{k}-\boldsymbol{b}\right)^{\mathsf{T}}B\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)-\frac{\tau}{1+\tau}\left\|B(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1})\right\|^{2}\\
&+\frac{1}{\beta(1+\tau)}\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)^{\mathsf{T}}L\left[\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)-\left(\boldsymbol{y}^{k-1}-\boldsymbol{y}^{k}\right)\right]\\
\geq&\frac{1-s}{1+\tau}\left(A\boldsymbol{x}^{k}+B\boldsymbol{y}^{k}-\boldsymbol{b}\right)^{\mathsf{T}}B\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)-\frac{\tau}{1+\tau}\left\|B(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1})\right\|^{2}\\
&+\frac{1}{2\beta(1+\tau)}\left(\left\|\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right\|_{L}^{2}-\left\|\boldsymbol{y}^{k-1}-\boldsymbol{y}^{k}\right\|_{L}^{2}\right).
\end{aligned}\tag{4.18}$$

Then, combining (4.17) and (4.18) we have

$$\begin{aligned}
\left\|\boldsymbol{w}^{k}-\widetilde{\boldsymbol{w}}^{k}\right\|_{\widetilde{G}_{k}}^{2}\geq&\left\|\boldsymbol{x}^{k}-\boldsymbol{x}^{k+1}\right\|_{\mathcal{D}_{k}}^{2}+(2-\tau-s)\beta\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^{2}\\
&+\frac{2\beta}{1+\tau}(1-\tau)(1-s)\left(A\boldsymbol{x}^{k}+B\boldsymbol{y}^{k}-\boldsymbol{b}\right)^{\mathsf{T}}B\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)\\
&+(1-\tau)\beta\left\|B\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)\right\|^{2}-\frac{2\tau(1-\tau)}{1+\tau}\beta\left\|B\left(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right)\right\|^{2}\\
&+\frac{1-\tau}{1+\tau}\left(\left\|\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right\|_{L}^{2}-\left\|\boldsymbol{y}^{k-1}-\boldsymbol{y}^{k}\right\|_{L}^{2}\right)\\
\geq&\left\|\boldsymbol{x}^{k}-\boldsymbol{x}^{k+1}\right\|_{\mathcal{D}_{k}}^{2}+\left(2-\tau-s-\frac{(1-s)^{2}}{1+\tau}\right)\beta\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^{2}\\
&+\frac{(1-s)^{2}}{1+\tau}\beta\left(\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^{2}-\left\|A\boldsymbol{x}^{k}+B\boldsymbol{y}^{k}-\boldsymbol{b}\right\|^{2}\right)\\
&+\frac{1-\tau}{1+\tau}\left(\left\|\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right\|_{L}^{2}-\left\|\boldsymbol{y}^{k-1}-\boldsymbol{y}^{k}\right\|_{L}^{2}\right)\\
&+\left(1-\tau-\frac{(1-\tau)^{2}}{1+\tau}-\frac{2\tau(1-\tau)}{1+\tau}\right)\beta\left\|B(\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1})\right\|^{2}\\
=&\left\|\boldsymbol{x}^{k}-\boldsymbol{x}^{k+1}\right\|_{\mathcal{D}_{k}}^{2}+\left(2-\tau-s-\frac{(1-s)^{2}}{1+\tau}\right)\beta\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^{2}\\
&+\frac{(1-s)^{2}}{1+\tau}\beta\left(\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^{2}-\left\|A\boldsymbol{x}^{k}+B\boldsymbol{y}^{k}-\boldsymbol{b}\right\|^{2}\right)\\
&+\frac{1-\tau}{1+\tau}\left(\left\|\boldsymbol{y}^{k}-\boldsymbol{y}^{k+1}\right\|_{L}^{2}-\left\|\boldsymbol{y}^{k-1}-\boldsymbol{y}^{k}\right\|_{L}^{2}\right),
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwartz inequality

$$\begin{aligned}&2(1-s)(1-\tau)\left(A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right)^{\mathsf{T}}B\left(\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right)\\ \geq&-(1-s)^2\left\|A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right\|^2-(1-\tau)^2\left\|B\left(\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right)\right\|^2.\end{aligned}$$

So, (4.15) holds with ω_i, $i=0,1,2$, defined as in (4.16). Moreover, for any $(\tau,s)\in\Delta$, we can derive $\omega_i\geq 0$ for $i=0,1,2$. This completes the whole proof. □

4.2 Iteration complexity in expectation

We now analyze the global ergodic convergence and the iteration complexity of Algorithm 1.

Theorem 4.1. *Suppose $L\succeq\mathbf{0}$ and $(\tau,s)\in\Delta$ defined in (3.1). If for some integers $\kappa,T>0$, the following conditions hold for all $k\in[\kappa,\kappa+T]$: (I) $\eta_k\in(0,1/(2\nu)]$ and the sequence $\{\eta_k m_k(m_k+1)\}$ is nondecreasing; (II) $\mathcal{D}_k\succeq\mathcal{D}_{k+1}\succeq\mathbf{0}$ and $\mathbb{E}(\|\boldsymbol{\delta}_t\|^2_{\mathcal{H}^{-1}})\leq\sigma^2$ for some $\sigma>0$, where $\boldsymbol{\delta}_t$ and $\mathcal{D}_k$ are defined in Lemma 4.1. Then, for any $\boldsymbol{w}\in\Omega$, we have*

$$\begin{aligned}&\mathbb{E}\left[F(\boldsymbol{w}_T)-F(\boldsymbol{w})+(\boldsymbol{w}_T-\boldsymbol{w})^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\right]\\ \leq&\frac{1}{2T}\left\{\sigma^2\sum_{k=\kappa}^{\kappa+T}\eta_k m_k+\frac{4}{m_\kappa(m_\kappa+1)\eta_\kappa}\|\boldsymbol{x}-\check{\boldsymbol{x}}^\kappa\|^2_{\mathcal{H}}+\|\boldsymbol{w}-\boldsymbol{w}^\kappa\|^2_{\widetilde{\mathcal{Q}}_\kappa}\right.\\ &\left.+\omega_1\|A\boldsymbol{x}^\kappa+B\boldsymbol{y}^\kappa-\boldsymbol{b}\|^2+\omega_2\left\|\boldsymbol{y}^{\kappa-1}-\boldsymbol{y}^\kappa\right\|^2_L\right\},\end{aligned}\tag{4.19}$$

where $\boldsymbol{w}_T=\frac{1}{T}\sum_{k=\kappa}^{\kappa+T}\widetilde{\boldsymbol{w}}^k$, $\omega_1\geq 0$ and $\omega_2\geq 0$ are defined in (4.16).

Proof. By the assumption, $\mathcal{D}_k\succeq\mathcal{D}_{k+1}\succeq\mathbf{0}$ implies $\widetilde{Q}_k\succeq\widetilde{Q}_{k+1}\succeq\mathbf{0}$ for the matrix $\widetilde{Q}_k$ given in (4.11). Substituting (4.15) into (4.10) and utilizing the relation $\widetilde{Q}_k\succeq\widetilde{Q}_{k+1}$, it follows from Lemma 4.4 that

$$\begin{aligned}&F(\widetilde{\boldsymbol{w}}^k)-F(\boldsymbol{w})+(\widetilde{\boldsymbol{w}}^k-\boldsymbol{w})^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\\ \leq&-\zeta^k+\frac{1}{2}\left\{\left\|\boldsymbol{w}-\boldsymbol{w}^k\right\|^2_{\widetilde{Q}_k}-\left\|\boldsymbol{w}-\boldsymbol{w}^{k+1}\right\|^2_{\widetilde{Q}_{k+1}}\right\}\\ &+\frac{\omega_1}{2}\left(\left\|A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right\|^2-\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^2\right)\\ &+\frac{\omega_2}{2}\left(\left\|\boldsymbol{y}^{k-1}-\boldsymbol{y}^k\right\|^2_L-\left\|\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right\|^2_L\right),\end{aligned}$$

where ω_1,ω_2 are defined in (4.16). Summing the above inequality over k between κ and

$\kappa+T$, we deduce by Lemma 4.3 that

$$\begin{aligned}&\sum_{k=\kappa}^{\kappa+T} F(\widetilde{\boldsymbol{w}}^k)-T\Big\{F(\boldsymbol{w})+(\boldsymbol{w}_T-\boldsymbol{w})^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\Big\}\\ &\leq-\sum_{k=\kappa}^{\kappa+T}\zeta^k+\frac{1}{2}\left\{\|\boldsymbol{w}-\boldsymbol{w}^\kappa\|^2_{\widetilde{Q}_\kappa}+\omega_1\|A\boldsymbol{x}^\kappa+B\boldsymbol{y}^\kappa-\boldsymbol{b}\|^2+\omega_2\left\|\boldsymbol{y}^{\kappa-1}-\boldsymbol{y}^\kappa\right\|^2_L\right\}.\end{aligned}\tag{4.20}$$

Then, it follows from convexity of F and the definition of $\mathbf{w}_T$ that

$$F(\boldsymbol{w}_T)\leq\frac{1}{T}\sum_{k=\kappa}^{\kappa+T}F(\widetilde{\boldsymbol{w}}^k).\tag{4.21}$$

Dividing (4.20) by T and using (4.21), we obtain

$$\begin{aligned}&F(\boldsymbol{w}_T)-F(\boldsymbol{w})+(\boldsymbol{w}_T-\boldsymbol{w})^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\\ &\leq\frac{1}{T}\left[-\sum_{k=\kappa}^{\kappa+T}\zeta^k+\frac{1}{2}\left\{\|\boldsymbol{w}-\boldsymbol{w}^\kappa\|^2_{\widetilde{Q}_\kappa}+\omega_1\|A\boldsymbol{x}^\kappa+B\boldsymbol{y}^\kappa-\boldsymbol{b}\|^2+\omega_2\left\|\boldsymbol{y}^{\kappa-1}-\boldsymbol{y}^\kappa\right\|^2_L\right\}\right].\end{aligned}\tag{4.22}$$

Let us focus on the terms involving ζ^k. By assumption, the sequence $\{m_k(m_k+1)\eta_k\}$ is nondecreasing for $k\in[\kappa,\kappa+T]$ and $\mathcal{H}\succ\mathbf{0}$, thus we have

$$\sum_{k=\kappa}^{\kappa+T}\frac{2}{m_k(m_k+1)\eta_k}\left(\left\|\boldsymbol{x}-\check{\boldsymbol{x}}^k\right\|^2_{\mathcal{H}}-\left\|\boldsymbol{x}-\check{\boldsymbol{x}}^{k+1}\right\|^2_{\mathcal{H}}\right)\leq\frac{2\|\boldsymbol{x}-\check{\boldsymbol{x}}^\kappa\|^2_{\mathcal{H}}}{m_\kappa(m_\kappa+1)\eta_\kappa}.\tag{4.23}$$

Note that

$$\boldsymbol{\delta}_t=\nabla f(\widehat{\boldsymbol{x}}_t)-\boldsymbol{d}_t=\nabla f(\widehat{\boldsymbol{x}}_t)-\nabla f_{\xi_t}(\widehat{\boldsymbol{x}}_t)-\boldsymbol{e}_t$$

only depends on the index ξ_t. So we have $\mathbb{E}[\boldsymbol{\delta}_t]=\mathbf{0}$ since the random variable $\xi_t\in\{1,2,\cdots,N\}$ is chosen with uniform probability and $\mathbb{E}[\boldsymbol{e}_t]=\mathbf{0}$. Also, since $\check{\boldsymbol{x}}_t$ depends on ξ_{t-1}, ξ_{t-2}, $\cdots$, we have $\mathbb{E}[\langle\boldsymbol{\delta}_t,\check{\boldsymbol{x}}_t-\boldsymbol{x}\rangle]=\mathbf{0}$. By the assumption that $\mathbb{E}(\|\boldsymbol{\delta}_t\|^2_{\mathcal{H}^{-1}})\leq\sigma^2$, we have

$$\mathbb{E}\left[\sum_{t=1}^{m_k}t^2\|\boldsymbol{\delta}_t\|^2_{\mathcal{H}^{-1}}\right]\leq\frac{\sigma^2m_k(m_k+1)(2m_k+1)}{6}\leq\frac{\sigma^2}{2}m_k^2(m_k+1)$$

since $m_k\geq1$. Combining these bounds for the terms in ζ^k with the condition $\eta_k\leq1/(2\nu)$ is to get

$$-\mathbb{E}\left[\sum_{k=\kappa}^{\kappa+T}\zeta^k\right]\leq\frac{2}{m_\kappa(m_\kappa+1)\eta_\kappa}\|\boldsymbol{x}-\check{\boldsymbol{x}}^\kappa\|^2_{\mathcal{H}}+\frac{\sigma^2}{2}\sum_{k=\kappa}^{\kappa+T}\eta_km_k.$$

Finally, applying the expectation operator to (4.22) and substituting this bound into the ζ^k term complete the proof. □

By properly setting the algorithm parameters, the following theorem shows the convergence rate of Algorithm 1 in the expectation of both the objective function value gap and the constraint violation.

Theorem 4.2. *Suppose the conditions in Theorem* 4.1 *hold. Let*

$$\eta_k=\min\left\{\frac{c_1}{m_k(m_k+1)},c_2\right\}\quad \textit{and}\quad m_k=\max\{\lceil c_3k^{\varrho}\rceil,m\},\tag{4.24}$$

where $c_1,c_2,c_3>0$, $\varrho\geq 1$ *are constants and* $m>0$ *is a given integer. Then, for every* $\boldsymbol{w}^*\in\Omega^*$, *we have*

$$\left|\mathbb{E}\left[F(\boldsymbol{w}_T)-F(\boldsymbol{w}^*)\right]\right|=E_{\varrho}(T)=\mathbb{E}\left[\|A\boldsymbol{x}_T+B\boldsymbol{y}_T-\boldsymbol{b}\|\right],\tag{4.25}$$

where $E_{\varrho}(T)=\mathcal{O}(1/T)$ *for* $\varrho>1$ *and* $E_{\varrho}(T)=\mathcal{O}(T^{-1}\log T)$ *for* $\varrho=1$.

Proof. The proof is same as that of [3, Theorem 4.2] and thus is omitted here. □

Remark 4.1. (I) In practice, the matrix $\mathcal{M}_k$ in Algorithm 1 could be adaptively adjusted as $\mathcal{M}_k=\rho_k\boldsymbol{I}$, where $\rho_k=\max\{\rho_{\min},\beta\delta_2^k/\delta_1^k\}$ with $\rho_{\min}>0$,

$$\delta_1^k=\left\|\boldsymbol{x}^k-\boldsymbol{x}^{k-1}\right\|^2\quad\text{and}\quad\delta_2^k=\left\|A(\boldsymbol{x}^k-\boldsymbol{x}^{k-1})\right\|^2.$$

Since δ_2^k/δ_1^k is an underestimate of the largest eigenvalue of $A^{\mathsf{T}}A$, to ensure convergence, the safeguard lower bound $\rho_{\min}$ should be increased during the optimization if necessary. One may see [3, Remark 4.2] for more details.

(II) When the set $\mathcal{Y}$ is bounded, we may even use a positive-indefinite proximal matrix

$$L=\tau\chi\boldsymbol{I}-\beta B^{\mathsf{T}}B,\quad\text{where}\quad\beta\|B^{\mathsf{T}}B\|\leq\chi<+\infty\quad\text{and}\quad\tau\in[-1,1],$$

in the update of $\boldsymbol{y}$-subproblem. In this case, denoting $\mathcal{N}_{\mathcal{Y}}=\sup\{\|\boldsymbol{y}_1-\boldsymbol{y}_2\|:\boldsymbol{y}_1,\boldsymbol{y}_2\in\mathcal{Y}\}<\infty$, analogous to Theorem 4.1 we can show

$$\begin{aligned}&\mathbb{E}\left[F(\boldsymbol{w}_T)-F(\boldsymbol{w})+(\boldsymbol{w}_T-\boldsymbol{w})^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\right]\\&\leq\frac{1}{2T}\left\{\sigma^2\sum_{k=\kappa}^{\kappa+T}\eta_km_k+\frac{4}{m_\kappa(m_\kappa+1)\eta_\kappa}\|\boldsymbol{x}-\check{\boldsymbol{x}}^\kappa\|_{\mathcal{H}}^2+\|\boldsymbol{w}-\boldsymbol{w}^\kappa\|_{\widetilde{\mathcal{Q}}_\kappa}^2\right.\\&\quad\left.+\omega_1\|A\boldsymbol{x}^\kappa+B\boldsymbol{y}^\kappa-\boldsymbol{b}\|^2+\omega_2\mathbf{const}\right\},\end{aligned}\tag{4.26}$$

where

$$\mathbf{const}=\begin{cases}\mathcal{N}_{\mathcal{Y}}^2(\beta\|B\|^2-\tau\chi), & \text{if }\ \tau\in[-1,0],\\ \beta(\mathcal{N}_{\mathcal{Y}}\|B\|)^2+\tau\chi\left\|\boldsymbol{y}^{\kappa-1}-\boldsymbol{y}^\kappa\right\|^2, & \text{if }\ \tau\in(0,1].\end{cases}$$

This means that the results of Theorem 4.2 could still hold even when the proximal matrix L is positive-indefinite.

(III) Similar ideas of Algorithm 1 can be further generalized to solve separable convex optimization with one or multi-block structures. For content focus of the paper, we leave these discussions in the Appendix.

5 Numerical experiments

In this section, we apply the proposed algorithm to solve the following graph-guided fused lasso problem in machine learning:

$$\min_{\mathbf{x}} \frac{1}{N}\sum_{j=1}^{N} f_j(\mathbf{x}) + \mu\|A\mathbf{x}\|_1,$$

where $f_j(\mathbf{x}) = \log\left(1+\exp(-b_j a_j^{\mathsf{T}}\mathbf{x})\right)$ denotes the logistic loss function on the feature-label pair $(a_j, b_j) \in \mathbb{R}^l \times \{-1,1\}$, $N(>l)$ is the data size, $\mu > 0$ is a given regularization parameter, and $A = [\mathbf{G};\mathbf{I}]$ is a matrix encoding the feature sparsity pattern. Here, $\mathbf{G}$ is the sparsity pattern of the graph that is obtained by sparse inverse covariance estimation [10]. Introducing an auxiliary variable $\boldsymbol{y}$, the above problem is equivalent to the problem

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & F(\mathbf{x},\mathbf{y}) := \tfrac{1}{N}\sum_{j=1}^{N} f_j(\mathbf{x}) + \mu\|\mathbf{y}\|_1 \\ \text{s.t.} \quad & A\mathbf{x} - \mathbf{y} = \mathbf{0}, \end{aligned} \tag{5.1}$$

which has the format of our model (1.1). In addition, it can be easily verified that the Assumptions 2.1-2.2 hold. Since the coefficient matrix of the $\boldsymbol{y}$ variable in the constraints of (5.1) is $-\boldsymbol{I}$, the $\boldsymbol{y}$-subproblem will have a closed-form solution by simply setting $L = \mathbf{0}$ in Algorithm 1. Otherwise, the linearization techniques discussed in (3.2) on choosing L can be applied to obtain a closed-form solution of the $\boldsymbol{y}$-subproblem. With $L = \mathbf{0}$, the subproblems in Algorithm 1 would have the following closed-form solution:

$$\begin{cases} \check{\boldsymbol{x}}_{t+1} = \left[\gamma_t\mathcal{H} + \mathcal{M}_k\right]^{-1}\left[\gamma_t\mathcal{H}\check{\boldsymbol{x}}_t + \mathcal{M}_k\mathbf{x}^k - \boldsymbol{d}_t - \boldsymbol{h}^k\right], \\ \mathbf{y}^{k+1} = \text{Shrink}\left(\frac{\mu}{\beta}, A\mathbf{x}^{k+1} - \frac{\boldsymbol{\lambda}^{k+\frac{1}{2}}}{\beta}\right). \end{cases} \tag{5.2}$$

Here, Shrink$(\cdot,\cdot)$ denotes the soft shrinkage operator and can be evaluated using the MATLAB built-in function "`wthresh`".

In the numerical experiments, the penalty parameter in SAS-ADMM is taken as $\beta = 0.001$, the matrices $\mathcal{M}_k$ are updated adaptively by the strategy explained in Remark 4.1 (I) with initial values $\rho_0 = 1$, $\rho_{\min} = 10^{-5}$ and $\mathcal{H} = 2\times 10^{-5}\boldsymbol{I}$. The other parameters as well as the vector $\boldsymbol{e}_t$ in SAS-ADMM (i.e. Algorithm 1) are chosen the same way as that used in [3, Section 7.1], that is

$$\boldsymbol{e}_t = \begin{cases} \nabla f(\boldsymbol{x}_{k-1}) - \nabla f_{\tilde{\zeta}_t}(\boldsymbol{x}_{k-1}), & \text{if } m_k > l, \\ \mathbf{0}, & \text{otherwise}, \end{cases} \tag{5.3}$$

where $\boldsymbol{x}_k$ is the ergodic mean of the $\boldsymbol{x}$-iterates. Motivated from Theorem 4.2, we use

$$\text{Obj\_err} = \frac{|F(\boldsymbol{x},\boldsymbol{y}) - F^*|}{\max\{F^*,1\}} \quad \text{and} \quad \text{Equ\_err} = \|A\boldsymbol{x} - \boldsymbol{y}\|$$

to denote the relative objective value error and the constraint violation error. Here, F^* is the approximate optimal objective function value obtained by running Algorithm 1 for more than 10 minutes. To measure the performance of a algorithm, we plot the maximum of the relative objective error and the constraint error, that is

$$\text{Opt\_err} = \max(\text{Obj\_err}, \text{Equ\_err}),$$

against the CPU time used. All experiments are implemented in MATLAB R2018a (64-bit) with the same starting point $(\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\lambda}^0) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$ and performed on a PC with Windows 10 operating system, with an Intel i7-8700K CPU and 16GB RAM.

We compare the numerical performance of the proposed algorithm SAS-ADMM[‡] using stepsizes $(\tau, s) = (0.9, 1.09)$, which is suggested in [1] for GS-ADMM, and AS-ADMM [3] for solving problem (5.1) on the dataset *mnist* (including 11,791 samples and 784 features, that is, $(N, l) = (11791, 784)$) downloaded from LIBSVM website. The regularization parameter μ in (5.1) is set as 10^{-5}. For both SAS-ADMM and AS-ADMM, we plot the error associated with the iterates over the first 1/3 of the total CPU time budget, followed by the error associated with the ergodic iterates over the last 2/3 of the budget. We make 10 and 20 successive runs of each algorithm under the CPU time budgets 120s and 200s, respectively. The average comparison results on Opt_err are shown in Fig. 1, and the comparison of the finally obtained iterative solution $\boldsymbol{x}^{k+1}$ and hist$(\boldsymbol{x}^{k+1})$ are shown in Figs. 2-3. Here, we only compare SAS-ADMM with AS-ADMM since in [3] AS-ADMM was shown competitive or better than other state-of-the-art deterministic and stochastic methods. Note that Opt_err has a big drop at around 1/3 of the CPU time budget, the point where the ergodic iterates are started to use for reporting the objective value. From Fig. 1, we can see that SAS-ADMM initially performs worse than AS-ADMM at the beginning iterations. But after the first 1/3 of the total CPU time budget, the SAS-ADMM eventually seems to perform better than AS-ADMM. Finally, Figs. 2-3 show that both the comparison algorithms indeed get sparse solutions.

6 Conclusion

We proposed a symmetric accelerated stochastic alternating direction method of multipliers, called SAS-ADMM, whose dual variables are symmetrically updated. We gave the specific dual stepsizes region ensuring the global convergence, which is larger than those in the literature. Under proper choice of the algorithm parameters, we proved the convergence of SAS-ADMM in expectation with the worst-case $\mathcal{O}(1/T)$ convergence rate, where T represents the number of iterations. Our preliminary experiments showed that by symmetrically updating the dual variables using a more flexible region, SAS-ADMM could outperform AS-ADMM, which only updates the dual variable once, for solving some structured optimization problems arising in machine learning.

[‡]All codes are available at https://github.com/bjc1987/bjc1987.github.io

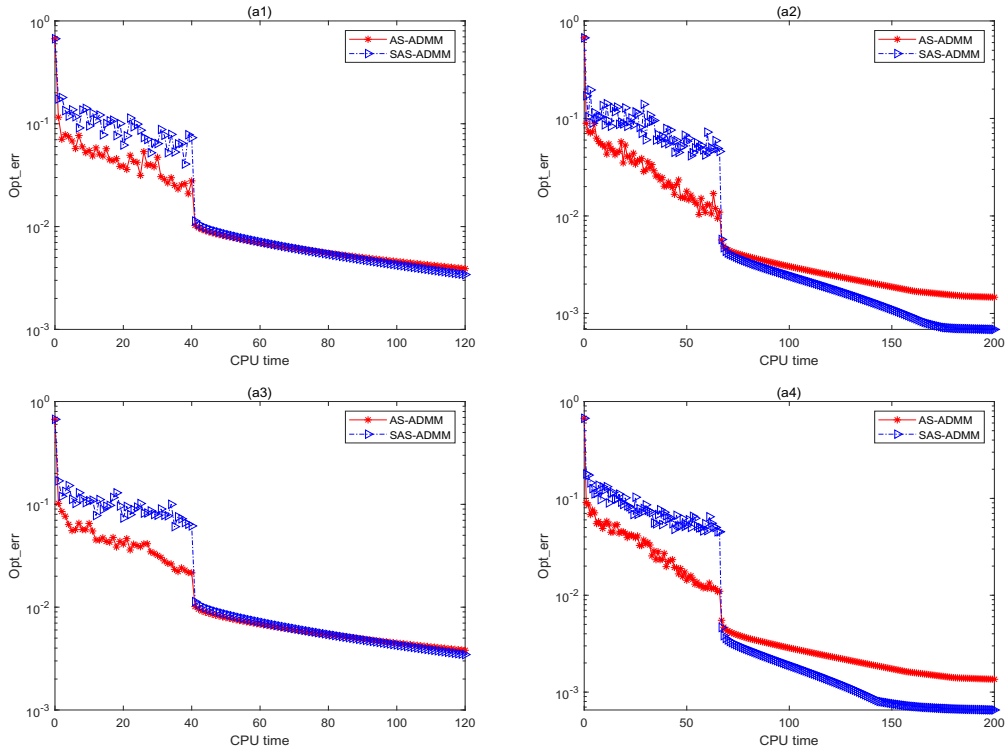

Figure 1: Comparison of Opt_err vs CPU time for Problem (5.1) on the *mnist* dataset: (a1)-(a2) are the results after 10 successive runs; (a3)-(a4) are the results after 20 successive runs.

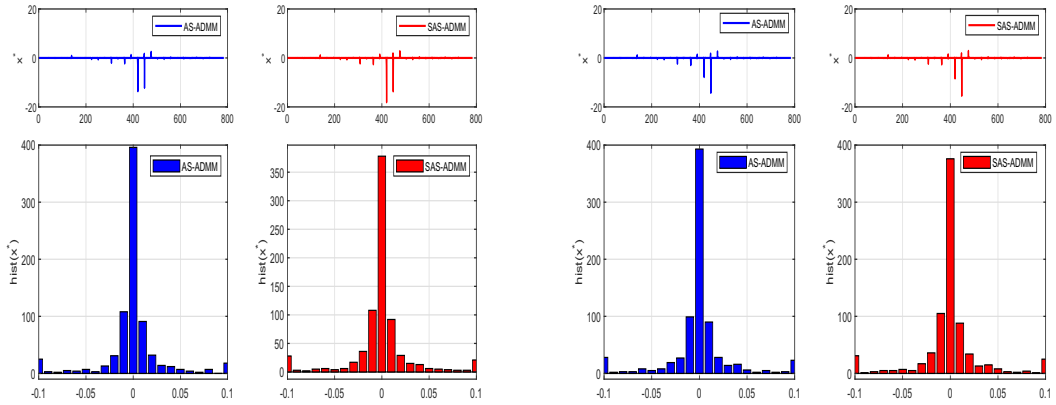

Figure 2: Comparison of the finally obtained iterate $\boldsymbol{x}^{k+1}$ and hist$(\boldsymbol{x}^{k+1})$ after 10 successive runs: the left two subfigures correspond to (a1); the right two subfigures correspond to (a2).

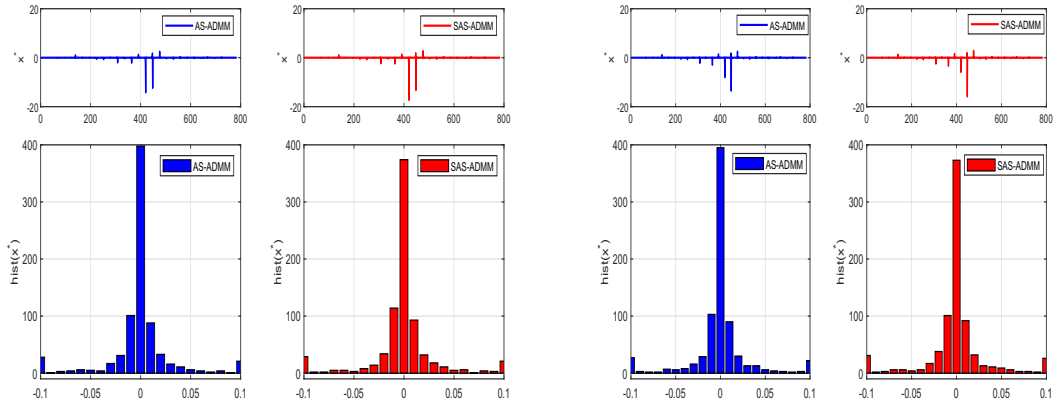

Figure 3: Comparison of the finally obtained iterate x^{k+1} and hist(x^{k+1}) after 20 successive runs: the left two subfigures correspond to (a3); the right two subfigures correspond to (a4).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (12001430, 12131004, 12126603, 72071158), the China Postdoctoral Science Foundation (2020M683545), the Foundation of National Key Laboratory of Science and Technology on Aerodynamic Design and Research (614220119040101) and the USA National Science Foundation (1819161, 2110722).

Appendix: Further discussions

In this section, we discuss 3-block extensions of Algorithm 1 and its variance of a stochastic augmented Lagrangian method.

A.1 A stochastic ALM

We first consider a stochastic augmented Lagrangian method, a variant of SAS-ADMM, to solve

$$\min\{f(\boldsymbol{x}) \,|\, A\boldsymbol{x}=\boldsymbol{b},\, \boldsymbol{x}\in\mathcal{X}\}, \tag{A.1}$$

where $\mathcal{X}\subset\mathbb{R}^{n_1}$ is a closed convex subset, and f is an average of N smooth convex functions as defined in (1.1). Now, the augmented Lagrangian of (A.1) is

$$\mathcal{L}_\beta(\boldsymbol{x},\boldsymbol{\lambda}) := \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) + \frac{\beta}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|^2,$$

where $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) - \boldsymbol{\lambda}^{\mathsf{T}}(A\boldsymbol{x}-\boldsymbol{b})$. Then, based on Algorithm 1, we can propose the following Accelerated Stochastic ALM (AS-ALM), Algorithm 2). Similar to SAS-ADMM, we

Algorithm 2 Accelerated Stochastic ALM (AS-ALM)

Parameters: $\beta>0, s\in(0,2]$ and $\mathcal{H}\succ\mathbf{0}$.
Initialization: $(\boldsymbol{x}^0,\boldsymbol{\lambda}^0)\in\mathcal{X}\times\mathbb{R}^n:=\Omega$ and $\check{\boldsymbol{x}}^0=\boldsymbol{x}^0$.
For $k=0,1,\cdots$
  Choose $m_k>0$, $\eta_k>0$ and $\mathcal{M}_k$ such that $\mathcal{M}_k-\beta A^{\mathsf{T}}A\succeq\mathbf{0}$.
  $\boldsymbol{h}^k:=-A^{\mathsf{T}}\left[\boldsymbol{\lambda}^k-\beta(A\boldsymbol{x}^k-\boldsymbol{b})\right]$.
  $(\boldsymbol{x}^{k+1},\check{\boldsymbol{x}}^{k+1})=$ **xsub** $(\boldsymbol{x}^k,\check{\boldsymbol{x}}^k,\boldsymbol{h}^k)$ with **xsub** given in Algorithm 1.
  $\boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^k-s\beta\left(A\boldsymbol{x}^{k+1}-\boldsymbol{b}\right)$.
end

can easily establish the following lemmas on AS-ALM. However, in this case, the convergence region for the dual stepsize can be enlarged from $(0,(\sqrt{5}+1)/2]$ of AS-ADMM [3] to $(0,2]$.

Lemma A.1. *Let $\{\boldsymbol{x}^k\}$ be generated by Algorithm 2 and $\eta_k\in(0,1/\nu)$. Then, the inequality (4.1) holds with*

$$\widetilde{\boldsymbol{\lambda}}^k=\boldsymbol{\lambda}^k-\beta\left(A\boldsymbol{x}^{k+1}-\boldsymbol{b}\right). \tag{A.2}$$

For the iterates generated by Algorithm 2, in this subsection let $\boldsymbol{w}^k=(\boldsymbol{x}^k;\boldsymbol{\lambda}^k)$ and $\widetilde{\boldsymbol{w}}^k=(\boldsymbol{x}^{k+1};\widetilde{\boldsymbol{\lambda}}^k)$, where $\widetilde{\boldsymbol{\lambda}}^k$ is defined in (A.2). Then, we have the following lemma.

Lemma A.2. *Let $\{\boldsymbol{w}^k\}$ be generated by Algorithm 2 and $\eta_k\in(0,1/\nu)$. Then, we have $\widetilde{\boldsymbol{w}}^k\in\Omega$ and*

$$f(\boldsymbol{x})-f(\boldsymbol{x}^{k+1})+\left\langle\boldsymbol{w}-\widetilde{\boldsymbol{w}}^k,\mathcal{J}(\boldsymbol{w})\right\rangle\geq(\boldsymbol{w}-\widetilde{\boldsymbol{w}}^k)^{\mathsf{T}}Q_k(\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k)+\zeta^k$$

for all $\boldsymbol{w}\in\Omega$, where ζ^k is given by (4.3),

$$\mathcal{J}(\boldsymbol{w})=\begin{pmatrix}-A^{\mathsf{T}}\boldsymbol{\lambda}\\ A\boldsymbol{x}-\boldsymbol{b}\end{pmatrix}\quad\text{and}\quad Q_k=\begin{bmatrix}\mathcal{D}_k & \\ & \frac{1}{\beta}\boldsymbol{I}\end{bmatrix}.$$

Proof. Combining the inequality (4.1) and the relation $A\boldsymbol{x}^{k+1}-\boldsymbol{b}=\frac{1}{\beta}(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k)$ gives the results. □

Lemma A.3. *Let $\{\boldsymbol{w}^k\}$ be generated by Algorithm 2 and $\eta_k\in(0,1/\nu)$. Then, for any $\boldsymbol{w}\in\Omega$, we have*

$$\begin{aligned}&f(\boldsymbol{x})-f(\boldsymbol{x}^{k+1})+(\boldsymbol{w}-\widetilde{\boldsymbol{w}}^k)^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\\ \geq&\frac{1}{2}\left\{\left\|\boldsymbol{w}-\boldsymbol{w}^{k+1}\right\|^2_{\widetilde{Q}_k}-\left\|\boldsymbol{w}-\boldsymbol{w}^k\right\|^2_{\widetilde{Q}_k}+\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|^2_{\widetilde{G}_k}\right\}+\zeta^k,\end{aligned}$$

where ζ^k is given by (4.3),

$$\widetilde{Q}_k=\begin{bmatrix}\mathcal{D}_k & \\ & \frac{1}{s\beta}\boldsymbol{I}\end{bmatrix} \quad \textit{and} \quad \widetilde{G}_k=\begin{bmatrix}\mathcal{D}_k & \\ & \frac{2-s}{\beta}\boldsymbol{I}\end{bmatrix}.$$

Proof. The proof is similar to that of Corollary 4.1 and is omitted. □

Finally, under the conditions of Theorem 4.2, by Lemma A.3 and a similar proof of Theorem 4.2, we can deduce that for any $\boldsymbol{w}_T:=\frac{1}{1+T}\sum_{k=\kappa}^{\kappa+T}\widetilde{\boldsymbol{w}}^k$ and $\kappa\geq 0$, it has

$$|\mathbb{E}[f(\boldsymbol{x}_T)-f(\boldsymbol{x}^*)]|=E_\varrho(T)=\mathbb{E}\big[\|A\boldsymbol{x}_T-\boldsymbol{b}\|\big],$$

where $E_\varrho(T)=\mathcal{O}(1/T)$ for $\varrho>1$ and $E_\varrho(T)=\mathcal{O}(T^{-1}\log T)$ for $\varrho=1$.

A.2 Three-block extensions

Consider a 3-block extension of problem (1.1)

$$\begin{aligned}\min\quad & F(\boldsymbol{w}):=f(\boldsymbol{x})+g(\boldsymbol{y})+l(\boldsymbol{z})\\ \text{s.t.}\quad & \mathcal{K}\boldsymbol{w}:=A\boldsymbol{x}+B\boldsymbol{y}+C\boldsymbol{z}=\boldsymbol{b},\\ & \boldsymbol{x}\in\mathcal{X},\quad \boldsymbol{y}\in\mathcal{Y},\quad \boldsymbol{z}\in\mathcal{Z},\end{aligned} \tag{A.3}$$

where l is a closed convex function, $C\in\mathbb{R}^{n\times n_3}$ is a given matrix, $\mathcal{Z}\subset\mathbb{R}^{n_3}$ is a simple closed convex subset, and the other functions and variables remain the same definitions as those in problem (1.1). Here, the additional function l can be possibly used to promote some data structure different from the structure promoted by g. For convenience, in this subsection, let us define $\mathcal{K}\boldsymbol{w}:=A\boldsymbol{x}+B\boldsymbol{y}+C\boldsymbol{z}$, denote $\boldsymbol{w}=(\boldsymbol{z};\boldsymbol{x};\boldsymbol{y};\boldsymbol{\lambda})$,

$$\boldsymbol{w}^k=\begin{pmatrix}\boldsymbol{x}^k\\ \boldsymbol{y}^k\\ \boldsymbol{z}^k\\ \boldsymbol{\lambda}^k\end{pmatrix},\quad \widetilde{\boldsymbol{w}}^k:=\begin{pmatrix}\widetilde{\boldsymbol{x}}^k\\ \widetilde{\boldsymbol{y}}^k\\ \widetilde{\boldsymbol{z}}^k\\ \widetilde{\boldsymbol{\lambda}}^k\end{pmatrix}=\begin{pmatrix}\boldsymbol{x}^{k+1}\\ \boldsymbol{y}^{k+1}\\ \boldsymbol{z}^{k+1}\\ \widetilde{\boldsymbol{\lambda}}^k\end{pmatrix}\quad\text{and}\quad \mathcal{J}(\boldsymbol{w})=\begin{pmatrix}-A^{\mathsf{T}}\boldsymbol{\lambda}\\ -B^{\mathsf{T}}\boldsymbol{\lambda}\\ -C^{\mathsf{T}}\boldsymbol{\lambda}\\ \mathcal{K}\boldsymbol{w}-\boldsymbol{b}\end{pmatrix}, \tag{A.4}$$

where $\widetilde{\boldsymbol{\lambda}}^k$ will be specified differently in the following two discussion cases.

A.2.1 Extension in Gauss-Seidel update

For this case, we need an assumption that $C^{\mathsf{T}}A=\boldsymbol{0}$. Then SAS-ADMM can be directly extended to Algorithm 3 for solving the 3-block problem (A.3), where the variable updating order is $\boldsymbol{z}^{k+1}\to\boldsymbol{x}^{k+1}\to\boldsymbol{y}^{k+1}\to\boldsymbol{\lambda}^{k+1}$ in a Gauss-Seidel scheme. Now, let

$$\widetilde{\boldsymbol{\lambda}}^k=\boldsymbol{\lambda}^k-\beta\Big(C\boldsymbol{z}^{k+1}+A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}\Big), \tag{A.5}$$

$$\boldsymbol{v}^{k+1}=(\boldsymbol{x}^{k+1};\boldsymbol{y}^{k+1};\boldsymbol{\lambda}^{k+1})\quad\text{and}\quad \widetilde{\boldsymbol{v}}^k=(\widetilde{\boldsymbol{x}}^k;\widetilde{\boldsymbol{y}}^k;\widetilde{\boldsymbol{\lambda}}^k). \tag{A.6}$$

Then, we have the following main lemma for the convergence of Algorithm 3.

Algorithm 3 Extension of SAS-ADMM in Gauss-Seidel update

Parameters: $\beta>0,\ \mathcal{H}\succ\mathbf{0},\ L\succeq\mathbf{0}$ and $(\tau,s)\in\Delta$.
Initialization: $(\boldsymbol{x}^0,\boldsymbol{y}^0,\boldsymbol{z}^0,\boldsymbol{\lambda}^0)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\times\mathbb{R}^n:=\Omega,\ \breve{\boldsymbol{x}}^0=\boldsymbol{x}^0$.
For $k=0,1,\cdots$
Choose $m_k>0,\ \eta_k>0$ and $\mathcal{M}_k$ such that $\mathcal{M}_k-\beta A^{\mathsf{T}}A\succeq\mathbf{0}$.

$$\boldsymbol{z}^{k+1}\in\operatorname*{argmin}_{\boldsymbol{z}\in\mathcal{Z}} l(\boldsymbol{z})+\tfrac{\beta}{2}\left\|C\boldsymbol{z}+A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}-\frac{\boldsymbol{\lambda}^k}{\beta}\right\|^2.$$

$$\boldsymbol{h}^k:=-A^{\mathsf{T}}\left[\boldsymbol{\lambda}^k-\beta(C\boldsymbol{z}^{k+1}+A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b})\right].$$

$(\boldsymbol{x}^{k+1},\breve{\boldsymbol{x}}^{k+1})=$ **xsub** $(\boldsymbol{x}^k,\breve{\boldsymbol{x}}^k,\boldsymbol{h}^k)$ with **xsub** given in ALG.1.

$$\boldsymbol{\lambda}^{k+\frac{1}{2}}=\boldsymbol{\lambda}^k-\tau\beta\left(C\boldsymbol{z}^{k+1}+A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}\right).$$

$$\boldsymbol{y}^{k+1}\in\operatorname*{argmin}_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{y})+\tfrac{\beta}{2}\left\|C\boldsymbol{z}^{k+1}+A\boldsymbol{x}^{k+1}+B\boldsymbol{y}-\boldsymbol{b}-\frac{\boldsymbol{\lambda}^{k+\frac{1}{2}}}{\beta}\right\|^2+\tfrac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^k\right\|_L^2.$$

$$\boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^{k+\frac{1}{2}}-s\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}+C\boldsymbol{z}^{k+1}-\boldsymbol{b}\right).$$

end

Lemma A.4. *Assume $C^{\mathsf{T}}A=\mathbf{0}$ and $\eta_k\in(0,1/\nu)$. Then, the iterates generated by Algorithm 3 satisfy $\widetilde{\boldsymbol{w}}^k\in\Omega$ and*

$$\begin{aligned}&F(\boldsymbol{w})-F(\widetilde{\boldsymbol{w}}^k)+\left(\boldsymbol{w}-\widetilde{\boldsymbol{w}}^k\right)^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\\ \geq&\frac{1}{2}\left\{\left\|\boldsymbol{v}-\boldsymbol{v}^{k+1}\right\|_{\widetilde{Q}_k}^2-\left\|\boldsymbol{v}-\boldsymbol{v}^k\right\|_{\widetilde{Q}_k}^2+\left\|\boldsymbol{v}^k-\widetilde{\boldsymbol{v}}^k\right\|_{\widetilde{G}_k}^2\right\}+\zeta^k\end{aligned}$$

for any $\boldsymbol{w}\in\Omega$, where $\widetilde{Q}_k,\widetilde{G}_k$ and ζ^k are given in Corollary 4.1 and (4.3), respectively. Moreover, we have

$$\begin{aligned}\left\|\boldsymbol{v}^k-\widetilde{\boldsymbol{v}}^k\right\|_{\widetilde{G}_k}^2\geq&\left\|\boldsymbol{x}^k-\boldsymbol{x}^{k+1}\right\|_{\mathcal{D}_k}^2+\omega_0\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^2\\&+\omega_1\left(\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right\|^2-\left\|A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right\|^2\right)\\&+\omega_2\left(\left\|\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\right\|_L^2-\left\|\boldsymbol{y}^{k-1}-\boldsymbol{y}^k\right\|_L^2\right),\end{aligned}$$

where $\omega_0,\omega_1,\omega_2\geq0$ are given in (4.16).

Proof. By the updates of $\boldsymbol{h}^k$ and $\widetilde{\boldsymbol{\lambda}}^k$ in Algorithm 3, it is easy to derive (4.1) as before. Then, according to the first-order optimality condition of z-subproblem and the assumption that $C^{\mathsf{T}}A=\mathbf{0}$, we have

$$\boldsymbol{z}^{k+1}\in\mathcal{Z},\quad l(\boldsymbol{z})-l(\boldsymbol{z}^{k+1})+\left\langle\boldsymbol{z}-\boldsymbol{z}^{k+1},\boldsymbol{p}_{\boldsymbol{z}}^k\right\rangle\geq0,\quad\forall\boldsymbol{z}\in\mathcal{Z},\tag{A.7}$$

where

$$\begin{aligned}\boldsymbol{p}_z^k &= -C^{\mathsf{T}}\boldsymbol{\lambda}^k+\beta C^{\mathsf{T}}\left(C\boldsymbol{z}^{k+1}+A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b}\right)\\ &= -C^{\mathsf{T}}\widetilde{\boldsymbol{\lambda}}^k-\beta C^{\mathsf{T}}A(\boldsymbol{x}^{k+1}-\boldsymbol{x}^k)=-C^{\mathsf{T}}\widetilde{\boldsymbol{\lambda}}^k.\end{aligned}$$

Similarly, we have by the $\boldsymbol{y}$-update that

$$\boldsymbol{y}^{k+1}\in\mathcal{Y},\quad g(\boldsymbol{y})-g(\boldsymbol{y}^{k+1})+\left\langle \boldsymbol{y}-\boldsymbol{y}^{k+1},\boldsymbol{p}_y^k\right\rangle\geq 0,\quad \forall \boldsymbol{y}\in\mathcal{Y}, \tag{A.8}$$

where

$$\begin{aligned}\boldsymbol{p}_y^k &= -B^{\mathsf{T}}\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B^{\mathsf{T}}\left(\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b}\right)+L(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k)\\ &= -B^{\mathsf{T}}\boldsymbol{\lambda}^{k+\frac{1}{2}}+B^{\mathsf{T}}\left(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k\right)+\left[L+\beta B^{\mathsf{T}}B\right]\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right)\\ &= -B^{\mathsf{T}}\widetilde{\boldsymbol{\lambda}}^k+\tau B^{\mathsf{T}}(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k)+\left[L+\beta B^{\mathsf{T}}B\right]\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right).\end{aligned}$$

Besides, it follows from the updates of $\widetilde{\boldsymbol{\lambda}}^k$ that

$$\left\langle \boldsymbol{\lambda}-\widetilde{\boldsymbol{\lambda}}^k,\mathcal{K}\widetilde{\boldsymbol{w}}^k-\boldsymbol{b}+\frac{1}{\beta}\left(\widetilde{\boldsymbol{\lambda}}^k-\boldsymbol{\lambda}^k\right)-B\left(\widetilde{\boldsymbol{y}}^k-\boldsymbol{y}^k\right)\right\rangle=0,\quad \forall\boldsymbol{\lambda}\in\mathbb{R}^n. \tag{A.9}$$

Combining the above inequalities (A.7), (A.8), (A.9) with (4.1), we can get

$$F(\boldsymbol{w})-F(\widetilde{\boldsymbol{w}}^k)+\left(\boldsymbol{w}-\widetilde{\boldsymbol{w}}^k\right)^{\mathsf{T}}\mathcal{J}(\widetilde{\boldsymbol{w}}^k)\geq(\boldsymbol{v}-\widetilde{\boldsymbol{v}}^k)^{\mathsf{T}}Q_k(\boldsymbol{v}^k-\widetilde{\boldsymbol{v}}^k)+\zeta^k, \tag{A.10}$$

where ζ^k, Q_k are given by (4.3) and (4.5), respectively. Then, the rest proof will be similar to that of Corollary 4.1 and Lemma 4.4. □

Based on the above Lemma A.4, the ergodic convergence of Algorithm 3 with a sublinear convergence rate can be similarly established under the conditions of Theorem 4.2. Here, we omit the detailed proof. Note that, if the $\boldsymbol{z}$-subproblem is not easily solvable, one could also add a positive semidefinite proximal term to linearize it. However, the requirement $C^{\mathsf{T}}A=\boldsymbol{0}$ is quite strict in applications. In the next subsection we will propose a partially Jacobi update for the primal variables, for which $C^{\mathsf{T}}A=\boldsymbol{0}$ is not required.

A.2.2 Extension in partially Jacobi update

Now, let us consider Algorithm 4, where the block variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are updated in a Jacobi fashion.

To establish the global convergence of Algorithm 4, we first have the following observations. Denoting

$$\widetilde{\boldsymbol{\lambda}}^k=\boldsymbol{\lambda}^k-\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k+C\boldsymbol{z}^k-\boldsymbol{b}\right) \tag{A.11}$$

Algorithm 4 Extension of SAS-ADMM in partially Jacobi update

Parameters: $\beta>0,(\tau,s)\in\Delta$, L_1 and L_2 satisfy (A.22).
Initialization: $(\boldsymbol{x}^0,\boldsymbol{y}^0,\boldsymbol{z}^0,\boldsymbol{\lambda}^0)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\times\mathbb{R}^n:=\Omega$, $\check{\boldsymbol{x}}^0=\boldsymbol{x}^0$.
For $k=0,1,\cdots$
Choose $m_k>0$, $\eta_k>0$ and $\mathcal{M}_k$ such that $\mathcal{M}_k-\beta A^{\mathsf{T}}A\succeq\mathbf{0}$.
$\boldsymbol{h}^k:=-A^{\mathsf{T}}\left[\boldsymbol{\lambda}^k-\beta(A\boldsymbol{x}^k+B\boldsymbol{y}^k+C\boldsymbol{z}^k-\boldsymbol{b})\right]$.
$(\boldsymbol{x}^{k+1},\check{\boldsymbol{x}}^{k+1})=$ **xsub** $(\boldsymbol{x}^k,\check{\boldsymbol{x}}^k)$ with **xsub** given in ALG.1.
$\boldsymbol{\lambda}^{k+\frac{1}{2}}=\boldsymbol{\lambda}^k-\tau\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k+C\boldsymbol{z}^k-\boldsymbol{b}\right)$.
$\boldsymbol{y}^{k+1}\in\underset{\boldsymbol{y}\in\mathcal{Y}}{\arg\min}\, g(\boldsymbol{y})+\frac{\beta}{2}\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}+C\boldsymbol{z}^k-\boldsymbol{b}-\frac{\boldsymbol{\lambda}^{k+\frac{1}{2}}}{\beta}\right\|^2+\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^k\right\|_{L_1}^2$.
$\boldsymbol{z}^{k+1}\in\underset{\boldsymbol{z}\in\mathcal{Z}}{\arg\min}\, l(\boldsymbol{z})+\frac{\beta}{2}\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k+C\boldsymbol{z}-\boldsymbol{b}-\frac{\boldsymbol{\lambda}^{k+\frac{1}{2}}}{\beta}\right\|^2+\frac{1}{2}\left\|\boldsymbol{z}-\boldsymbol{z}^k\right\|_{L_2}^2$.
$\boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^{k+\frac{1}{2}}-s\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}+C\boldsymbol{z}^{k+1}-\boldsymbol{b}\right)$.
end

and using the first-order optimality condition of the $\boldsymbol{y}$-subproblem, we have

$$\boldsymbol{y}^{k+1}\in\mathcal{Y},\quad g(\boldsymbol{y})-g(\boldsymbol{y}^{k+1})+\left\langle\boldsymbol{y}-\boldsymbol{y}^{k+1},\boldsymbol{p}_{\boldsymbol{y}}^k\right\rangle\geq 0,\quad\forall\boldsymbol{y}\in\mathcal{Y},\tag{A.12}$$

where

$$\begin{aligned}\boldsymbol{p}_{\boldsymbol{y}}^k&=-B^{\mathsf{T}}\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B^{\mathsf{T}}\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}+C\boldsymbol{z}^k-\boldsymbol{b}\right)+L_1\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right)\\&=-B^{\mathsf{T}}\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B^{\mathsf{T}}\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k+C\boldsymbol{z}^k-\boldsymbol{b}\right)+(L_1+\beta B^{\mathsf{T}}B)\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right)\\&=-B^{\mathsf{T}}\widetilde{\boldsymbol{\lambda}}^k+\tau B^{\mathsf{T}}(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k)+(L_1+\beta B^{\mathsf{T}}B)\left(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\right),\end{aligned}$$

and we use the relationship

$$\boldsymbol{\lambda}^{k+\frac{1}{2}}=\boldsymbol{\lambda}^k-\tau(\boldsymbol{\lambda}^k-\widetilde{\boldsymbol{\lambda}}^k).\tag{A.13}$$

Combining (A.12) and the definition of $\boldsymbol{p}_{\boldsymbol{y}}^k$, we have

$$\begin{aligned}g(\boldsymbol{y})-g(\boldsymbol{y}^{k+1})+\Big\langle\boldsymbol{y}-\boldsymbol{y}^{k+1},&-B^{\mathsf{T}}\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B^{\mathsf{T}}(\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b})\\&-\beta B^{\mathsf{T}}C(\boldsymbol{z}^{k+1}-\boldsymbol{z}^k)+L_1(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k)\Big\rangle\geq 0.\end{aligned}\tag{A.14}$$

Similarly, by the first-order optimality condition of the $\boldsymbol{z}$-subproblem, we have

$$\begin{aligned}l(\boldsymbol{z})-l(\boldsymbol{z}^{k+1})+\Big\langle\boldsymbol{z}-\boldsymbol{z}^{k+1},&-B^{\mathsf{T}}\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta C^{\mathsf{T}}(\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b})\\&-\beta C^{\mathsf{T}}B(\boldsymbol{y}^{k+1}-\boldsymbol{y}^k)+L_2(\boldsymbol{z}^{k+1}-\boldsymbol{z}^k)\Big\rangle\geq 0.\end{aligned}\tag{A.15}$$

Adding the above two inequalities (A.14) and (A.15), we can see $(\boldsymbol{y}^{k+1},\boldsymbol{z}^{k+1})$ satisfies the first-order optimality condition, hence is a solution, of the following problem

$$(\boldsymbol{y}^{k+1},\boldsymbol{z}^{k+1})\in\arg\min_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{z}\in\mathcal{Z}}\left\{\begin{array}{c} g(\boldsymbol{y})+l(\boldsymbol{z})+\frac{1}{2}\left\|(\boldsymbol{y}-\boldsymbol{y}^k,\boldsymbol{z}-\boldsymbol{z}^k)\right\|_{\overline{L}}^2 \\ +\frac{\beta}{2}\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}+C\boldsymbol{z}-\boldsymbol{b}-\frac{\boldsymbol{\lambda}^{k+\frac{1}{2}}}{\beta}\right\|^2 \end{array}\right\}, \tag{A.16}$$

where

$$\overline{L}=\left[\begin{array}{cc} L_1 & -\beta B^{\mathsf{T}}C \\ -\beta C^{\mathsf{T}}B & L_2 \end{array}\right]. \tag{A.17}$$

Hence, by considering $(\boldsymbol{y},\boldsymbol{z})$ as one block variable, Algorithm 4 is essentially a particular version of Algorithm 1 for solving a 2-block problem with L and B being replaced by $\overline{L}$ and (B,C), respectively.

From the above observations, we can directly establish the following properties of Algorithm 4.

Lemma A.5. *The iterates generated by Algorithm* 4 *satisfy*

$$F(\boldsymbol{w})-F(\widetilde{\boldsymbol{w}}^k)+\left\langle \boldsymbol{w}-\widetilde{\boldsymbol{w}}^k,\mathcal{J}(\boldsymbol{w})\right\rangle\geq(\boldsymbol{w}-\widetilde{\boldsymbol{w}}^k)^{\mathsf{T}}Q_k(\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k)+\zeta^k$$

for any $\boldsymbol{w}\in\Omega$, *where* ζ^k *is given by* (4.3),

$$Q_k=\left[\begin{array}{ccc} \mathcal{D}_k & & \\ & L_1+\beta B^{\mathsf{T}}B & -\tau B^{\mathsf{T}} \\ & & L_2+\beta C^{\mathsf{T}}C \quad -\tau C^{\mathsf{T}} \\ & -B \quad -C & \frac{1}{\beta}\boldsymbol{I} \end{array}\right]. \tag{A.18}$$

Proof. Notice that

$$\overline{L}+\beta(B,C)^{\mathsf{T}}(B,C)=\left[\begin{array}{cc} L_1+\beta B^{\mathsf{T}}B & \\ & L_2+\beta C^{\mathsf{T}}C \end{array}\right].$$

So, replacing L and B in Lemma 4.2 by $\overline{L}$ and (B,C), respectively, this lemma directly follows from Lemma 4.2. □

Similarly, identifying L and B in (4.11) with $\overline{L}$ and (B,C), respectively, it follows from Corollary 4.1 that

$$\begin{aligned} &F(\boldsymbol{w})-F(\widetilde{\boldsymbol{w}}^k)+\left\langle \boldsymbol{w}-\widetilde{\boldsymbol{w}}^k,\mathcal{J}(\boldsymbol{w})\right\rangle \\ &\geq\frac{1}{2}\left\{\left\|\boldsymbol{w}-\boldsymbol{w}^{k+1}\right\|_{\widetilde{Q}_k}^2-\left\|\boldsymbol{w}-\boldsymbol{w}^k\right\|_{\widetilde{Q}_k}^2+\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{G}_k}^2\right\}+\zeta^k, \end{aligned} \tag{A.19}$$

where ζ^k is given by (4.3) and

$$\widetilde{Q}_k=\begin{bmatrix} \mathcal{D}_k & & & \\ & L_1+(1-\frac{\tau s}{\tau+s})\beta B^\mathsf{T}B & -\frac{\tau s}{\tau+s}\beta B^\mathsf{T}C & -\frac{\tau}{\tau+s}B^\mathsf{T} \\ & -\frac{\tau s}{\tau+s}\beta C^\mathsf{T}B & L_2+(1-\frac{\tau s}{\tau+s})\beta C^\mathsf{T}C & -\frac{\tau}{\tau+s}C^\mathsf{T} \\ & -\frac{\tau}{\tau+s}B & -\frac{\tau}{\tau+s}C & \frac{1}{(\tau+s)\beta}\boldsymbol{I} \end{bmatrix}, \tag{A.20}$$

$$\widetilde{G}_k=\begin{bmatrix} \mathcal{D}_k & & & \\ & L_1+(1-s)\beta B^\mathsf{T}B & -s\beta B^\mathsf{T}C & (s-1)B^\mathsf{T} \\ & -s\beta C^\mathsf{T}B & L_2+(1-s)\beta C^\mathsf{T}C & (s-1)C^\mathsf{T} \\ & (s-1)B & (s-1)C & \frac{2-\tau-s}{\beta}\boldsymbol{I} \end{bmatrix}. \tag{A.21}$$

Then, we have the following estimate on a lower bound of $\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\|_{\widetilde{G}_k}$.

Lemma A.6. *Suppose there exist $\gamma_1>0$ and $\gamma_2>0$ with $\gamma_1\gamma_2\geq 1$ such that*

$$L_1\succeq\gamma_1\beta B^\mathsf{T}B \quad \textit{and} \quad L_2\succeq\gamma_2\beta C^\mathsf{T}C. \tag{A.22}$$

Then, for any $(\tau,s)\in\Delta$ defined in (3.1), we have $\widetilde{Q}_k$ defined in (A.20) is positive semidefinite and

$$\begin{aligned}\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|^2_{\widetilde{G}_k}\geq&\left\|\boldsymbol{x}^k-\boldsymbol{x}^{k+1}\right\|^2_{\mathcal{D}_k}+\omega_0\left\|\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b}\right\|^2\\&+\omega_1\left(\left\|\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b}\right\|^2-\left\|\mathcal{K}\boldsymbol{w}^{k}-\boldsymbol{b}\right\|^2\right)\\&+\omega_2\left(\left\|\begin{pmatrix}\boldsymbol{y}^k-\boldsymbol{y}^{k+1}\\ \boldsymbol{z}^k-\boldsymbol{z}^{k+1}\end{pmatrix}\right\|^2_{\overline{L}}-\left\|\begin{pmatrix}\boldsymbol{y}^k-\boldsymbol{y}^{k-1}\\ \boldsymbol{z}^k-\boldsymbol{z}^{k-1}\end{pmatrix}\right\|^2_{\overline{L}}\right),\end{aligned} \tag{A.23}$$

where $\omega_0,\omega_1,\omega_2\geq 0$ is defined in (4.16) and $\overline{L}$ is defined in (A.17).

Proof. First, since $L_1\succeq\gamma_1\beta B^\mathsf{T}B$ and $L_2\succeq\gamma_2\beta C^\mathsf{T}C$, it follows from $\gamma_1>0,\gamma_2>0$ and $\gamma_1\gamma_2\geq 1$ that

$$\overline{L}=\begin{bmatrix} L_1 & -\beta B^\mathsf{T}C \\ -\beta C^\mathsf{T}B & L_2 \end{bmatrix}\succeq\beta\begin{bmatrix} \gamma_1 B^\mathsf{T}B & -B^\mathsf{T}C \\ -C^\mathsf{T}B & \gamma_2 C^\mathsf{T}C \end{bmatrix}\succeq\boldsymbol{0}. \tag{A.24}$$

By Lemma 4.3, we have $\widetilde{Q}_k$ defined in (A.20) is positive semidefinite if

$$\overline{L}\succeq(\tau-1)\beta(B,C)^\mathsf{T}(B,C). \tag{A.25}$$

Since $\tau\leq 1$ for any $(\tau,s)\in\Delta$, we have $\boldsymbol{0}\succeq(\tau-1)\beta(B,C)^\mathsf{T}(B,C)$. Therefore, we have from (A.24) that (A.25) holds automatically and therefore, $\widetilde{Q}_k$ defined in (A.20) is positive semidefinite. Furthermore, it follows from Lemma 4.4 that (A.23) holds as long as $\overline{L}\succeq\boldsymbol{0}$ which is verified by (A.24). □

Now, defining $\boldsymbol{w}_T := \frac{1}{T}\sum_{k=\kappa}^{\kappa+T}\widetilde{\boldsymbol{w}}^k$ for some integers $T>0$ and $\kappa>0$, under the same conditions in Theorem 4.1, by Lemma A.6 and similar to the proof of Theorem 4.1, we can obtain

$$\begin{aligned}
&\mathbb{E}\Big[F(\boldsymbol{w}_T)-F(\boldsymbol{w})+(\boldsymbol{w}_T-\boldsymbol{w})^{\mathsf{T}}\mathcal{J}(\boldsymbol{w})\Big]\\
\le&\frac{1}{2T}\Big\{\sigma^2\sum_{k=\kappa}^{\kappa+T}\eta_k m_k+\frac{4}{m_\kappa(m_\kappa+1)\eta_\kappa}\|\boldsymbol{x}-\breve{\boldsymbol{x}}^\kappa\|_{\mathcal{H}}^2+\|\boldsymbol{w}-\boldsymbol{w}^\kappa\|_{\widetilde{\mathcal{Q}}_\kappa}^2\\
&+\omega_1\|\mathcal{K}\boldsymbol{w}^\kappa-\boldsymbol{b}\|^2+\omega_2\left\|\begin{pmatrix}\boldsymbol{y}^\kappa-\boldsymbol{y}^{\kappa-1}\\ \boldsymbol{z}^\kappa-\boldsymbol{z}^{\kappa-1}\end{pmatrix}\right\|_{\overline{L}}^2\Big\},
\end{aligned}$$

where $\omega_1\ge 0$ and $\omega_2\ge 0$ given in (4.16). So, by the choice of the parameters (η_k,m_k) chosen in Theorem 4.2, we can obtain

$$\big|\mathbb{E}\big[F(\boldsymbol{w}_T)-F(\boldsymbol{w}^*)\big]\big|=E_\varrho(T)=\mathbb{E}\big[\|A\boldsymbol{x}_T+B\boldsymbol{y}_T+C\boldsymbol{z}_T-\boldsymbol{b}\|\big],$$

where $E_\varrho(T)=\mathcal{O}(1/T)$ for the parameter $\varrho>1$ and $E_\varrho(T)=\mathcal{O}(T^{-1}\log T)$ for $\varrho=1$.

Remark A.1. Observing from the above analysis, Algorithm 4 could be in fact generalized to Algorithm 5 for solving the multi-block separable convex optimization:

$$\begin{aligned}
\min\quad & F(\boldsymbol{w}):=f(\boldsymbol{x})+\sum_{i=1}^{q}g_i(\boldsymbol{y}_i)\\
\text{s.t.}\quad & \mathcal{K}\boldsymbol{w}:=A\boldsymbol{x}+\sum_{i=1}^{q}B_i\boldsymbol{y}_i=\boldsymbol{b},\\
& \boldsymbol{x}\in\mathcal{X},\quad \boldsymbol{y}_i\in\mathcal{Y}_i,\quad i=1,2,\cdots,q,
\end{aligned}\tag{A.26}$$

where f has the same definition as in (1.1), $g_i:\mathcal{Y}_i\to\mathbb{R}\cup\{+\infty\}$ is a convex but possibly nonsmooth function, $B_i\in\mathbb{R}^{n\times n_i}$ and $\mathcal{Y}_i\subset\mathbb{R}^{n_i}$ is a closed convex subset.

The convergence of Algorithm 5 can be analogously established with proper modifications on the convergence proof of Algorithm 4. Here, we only give a very brief explanation. Denote $g(\boldsymbol{y})=\sum_{i=1}^{q}g_i(\boldsymbol{y}_i)$, $B=(B_1,\cdots,B_q)$, $\boldsymbol{y}=(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_q)$, $\boldsymbol{y}^k=(\boldsymbol{y}_1^k;\cdots;\boldsymbol{y}_q^k)$ and $\mathcal{Y}=\mathcal{Y}_1\times\cdots\times\mathcal{Y}_q$. Then, by the first-order optimality condition of $\boldsymbol{y}_i$-subproblem, we have $\boldsymbol{y}_i^{k+1}\in\mathcal{Y}_i$ and

$$\begin{aligned}
&g_i(\boldsymbol{y}_i)-g_i(\boldsymbol{y}_i^{k+1})+\Big\langle\boldsymbol{y}_i-\boldsymbol{y}_i^{k+1},-B_i^{\mathsf{T}}\boldsymbol{\lambda}^{k+\frac{1}{2}}+\beta B_i^{\mathsf{T}}(\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b})-\\
&\qquad\beta\sum_{l\ne i,l=1}^{q}B_i^{\mathsf{T}}B_l(\boldsymbol{y}_l^{k+1}-\boldsymbol{y}_l^k)+L_i(\boldsymbol{y}_i^{k+1}-\boldsymbol{y}_i^k)\Big\rangle\ge 0,\quad\forall\boldsymbol{y}_i\in\mathcal{Y}_i.
\end{aligned}$$

After adding the above inequality from $i=1$ to q, we can see $\boldsymbol{y}^{k+1}$ satisfies the first-order optimality condition, hence is a solution, of the following problem:

$$\boldsymbol{y}^{k+1}\in\operatorname*{argmin}_{\boldsymbol{y}\in\mathcal{Y}}g(\boldsymbol{y})+\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^k\right\|_{\widetilde{L}}^2+\frac{\beta}{2}\left\|A\boldsymbol{x}^{k+1}+B\boldsymbol{y}-\boldsymbol{b}-\frac{\boldsymbol{\lambda}^{k+\frac{1}{2}}}{\beta}\right\|^2,$$

Algorithm 5 Multi-block extension of SAS-ADMM in partially Jacobi update

Parameters: $\beta>0,(\tau,s)\in\Delta$ and $L_i\succeq(q-1)\beta B_i^{\mathsf{T}}B_i$ for all $i=1,\cdots,q$.
Initialization: $(\boldsymbol{x}^0,\boldsymbol{y}^0,\boldsymbol{\lambda}^0)\in\mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n$, $\check{\boldsymbol{x}}^0=\boldsymbol{x}^0$.
For $k=0,1,\cdots$
Choose $m_k>0$, $\eta_k>0$ and $\mathcal{M}_k$ such that $\mathcal{M}_k-\beta A^{\mathsf{T}}A\succeq\mathbf{0}$.
$\boldsymbol{h}^k:=-A^{\mathsf{T}}\left[\boldsymbol{\lambda}^k-\beta(A\boldsymbol{x}^k+B\boldsymbol{y}^k-\boldsymbol{b})\right]$.
$(\boldsymbol{x}^{k+1},\check{\boldsymbol{x}}^{k+1})=$ **xsub** $(\boldsymbol{x}^k,\check{\boldsymbol{x}}^k)$ with **xsub** given in ALG.1.
$\boldsymbol{\lambda}^{k+\frac{1}{2}}=\boldsymbol{\lambda}^k-\tau\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^k-\boldsymbol{b}\right)$.
For $i=1,2,\cdots,q$,

$$\boldsymbol{y}_i^{k+1}\in\arg\min_{\boldsymbol{y}_i\in\mathcal{Y}_i} g_i(\boldsymbol{y}_i)+\frac{\beta}{2}\left\|A\boldsymbol{x}^{k+1}+B_i\boldsymbol{y}_i+\sum_{l\neq i,l=1}^{q}B_l\boldsymbol{y}_l^k-\boldsymbol{b}-\frac{\boldsymbol{\lambda}^{k+\frac{1}{2}}}{\beta}\right\|^2+\frac{1}{2}\left\|\boldsymbol{y}_i-\boldsymbol{y}_i^k\right\|_{L_i}^2.$$

end
$\boldsymbol{\lambda}^{k+1}=\boldsymbol{\lambda}^{k+\frac{1}{2}}-s\beta\left(A\boldsymbol{x}^{k+1}+B\boldsymbol{y}^{k+1}-\boldsymbol{b}\right)$.
end

where

$$\widetilde{L}=\begin{bmatrix} L_1 & -\beta B_1^{\mathsf{T}}B_2 & \cdots & -\beta B_1^{\mathsf{T}}B_q \\ -\beta B_2^{\mathsf{T}}B_1 & L_2 & \cdots & -\beta B_2^{\mathsf{T}}B_q \\ \vdots & \vdots & \ddots & \vdots \\ -\beta B_q^{\mathsf{T}}B_1 & -\beta B_q^{\mathsf{T}}B_2 & \cdots & L_q \end{bmatrix}. \tag{A.27}$$

So, by a similar analysis to Algorithm 4, the inequality (A.19) holds with

$$\widetilde{Q}_k=\left[\begin{array}{c|cccc|c} \mathcal{D}_k & & & & & \\ \hline & L_1+(1-\frac{\tau s}{\tau+s})\beta B_1^{\mathsf{T}}B_1 & -\frac{\tau s}{\tau+s}\beta B_1^{\mathsf{T}}B_2 & \cdots & -\frac{\tau s}{\tau+s}\beta B_1^{\mathsf{T}}B_q & -\frac{\tau}{\tau+s}B_1^{\mathsf{T}} \\ & -\frac{\tau s}{\tau+s}\beta B_2^{\mathsf{T}}B_1 & L_2+(1-\frac{\tau s}{\tau+s})\beta B_2^{\mathsf{T}}B_2 & \cdots & -\frac{\tau s}{\tau+s}\beta B_2^{\mathsf{T}}B_q & -\frac{\tau}{\tau+s}B_2^{\mathsf{T}} \\ & \vdots & \vdots & \ddots & \vdots & \vdots \\ & -\frac{\tau s}{\tau+s}\beta B_q^{\mathsf{T}}B_1 & -\frac{\tau s}{\tau+s}\beta B_q^{\mathsf{T}}B_2 & \cdots & L_q+(1-\frac{\tau s}{\tau+s})\beta B_q^{\mathsf{T}}B_q & -\frac{\tau}{\tau+s}B_q^{\mathsf{T}} \\ \hline & -\frac{\tau}{\tau+s}B_1 & -\frac{\tau}{\tau+s}B_2 & \cdots & -\frac{\tau}{\tau+s}B_q & \frac{1}{(\tau+s)\beta}\boldsymbol{I} \end{array}\right],$$

$$\widetilde{G}_k=\left[\begin{array}{c|cccc|c} \mathcal{D}_k & & & & & \\ \hline & L_1+(1-s)\beta B_1^{\mathsf{T}}B_1 & -s\beta B_1^{\mathsf{T}}B_2 & \cdots & -s\beta B_1^{\mathsf{T}}B_q & (s-1)B_1^{\mathsf{T}} \\ & -s\beta B_2^{\mathsf{T}}B_1 & L_2+(1-s)\beta B_2^{\mathsf{T}}B_2 & \cdots & -s\beta B_2^{\mathsf{T}}B_q & (s-1)B_2^{\mathsf{T}} \\ & \vdots & \vdots & \ddots & \vdots & \vdots \\ & -s\beta B_q^{\mathsf{T}}B_1 & -s\beta B_q^{\mathsf{T}}B_2 & \cdots & L_q+(1-s)\beta B_q^{\mathsf{T}}B_q & (s-1)B_q^{\mathsf{T}} \\ \hline & (s-1)B_1 & (s-1)B_2 & \cdots & (s-1)B_q & \frac{2-\tau-s}{\beta}\boldsymbol{I} \end{array}\right].$$

If $L_i \succeq (q-1)\beta B_i^{\mathsf{T}} B_i$ for $i=1,\cdots,q$, then for any $(\tau,s)\in\Delta$ defined by (3.1), the above matrix $\widetilde{Q}_k$ is positive semidefinite and

$$\begin{aligned}\left\|\boldsymbol{w}^k-\widetilde{\boldsymbol{w}}^k\right\|_{\widetilde{G}_k}^2 \geq &\left\|\boldsymbol{x}^k-\boldsymbol{x}^{k+1}\right\|_{\mathcal{D}_k}^2+\omega_0\left\|\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b}\right\|^2\\ &+\omega_1\left(\left\|\mathcal{K}\boldsymbol{w}^{k+1}-\boldsymbol{b}\right\|^2-\left\|\mathcal{K}\boldsymbol{w}^{k}-\boldsymbol{b}\right\|^2\right)\\ &+\omega_2\left(\left\|\boldsymbol{y}^{k+1}-\boldsymbol{y}^{k}\right\|_{\widetilde{L}}^2-\left\|\boldsymbol{y}^{k}-\boldsymbol{y}^{k-1}\right\|_{\widetilde{L}}^2\right),\end{aligned}$$

where $\omega_0,\omega_1,\omega_2\geq 0$ is defined in (4.16) and $\widetilde{L}$ is defined in (A.27). The above discussions imply that Algorithm 5 has the same convergence properties as Algorithm 4 and can be also considered as a stochastic extension of the deterministic GS-ADMM [1] for solving the grouped multi-block separable convex optimization problem.

References

[1] J. Bai, J. Li, F. Xu and H. Zhang, Generalized symmetric ADMM for separable convex optimization, Comput. Optim. Appl., 70 (2018), 129-170.

[2] J. Bai, X. Chang, J. Li and F. Xu, Convergence revisit on generalized symmetric ADMM, Optimization, 70 (2021), 149-168.

[3] J. Bai, W. Hager and H. Zhang, Accelerated stochastic ADMM for separable convex optimization, Comput. Optim. Appl., (2022), DOI:10.1007/s10589-021-00338-8.

[4] J. Bai, Y. Ma, H. Sun and M. Zhang, Iteration complexity analysis of a partial LQP-based alternating direction method of multipliers, Appl. Numer. Math., 165 (2021), 500-518.

[5] X. Cai, D. Han and X. Yuan, On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function, Comput. Optim. Appl., 66 (2017), 39-73.

[6] X. Chang, J. Bai, D. Song and S. Liu, Linearized symmetric multi-block ADMM with indefinite proximal regularization and optimal proximal parameter, Calcolo, 57 (2020), 1-36.

[7] Y. Dai, D. Han, X. Yuan, and W. Zhang, A sequential updating scheme of the lagrange multiplier for separable convex programming, Math. Comput., 86 (2017), 315-343.

[8] J. Eckstein and D. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, Math. Program., 55 (1992), 293-318.

[9] X. Fang, B. He, H. Liu and X. Yuan, Generalized alternating direction method of multipliers: New theoretical insights and applications, Math. Prog. Comp., 7 (2015), 149-187.

[10] J. Friedman, T. Hastie and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 9 (2008), 432-441.

[11] R. Glowinski and A. Marrocco, Approximation par*éléments* finis d'rdre un et r*ésolution*, par p*énalisation-dualité* d'une classe de probl*èmes* de Dirichlet non lin*éaires*, Rev. Fr. Autom. Inform. Rech. Op*ér.* Anal. Num*ér.*, 2 (1975), 41-76.

[12] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximations, Comput. Math. Appl., 2 (1976), 17-40.

[13] T. Goldstein, B. Donoghue, S. Setzer and R. Baraniuk, Fast alternating direction optimization methods, SIAM J. Imaging Sci., 7 (2014), 1588-1623.

[14] Y. Gu, B. Jiang and D. Han, A semi-proximal-based strictly contractive Peaceman-Rachford splitting method, arXiv:1506.02221, (2015), 1-20.
[15] G. Gu, B. He and J. Yang, Inexact alternating-direction-based contraction methods for separable linearly constrained convex optimization, J. Optim. Theory Appl., 163 (2014), 105-129.
[16] W. Hager and H. Zhang, Inexact alternating direction multiplier methods for separable convex optimization, Comput. Optim. Appl., 73 (2019), 201-235.
[17] D. Han, A hybrid entropic proximal decomposition method with self-adaptive strategy for solving variational inequality problems, Comput. Math. Appl., 55 (2008), 101-115.
[18] B. He, H. Liu, Z. Wang and X. Yuan, A strictly contractive Peaceman-Rachford splitting method for convex programming, SIAM J. Optim., 24 (2014), 1011-1040.
[19] B. He, F. Ma and X. Yuan, Convergence study on the symmetric version of ADMM with larger step sizes, SIAM J. Imaging Sci., 9 (2016), 1467–1501.
[20] M. Hong and Z. Luo, On the linear convergence of alternating direction method of multipliers, Math. Program., 162 (2017), 165-199.
[21] F. Jiang, Z. Wu and X. Cai, Generalized ADMM with optimal indefinite proximal term for linearly constrained convex optimization, J. Indust. Manag. Optim., 16 (2020), 835-856.
[22] Z. Jia, X. Gao, X. Cai and D. Han, The convergence rate analysis of the symmetric ADMM for the nonconvex separable optimization problems, J. Indust. Manag. Optim., 17 (2021), 1943-1971.
[23] T. Lin, S. Ma and S. Zhang, On the global linear convergence of the ADMM with multi-block variables, SIAM J. Optim., 25 (2015), 1478-1497.
[24] Y. Liu, F. Shang, and J. Cheng, Accelerated variance reduced stochastic ADMM, in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI17, AAAI Press, 2017, 2287-2293.
[25] G. Luo and Q. Yang, A fast symmetric alternating direction method of multipliers, Numer. Math. Theor. Meth. Appl., 13 (2020), 200-219.
[26] H. Ouyang, N. He, L. Tran and A. Gray, Stochastic alternating direction method of multipliers, Proc. 30th Int. Conf. Mach. Learn., (2013), 80-88.
[27] D. Peaceman and H. Rachford JR., The numerical solution of parabolic and elliptic differential equations, J. Soc. Indust. Appl. Math., 3 (1955), 28-41.
[28] H. Sun, M. Sun and Y. Wang, Proximal ADMM with larger step size for two-block separable convex programming and its application to the correlation matrices calibrating problems, J. Nonlinear Sci. Appl., 10 (2017), 5038-5051.
[29] Z. Wu and M. Li, An LQP-based symmetric alternating direction method of multipliers with larger step sizes, J. Oper. Res. Soc. China, 7 (2019), 365-383.
[30] Y. Xiao, L. Chen and D. Li, A generalized alternating direction method of multipliers with semi-proximal terms for convex composite conic programming, Math. Prog. Comp., 10 (2018), 533-555.
[31] M. Xu and T. Wu, A class of linearized proximal alternating direction methods, J. Optim. Theory Appl., 151 (2011), 321-337.
[32] Z. Yang and Z. Yan, Fast linearized alternating direction method of multipliers for the augmented l_1-regularized problem, SIViP., 9 (2015), 1601-1612.
[33] X. Yuan, S. Zeng and J. Zhang, Discerning the linear convergence of ADMM for sturcuted convex optimization through the lens of cariational analysis, J. Mach. Learn. Res., 21 (2020), 1-74.
[34] W. Yang and D. Han, Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems, SIAM J. Numer. Anal., 54 (2016), 625-640.

[35] S. Zhao, W. Li and Z. Zhou, Scalable stochastic alternating direction method of multipliers, arXiv:1502.03529, (2015), 1-24.
[36] N. Zhang, J. Wu, L. Zhang, A linearly convergent majorized ADMM with indefinite proximal terms for convex composite programming and its applications, Math. Comput., 89 (2020), 1867-1894.