CONVERGENCE ANALYSIS OF AN ACCELERATED STOCHASTIC ADMM WITH LARGER STEPSIZES * 
JIANCHAO BAI1, DEREN HAN2, HONGCHAO ZHANG3, AND HAO SUN4

Abstract. In this paper, we develop an accelerated stochastic Alternating Direction Method of Multipliers (ADMM) for solving the structured convex optimization problem whose objective is the sum of smooth and nonsmooth terms. The proposed algorithm uses the ideas of both ADMM and accelerated stochastic gradient method, and the involved dual variables are updated twice with a larger stepsize region than the so-called \((0, \sqrt{\frac{L_f}{B}}])\). To simplify computation of the nonsmooth subproblem, we add a nonnegative proximal term to transform the subproblem into a proximal mapping. By a variational analysis, we show that the proposed stochastic algorithm converges in expectation with the worst-case \(O(1/T)\) convergence rate, where \(T\) denotes the number of outer iterations. Convergence rates of a special stochastic augmented Lagrange method and 3-block extensions of the proposed algorithm are also investigated. Numerical experiments on testing a big-data problem in machine learning verify that the proposed method is effective and promising.

Key words. convex optimization, stochastic ADMM, larger stepsize, proximal term, complexity

AMS subject classifications. 65K10, 68W40, 90C25

1. Introduction. Consider a family of “smooth + nonsmooth” structured convex optimization problems

\[
\min \{ f(\mathbf{x}) + g(\mathbf{y}) \mid \mathbf{Ax} + \mathbf{By} = \mathbf{b}, \ \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y} \},
\]

where \(\mathcal{X} \subset \mathbb{R}^{n_1}\), \(\mathcal{Y} \subset \mathbb{R}^{n_2}\) are closed convex subsets, \(g : \mathcal{Y} \to \mathbb{R} \cup \{ +\infty \} \) is a convex but possibly nonsmooth function, \(A \in \mathbb{R}^{n \times n_1}\), \(B \in \mathbb{R}^{n \times n_2}\), \(\mathbf{b} \in \mathbb{R}^n\) are given data, and \(f\) is an average of component convex functions:

\[
f(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} f_j(\mathbf{x})
\]

with \(f_j : \mathcal{X} \to \mathcal{R}\) convex and Lipschitz continuously differentiable. Problem (1.1) corresponds to regularized empirical risk minimization in big-data applications [16, 22], including classification and regression models in machine learning, where \(N\) denotes the sample size and \(f_j\) is the empirical loss. A major difficulty to solve (1.1) is that \(N\) can be huge so that it would be very expensive to evaluate either \(f\) or its full gradient.

The augmented Lagrangian function of (1.1) is given by

\[
L_\beta(\mathbf{x}, \mathbf{y}, \lambda) = L(\mathbf{x}, \mathbf{y}, \lambda) + \frac{\beta}{2} \| \mathbf{Ax} + \mathbf{By} - \mathbf{b} \|^2,
\]

*Feb. 13, 2020. This research was supported by the Fundamental Research Funds for the Central Universities (G2019KY05106), the National Natural Science Foundation of China (11625105) and the USA National Science Foundation (1522654, 1819161).

1jianchaobai@nwpu.edu.cn, https://teacher.nwpu.edu.cn/jcbai, Department of Applied Mathematics and the MIIT Key Laboratory of Dynamics and Control of Complex Systems, Northwestern Polytechnical University, Xi’an 710129, China.

2handr@buaa.edu.cn, http://math.buaa.edu.cn/szdw/azcck/js/hdr.htm, School of Mathematics and System Sciences, Beihang University, Beijing 100191, China

3hoxhang@math.lsu.edu, http://www.math.lsu.edu/~hoxhang, Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, Phone (225) 578-1982, Fax (225) 578-4276.

4hsun@nwpu.edu.cn, Department of Applied Mathematics Northwestern Polytechnical University, Xi’an 710129, China.
where $\beta > 0$ is the penalty parameter, $\lambda$ is the Lagrange multiplier and

\[ L(x, y, \lambda) = f(x) + g(y) - \lambda^T (Ax + By - b). \]

Although the Augmented Lagrange Method (ALM) can be used to handle Problem (1.1), without utilizing the separable structure of (1.1) it is time-consuming or even infeasible to derive closed-form solution of the core subproblem of ALM, especially when the constraints are complex. As a splitting iteration of ALM, the classic Alternating Direction Method of Multipliers (ADMM, [6, 7]) is an effective approach to exploit the separable structure of the problem, and its iterates read

\[
\begin{align*}
    x^{k+1} &= \arg \min_{x \in X} L_\beta(x, y^k, \lambda^k), \\
    y^{k+1} &= \arg \min_{y \in Y} L_\beta(x^{k+1}, y, \lambda^k), \\
    \lambda^{k+1} &= \lambda^k - s\beta (Ax^{k+1} + By^{k+1} - b),
\end{align*}
\]

where $s \in (0, \frac{1 + \sqrt{5}}{2})$ denotes the stepsize of the dual variable.

The above classic ADMM was proved convergent for the two-block case [7], but its direct extension for solving the cases with more than two variables are not necessarily convergent [4]. With the purpose of enlarging stepsize region of the dual variable for the work in [12], Gu, et al. [8] firstly constructed a symmetric proximal ADMM whose dual variable is updated twice with different stepsizes. He, et al. [13] proposed the following symmetric ADMM (S-ADMM) without proximal terms:

\[
\begin{align*}
    x^{k+1} &= \arg \min_{x \in X} L_\beta(x, y^k, \lambda^k), \\
    \lambda^{k+\frac{1}{2}} &= \lambda^k - \tau\beta (Ax^{k+1} + By^k - b), \\
    y^{k+1} &= \arg \min_{y \in Y} L_\beta(x^{k+1}, y, \lambda^{k+\frac{1}{2}}), \\
    \lambda^{k+1} &= \lambda^{k+\frac{1}{2}} - s\beta (Ax^{k+1} + By^{k+1} - b),
\end{align*}
\]

where $(\tau, s)$ denotes the stepsizes parameter satisfying

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

Moreover, it was shown by testing the basis pursuit problem and the total-variational image deblurring problem that using relatively larger stepsizes could accelerate convergence of S-ADMM numerically. Based on the seminal work [8, 13], Bai, et al. [1] further designed a generalized S-ADMM for a grouped multi-block separable convex optimization and enlarged the above region $\Delta_0$ to

\[ \Delta = \left\{ (\tau, s) \mid \tau + s > 0, \tau \leq 1, -\tau^2 - s^2 - \tau s + s + 1 > 0 \right\}. \]

To the best of our knowledge, this $\Delta$ is the largest convergence range so far for symmetric ADMM-type algorithms and it has been used to design the logarithmic-quadratic proximal based S-ADMM [18]. Most of deterministic ADMM-type algorithms [1, 9, 11, 17, 19, 20] enjoy global convergence with the worst-case $O(1/t)$ convergence rate for the convex optimization such as (1.1), where $t$ denotes the iteration number. Under the mild assumption that the subdifferential of each component objective function is piecewise linear, Yang-Han [21] established the global linear convergence of ADMM for two-block separable convex optimization. Assuming that an error bound condition holds, the dual stepsize is sufficiently small and each coefficient
Parameters: \( \beta > 0, \mathcal{H} \succeq 0, L \succeq 0 \) and \((\tau, s) \in \Delta\) given by (1.3);
Initialization: \((x^0, y^0, \lambda^0) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n, \hat{x}^0 = x^0\);
For \( k = 0, 1, \ldots \)

Choose \( m_k > 0, \eta_k \in (0, \frac{1}{m_k}], \mathcal{M}_k \) such that \( \mathcal{M}_k - \beta A^T A \succeq 0; \)
\[ h^k := -A^T \left[ \lambda^k - \beta (Ax^k + By^k - b) \right]; \]
\[ (x^{k+1}, \hat{x}^{k+1}) = \text{xsub} \ (x^k, \hat{x}^k, h^k); \]
\[ \lambda^{k+\frac{1}{2}} = \lambda^k - \tau \beta (Ax^{k+1} + By^{k+1} - b); \]
\[ y^{k+1} = \arg\min_{y \in \mathcal{Y}} \left( \mathcal{L}_\beta \left( x^{k+1}, y, \lambda^{k+\frac{1}{2}} \right) + \frac{1}{2} \|y - y^k\|_L^2 \right); \]
\[ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \beta \left( Ax^{k+1} + By^{k+1} - b \right). \]

\( (x_{t+1}, \hat{x}_{t+1}) = \text{xsub} \ (x_t, \hat{x}_t, h_t) \)
For \( t = 1, 2, \ldots, m_k \)

Randomly select \( \xi_t \in \{1, 2, \ldots, N\} \) with uniform probability;
\( \beta_t = 2/(t+1), \gamma_t = 2/(tn_k), \tilde{x}_t = \beta_t \hat{x}_t + (1 - \beta_t)x_t; \)
\( d_t = \tilde{g}_t + e_t, \) where \( \tilde{g}_t = \nabla f_{\xi_t}(\hat{x}_t) \) and \( e_t \) is a random vector
satisfying \( \mathbb{E}[e_t] = 0; \)
\[ \hat{x}_{t+1} = \arg\min \left\{ \langle d_t + h, x \rangle + \frac{\gamma_t}{2} \|x - \hat{x}_t\|^2_H + \frac{1}{2} \|x - x^k\|_{\mathcal{M}_k}^2 : x \in \mathcal{X} \right\}; \]
\[ x_{t+1} = \beta_t \hat{x}_{t+1} + (1 - \beta_t)x_t. \]

\textbf{Alg. 1.1. Accelerated stochastic ADMM with large stepsizes (las-ADMM)}

matrix in the equality constraint has full column rank, Hong-Luo \[14\] showed a linear
convergence rate of their multi-block ADMM. More details about linear convergence
rate under strongly convex assumption can be found in e.g. \[3, 10, 15\].

Motivated by the interesting work \[2\] and the symmetric scheme \[1\], we would
study convergence properties of an accelerated stochastic ADMM with large stepsizes
(denoted by “las-ADMM”) as shown in Algorithm 1.1, which only requires the gradient
of a subset of the component functions \( f_i \) instead of its full gradient at each
iteration. Major features of our las-ADMM are summarized as three aspects:

- Firstly, las-ADMM inherits all merits of AS-ADMM \[2\], that is, it combines a
variant of the Nesterov’s accelerated gradient method with variance reduction
technique. It is assumed that the possibly nonsmooth function \( g(y) \) is proper closed
with a nonempty proximal mapping, while \( x\)-subproblem is solved by
an accelerated stochastic gradient method and has a closed-form solution.

- Secondly, unlike the classic ADMM the dual variable of las-ADMM is updated
twice with relatively larger stepsize region (this means the update of the dual
variable is more balance and flexible). If the \( x\)-subproblem is solved exactly
and \( \tau = 0 \), then las-ADMM becomes the classic ADMM with \( s \in (0, \frac{1.5}{\sqrt{2}}) \).
Note that the \( y\)-subproblem is solved by adding a proper proximal term for
simplifying its computation. A concrete case that constructing proximal
matrix \( L \) can be found in Lemma 3.2, which converts the \( y\)-subproblem to
the following proximal mapping for any \( \gamma > 0 \):

\[
\arg\min_{y \in \mathcal{Y}} g(y) + \frac{\gamma \sigma}{2} \left\| y - y^k - \beta B^T (Ax^{k+1} + By^k - b - \frac{\lambda^{k+1}}{\beta}) \right\|^2.
\]

- Finally, viewed from the point of variational inequality, we show the proposed stochastic algorithm converges in expectation with a sublinear convergence rate. Besides, we analyze some convergence properties of a special stochastic ALM for solving the one-block case and two extended stochastic ADMMs for solving the three-block case.

2. Preliminaries.

2.1. Notation and assumption. We follow the notations used in the previous work [2]. Let \( \mathcal{R} \), \( \mathcal{R}^n \), and \( \mathcal{R}^{n \times l} \) be the sets of real numbers, \( n \)-dimensional real column vectors, and \( n \times l \)-dimensional real matrices, respectively. The bold \( \mathbf{I} \) and \( \mathbf{0} \) denote the identity matrix and the zero matrix/vector, respectively. For 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 \( \| x \|^2_G := x^T G x \) and \( \| x \|_G := \sqrt{x^T G x} \) if \( G \succeq 0 \). We use \( \| \cdot \| \) to denote the standard Euclidean norm equipped with inner product \( \langle \cdot, \cdot \rangle \), \( \nabla f(x) \) to represent the gradient of \( f \) at \( x \) and \( E[\cdot] \) to denote mathematical expectation. Define

\[
(2.1) \quad u = \begin{pmatrix} x \\ y \end{pmatrix}, \quad w = \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix}, \quad \mathcal{J}(w) = \begin{pmatrix} -A^T \lambda \\ -B^T \lambda \\ Ax + By - b \end{pmatrix},
\]

and

\[
(2.2) \quad u^k = \begin{pmatrix} x^k \\ y^k \end{pmatrix}, \quad w^k = \begin{pmatrix} x^k \\ y^k \\ \lambda^k \end{pmatrix}, \quad \mathcal{J}(w^k) = \begin{pmatrix} -A^T \lambda^k \\ -B^T \lambda^k \\ Ax^k + By^k - b \end{pmatrix}.
\]

We also define

\[
F(w) = f(x) + g(y).
\]

Throughout the context, we make the following assumptions:

Assumption 2.1. All component functions \( f_j, j = 1, \ldots, N \) are continuously differentiable on \( \mathcal{X} \) and the solution set of the problem (1.1) is nonempty.

Assumption 2.2. The gradient \( \nabla f \) satisfies the Lipschitz condition

\[
(2.3) \quad \| \nabla f(x) - \nabla f(y) \|_{\mathcal{H}^{-1}} \leq \nu \| x - y \|_{\mathcal{H}}
\]

for any \( x, y \in \mathcal{X} \), where \( \nu > 0 \) and \( \mathcal{H} \succeq 0 \).

The first assumption is a basic assumption to illustrate the solvability of the problem. Under Assumption 2.2, it holds for any \( x, y \in \mathcal{X} \) that

\[
f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\nu}{2} \| x - y \|_{\mathcal{H}}^2,
\]

which is used to prove the forthcoming Lemma 2.1.
2.2. Optimality condition of (1.1). Denote $\Omega = \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n$. It is well-known in optimization that a point $w^* := (x^*, y^*, \lambda^*) \in \Omega$ is called the saddle-point of $L(x, y, \lambda)$ if it satisfies

\begin{equation}
L(x^*, y^*, \lambda) \leq L(x^*, y^*, \lambda^*) \leq L(x, y, \lambda^*), \quad w \in \Omega,
\end{equation}

which is equivalent to

\begin{equation*}
\begin{cases}
f(x) - f(x^*) + (x - x^*)^T(-A^T \lambda^*) \geq 0, \\
g(y) - g(y^*) + (y - y^*)^T(-B^T \lambda^*) \geq 0, \\
Ax^* + By^* - b = 0.
\end{cases}
\end{equation*}

Rewriting these inequalities as a more compact variational inequality gives

\begin{equation}
F(w) - F(w^*) + (w - w^*)^T J(w^*) \geq 0, \quad w \in \Omega.
\end{equation}

Notice that the affine mapping $J(\cdot)$ is skew-symmetric, so we have

\begin{equation}
(w - w)^T [J(w) - J(w)] \equiv 0, \quad \forall w, w' \in \Omega.
\end{equation}

Hence, (2.5) is rewritten as

\begin{equation}
F(w) - F(w^*) + (w - w^*)^T J(w) \geq 0, \quad w \in \Omega.
\end{equation}

Next, we recall a basic lemma before showing that the iterates generated by Algorithm 1.1 satisfy the variational-like inequality (2.7) with extra terms converging to zero in expectation.

**Lemma 2.1.** [2, Lemma 3.5] Let $\delta_k = \nabla f(x_k) - d_k$ and $D_k = M_k + \beta A^T A$. Then, the iterates generated by Algorithm 1.1 satisfy

\begin{equation}
f(x) - f(x^{k+1}) + \langle x - x^{k+1}, A^T \lambda_k \rangle \geq \langle x^{k+1} - x, D_k(x^{k+1} - x) \rangle + \zeta^k
\end{equation}

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

\begin{equation}
\hat{\lambda}_k = \lambda_k - \beta (Ax^{k+1} + By^k - b), \quad \text{and}
\end{equation}

\begin{equation}
\zeta^k = \frac{2}{m_k(m_k + 1)} \left[ \frac{1}{\eta_k} \left( \|x - x^{k+1}\|^2_H - \|x - x^k\|^2_H \right) - \sum_{t=1}^{m_k} t(\delta_t, \tilde{x}_t - x) - \frac{\eta_k}{4(1 - \eta_k \nu)} \sum_{t=1}^{m_k} t^2 \|\delta_t\|_{\mathcal{H}^{-1}}^2 \right].
\end{equation}

**Theorem 2.2.** The iterates generated by Algorithm 1.1 satisfy

\begin{equation}
F(w) - F(\tilde{w}^k) + \langle w - \tilde{w}^k, J(w) \rangle \geq (w - \tilde{w}^k)^T Q_k (w^k - \tilde{w}^k) + \zeta^k
\end{equation}

for all $w \in \Omega$, where $\zeta^k$ and $\hat{\lambda}$ are given by Lemma 2.1, and

\begin{equation}
\tilde{w}^k := \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \\ \hat{\lambda}_k \end{pmatrix} := \begin{pmatrix} x^{k+1} \\ y^{k+1} \\ \lambda_k \end{pmatrix}, \quad Q_k = \begin{bmatrix} D_k & L + \beta B^T B & -\tau B^T \\ -B & \frac{1}{\beta} I \end{bmatrix}.
\end{equation}
Proof. By the first-order optimality condition of \( y \)-subproblem, we have
\[
(2.13) \quad g(y) - g(y^{k+1}) + \langle y - y^{k+1}, p_k \rangle \geq 0, \quad \forall y \in \mathcal{Y},
\]
with \( p_k \) the gradient of the smooth terms:
\[
p_k = -B^T \lambda^{k+\frac{1}{2}} + \beta B^T (A x^{k+1} + B y^{k+1} - b) + L (y^{k+1} - y^k)
\]
\[
= -B^T \lambda^{k+\frac{1}{2}} + \beta B^T (A x^{k+1} + B y^{k} - b) + [L + \beta B^T B] (y^{k+1} - y^k)
\]
\[
= -B^T \lambda^{k+\frac{1}{2}} + \beta B^T (A x^k + B y^{k} - b) + [L + \beta B^T B] (y^{k+1} - y^k)
\]
where we use the following relation
\[
(2.14) \quad \lambda^{k+\frac{1}{2}} = \lambda^k - \tau (\lambda^k - \bar{\lambda}^k).
\]
By the definition of \( \bar{\lambda}^k \), we have
\[
(2.15) \quad (A \bar{x}^k + B \bar{y}^k - b) - B (\bar{y}^k - y^k) + \frac{1}{\beta} (\bar{\lambda}^k - \lambda^k) = 0.
\]
So, the inequality (2.11) is achieved by combining (2.8), (2.13) and (2.15) together with (2.6). \( \square \)

3. Convergence analysis.

3.1. Basic lemmas and theorems. We give some fundamental lemmas and theorems before analyzing convergence of Algorithm 1.1.

**Corollary 3.1.** The iterates generated by Algorithm 1.1 satisfy
\[
F(w) - F(\tilde{w}^k) + (w - \tilde{w}^k)^T J(w)
\]
\[
\geq \frac{1}{2} \left\{ \|w - w^{k+1}\|_Q^2 - \|w - w^k\|_Q^2 + \|w^k - \tilde{w}^k\|_{\tilde{Q}_k}^2 \right\} + \zeta^k,
\]
where
\[
\tilde{Q}_k = \begin{bmatrix} \mathcal{D}_k & L + (1 - s) \beta B^T B - \frac{\tau + s}{\beta} B^T B \frac{(s - 1) B^T}{\beta} I \\ \frac{\tau + s}{\beta} B^T B \frac{(s - 1) B^T}{\beta} I & \mathcal{D}_k \end{bmatrix}, \quad \tilde{G}_k = \begin{bmatrix} \mathcal{D}_k & L + (1 - s) \beta B^T B - \frac{\tau + s}{\beta} B^T B \frac{(s - 1) B^T}{\beta} I \\ \frac{\tau + s}{\beta} B^T B \frac{(s - 1) B^T}{\beta} I & \mathcal{D}_k \end{bmatrix}.
\]
\[
(3.2)
\]
Proof. By (2.14) and the way of generating \( \lambda^{k+1} \), we have
\[
-s \beta B (y^k - \bar{y}^k) + (\tau + s) (\lambda^k - \bar{\lambda}^k) = \lambda^k - \lambda^{k+1},
\]
which, by the definition of \( \tilde{w}^k \) in (2.12), further shows
\[
(3.3) \quad w^k - w^{k+1} = P (w^k - \tilde{w}^k)
\]
with
\[
P = \begin{bmatrix} I & I \\ -s \beta B & (\tau + s) I \end{bmatrix}.
\]
Hence, the relation $Q_k(w^k - \tilde{w}^k) = Q_kP^{-1}(w^k - w^{k+1})$ holds and

$$Q_kP^{-1} = \begin{bmatrix} D_k & L + \left(1 - \frac{\tau s}{\tau + s}\right) \beta B^TB - \frac{\tau}{\tau + s} B \frac{1}{\beta(\tau + s)} \mathbf{I} \\ \frac{\tau}{\tau + s} B & \frac{1}{\beta(\tau + s)} \mathbf{I} \end{bmatrix} = \tilde{Q}_k.$$ 

So, for any $w \in \Omega$, it follows from (2.11) and the above relation that

$$F(w) - F(\tilde{w}^k) + (w - \tilde{w}^k)^T J(w) \geq \zeta^k + (w - \tilde{w}^k)^T \tilde{Q}_k (w - w^{k+1})$$

$$= \zeta^k + \frac{1}{2} \left\lVert w - w^{k+1} \right\rVert^2_{\tilde{Q}_k} - \left\lVert w - w^{k} \right\rVert^2_{\tilde{Q}_k} + \left\lVert w^k - \tilde{w}^k \right\rVert^2_{\tilde{Q}_k} - \left\lVert w^{k+1} - \tilde{w}^k \right\rVert^2_{\tilde{Q}_k},$$

where the last equality uses the identity

$$2(a - b)^T \tilde{Q}_k(c - d) = \left\lVert a - d \right\rVert^2_{\tilde{Q}_k} - \left\lVert a - c \right\rVert^2_{\tilde{Q}_k} + \left\lVert c - b \right\rVert^2_{\tilde{Q}_k} - \left\lVert b - d \right\rVert^2_{\tilde{Q}_k},$$

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

Now, by (3.4) again, we deduce

$$\left\lVert w^k - \tilde{w}^k \right\rVert^2_{\tilde{Q}_k} - \left\lVert w^{k+1} - \tilde{w}^k \right\rVert^2_{\tilde{Q}_k}$$

$$= \left\lVert w^k - \tilde{w}^k \right\rVert^2_{\tilde{Q}_k} - \left\lVert w^k + w^k - \tilde{w}^k \right\rVert^2_{\tilde{Q}_k}$$

$$= \left\lVert w^k - \tilde{w}^k \right\rVert^2_{\tilde{Q}_k} - \left\lVert w^k - P(w^k - \tilde{w}^k) \right\rVert^2_{\tilde{Q}_k}$$

where we uses the relation $\tilde{Q}_k = P^T \tilde{Q}_k + \tilde{Q}_kP - P^T \tilde{Q}_kP$ for the last equation. As a result, we complete the whole proof. \(\square\)

**Lemma 3.2.** Let $L = \gamma \sigma \mathbf{I} - \beta B^TB$ with $\sigma \geq \beta B^TB$. Then, the matrix $\tilde{Q}_k$ given by (3.2) is symmetric positive semidefinite if $\gamma \geq \tau$ for any $(\tau, s) \in \Delta$.

**Proof.** Clearly, we just need to investigate the property of the lower-upper 2-by-2 block of $\tilde{Q}_k$, that is

$$\tilde{Q}_k^L = \begin{bmatrix} L + \left(1 - \frac{\tau s}{\tau + s}\right) \beta B^TB - \frac{\tau}{\tau + s} B \\ \frac{1}{\beta(\tau + s)} \mathbf{I} \end{bmatrix} \geq \begin{bmatrix} \frac{1}{\beta(\tau + s)} \mathbf{I} \\ \frac{\tau}{\tau + s} B \beta^{-\frac{s}{s+1}} \left( \frac{\tau}{\tau + s} \right) \mathbf{I} - \frac{\tau}{\tau + s} B \frac{1}{\beta(\tau + s)} \mathbf{I} \end{bmatrix} \geq \begin{bmatrix} \beta^{-\frac{s}{s+1}} B \\ \beta^{-\frac{s}{s+1}} \mathbf{I} \end{bmatrix}.$$ 

Meanwhile, we note that

$$\begin{bmatrix} \mathbf{I} \\ \tau \mathbf{I} \end{bmatrix}^T \begin{bmatrix} \frac{1}{\beta(\tau + s)} \mathbf{I} \\ \frac{\tau}{\tau + s} B \frac{1}{\beta(\tau + s)} \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \tau \mathbf{I} \end{bmatrix} = \begin{bmatrix} \gamma - \tau \mathbf{I} \\ \frac{1}{\tau + s} \mathbf{I} \end{bmatrix}.$$

It is clear that $\tilde{Q}_k^L$ is positive semidefinite if $\tau + s > 0$ and $\gamma \geq \tau$. \(\square\)
Remark 3.1. Followed by Lemma 3.2 and the inequality (3.6), we can deduce \( \tau \leq \frac{5-3\tau}{1+\tau} \geq 1 \) for any \( \tau \in \Delta \). This implies the proximal matrix \( L \) is always positive semidefinite. However, for the special case \( \tau = 0 \), our Algorithm 1.1 becomes AS-ADMM [2] and hence could use a positive indefinite proximal matrix. The reason why we have to add a positive semidefinite matrix \( L \) for Algorithm 1.1 in stead of a positive indefinite matrix is possibly that the matrix \( Q_k \) is not diagonal after using larger stepsize range while the the original matrix in [2, Theorem 3.7] is diagonal.

Although Lemma 3.2 provides a sufficient condition to ensure the positive semidefiniteness of the matrix \( \tilde{Q}_k \), it is still unclear whether Algorithm 1.1 is convergent or not because \( \tilde{G}_k \) is not necessarily positive definite for any \((\tau, s) \in \Delta \). Therefore, we need to estimate the lower bound of \( \|w^k - \tilde{w}^k\|_{\tilde{G}_k}^2 \).

**Theorem 3.3.** Let \( L = \gamma \sigma \mathbf{I} - \beta B^T B \) with \( \sigma \geq \beta B^T B \). Then, for any 

\[
(3.6) \quad (\tau, s) \in \Delta \quad \text{and} \quad \gamma \geq \frac{5 - 3\tau}{1 + \tau},
\]

there exist \( \omega_0 > 0 \), \( \omega_i (i = 1, 2, 3) \geq 0 \) such that

\[
(3.7) \quad \|w^k - \tilde{w}^k\|_{\tilde{G}_k}^2 \geq \|x^k - x^{k+1}\|_{\tilde{G}_k}^2 + \omega_0 \|Ax^{k+1} + By^{k+1} - b\|^2 + \omega_1 \|B(y^k - y^{k+1})\|^2 + \omega_2 \left( \|Ax^{k+1} + By^{k+1} - b\|^2 - \|Ax^k + By^k - b\|^2 \right) + \omega_3 \left( \|B(y^k - y^{k+1})\|^2 - \|B(y^{k-1} - y^k)\|^2 \right),
\]

Especially, we have

\[
\omega_0 = \left( \frac{2 - \tau - s - (1-s)^2}{1+\tau} \right) \beta, \quad \omega_1 = \left( \frac{5 - 3\tau}{1+\tau} \right) \beta,
\]

\[
\omega_2 = \frac{(1-s)^2}{1+\tau} \beta \quad \text{and} \quad \omega_3 = \frac{(1-\tau)(\gamma+1)}{1+\tau} \beta.
\]

**Proof.** Using the structure of \( \tilde{G}_k \) in (3.2), the conditions of Lemma 3.2 and the Eq. (2.15) gives

\[
(3.8) \quad \|w^k - \tilde{w}^k\|_{\tilde{G}_k}^2 = \|x^k - x^{k+1}\|_{\mathbb{D}_k}^2 + \|y^k - y^{k+1}\|_{L \mathbf{1} + (1-s)\beta B^T B}^2 + 2(s-1) \left( \lambda^k - \tilde{\lambda}^k \right)^T B(y^k - y^{k+1}) + \frac{2 - \tau - s}{\beta} \|\lambda^k - \tilde{\lambda}^k\|^2 \geq \|x^k - x^{k+1}\|_{\mathbb{D}_k}^2 + (2 - \tau - s)\beta \|Ax^{k+1} + By^{k+1} - b\|^2 + (\gamma - \tau)\beta \times \|B(y^k - y^{k+1})\|^2 + 2(1-\tau)\beta \left( Ax^{k+1} + By^{k+1} - b \right)^T B(y^k - y^{k+1}).
\]

Next, we estimate the last cross term in the inequality of (3.8). Letting \( y = y^k \) in the first-order optimality condition (2.13), we have

\[
g(y^k) - g(y^{k+1}) + \left\langle y^k - y^{k+1}, -B^T \lambda^k - \frac{1}{2} + \beta B^T (Ax^{k+1} + By^{k+1} - b) + L(y^{k+1} - y^k) \right\rangle \geq 0.
\]

Similarly, by choosing \( y = y^{k+1} \) in the first-order optimality condition of \( y \)-subproblem at the \((k-1)\)-th iteration, we have

\[
g(y^{k+1}) - g(y^k) + \left\langle y^{k+1} - y^k, -B^T \lambda^{k+1} - \frac{1}{2} + \beta B^T (Ax^{k+1} + By^k - b) + L(y^k - y^{k-1}) \right\rangle \geq 0.
\]
By the conditions of Lemma 3.2 again, it can be deduced that
\[ \sum_{}^{} \text{above two inequalities and noticing } \]
\[ \lambda^{k+\frac{1}{2}} - \lambda^{k+\frac{1}{2}} = \tau \beta (Ax^{k+1} + By^{k+1} - b) + s \beta (Ax^k + By^k - b) + \tau \beta B(y^k - y^{k+1}), \]
we obtain
\[
\begin{align*}
(3.9) \quad & (Ax^{k+1} + By^{k+1} - b)^T B (y^k - y^{k+1}) \\
& \geq \frac{1 - s}{1 + \tau} (Ax^k + By^k - b)^T B (y^k - y^{k+1}) - \frac{\tau}{1 + \tau} \|B(y^k - y^{k+1})\|^2 \\
& + \frac{1}{(1 + \tau)} (y^k - y^{k+1})^T L [(y^k - y^{k+1}) - (y^{k-1} - y^k)].
\end{align*}
\]
By the conditions of Lemma 3.2 again, it can be deduced that
\[
\begin{align*}
(3.10) \quad & \frac{1}{\beta} (y^k - y^{k+1})^T L [(y^k - y^{k+1}) - (y^{k-1} - y^k)] \\
& = \frac{\gamma \sigma}{\beta} \|y^k - y^{k+1}\|^2 - ||B(y^k - y^{k+1})||^2 \\
& - \frac{\gamma \sigma}{\beta} (y^k - y^{k+1})^T (y^{k-1} - y^k) + (y^k - y^{k+1})^T B^T B (y^{k-1} - y^k) \\
& \geq (\gamma - 1) \|B(y^k - y^{k+1})\|^2 - \frac{\gamma + 1}{2} \left( \|B(y^k - y^{k+1})\|^2 + \|B(y^{k-1} - y^k)\|^2 \right) \\
& = \frac{\gamma - 3}{2} \|B(y^k - y^{k+1})\|^2 - \frac{\gamma + 1}{2} \|B(y^{k-1} - y^k)\|^2.
\end{align*}
\]
Substituting the above (3.10) and (3.9) into (3.8) gives
\[
\begin{align*}
& \|w^k - \bar{w}^k\|^2_{G_k} \\
& \geq \|x^k - x^{k+1}\|^2_{D_k} + (2 - \tau - s)\beta \|Ax^{k+1} + By^{k+1} - b\|^2 \\
& + \frac{2\beta}{1 + \tau} (1 - \tau)(1 - s) (Ax^k + By^k - b)^T B (y^k - y^{k+1}) \\
& + (\gamma - \tau)\beta \|B (y^k - y^{k+1})\|^2 - \frac{2(1 - \tau)}{1 + \tau} \beta \|B (y^k - y^{k+1})\|^2 \\
& + \frac{2(1 - \tau)}{1 + \tau} \left( \frac{\gamma - 3}{2} \|B(y^k - y^{k+1})\|^2 - \frac{\gamma + 1}{2} \|B(y^{k-1} - y^k)\|^2 \right) \\
& \geq \|x^k - x^{k+1}\|^2_{D_k} + \left( 2 - \tau - s \left( \frac{1 - s}{1 + \tau} \right)^2 \right) \beta \|Ax^{k+1} + By^{k+1} - b\|^2 \\
& + \frac{(1 - s)^2}{1 + \tau} \beta \left( \|Ax^{k+1} + By^{k+1} - b\|^2 - \|Ax^k + By^k - b\|^2 \right) \\
& + \frac{(1 - \tau)(\gamma + 1)}{1 + \tau} \beta \left( \|B(y^k - y^{k+1})\|^2 - \|B(y^{k-1} - y^k)\|^2 \right) \\
& + \left( \frac{\gamma - 5 - 3\tau}{1 + \tau} \right) \beta \|B(y^k - y^{k+1})\|^2,
\end{align*}
\]
where the second inequality uses the following Cauchy-Schwartz inequality
\[
\begin{align*}
& 2(1 - s)(1 - \tau) (Ax^k + By^k - b)^T B (y^k - y^{k+1}) \\
& \geq -(1 - s)^2 \|Ax^k + By^k - b\|^2 - (1 - \tau)^2 \|B (y^k - y^{k+1})\|^2.
\end{align*}
\]
For any \((\tau, s, \gamma)\) satisfying (3.6), we have
\[-\tau^2 - s^2 - \tau s + \tau + s + 1 > 0 \quad \text{and} \quad \tau \leq 1\]
which implies
\[2 - \tau - s - \frac{(1 - s)^2}{1 + \tau} > 0, \quad \frac{(1 - s)^2}{1 + \tau} \geq 0 \quad \text{and} \quad \frac{(1 - \tau)(\gamma - 1)}{1 + \tau} \geq 0.\]
This together with \(\beta > 0\) ends the proof. \(\square\)

3.2. Iteration complexity in expectation. In this subsection, we analyze the iteration complexity of Algorithm 1.1. We first have the following lemma.

Lemma 3.4. \([2, \text{Lemma 4.1}]\) For \(\delta^t\) defined in Lemma 2.1, we have

\[
\mathbb{E}[\delta^t] = 0. \tag{3.11}
\]

Theorem 3.5. Suppose that Assumptions 2.1-2.2 hold and the conditions of Theorem 3.3 hold, \(\mathcal{D}_k \supseteq \mathcal{D}_{k+1} \supseteq 0\) for all \(k \geq 0\), the sequence \(\{\eta_k m_k(m_k + 1)\}\) is nondecreasing, and \(\mathbb{E}(\|\delta_k\|^2)^t_0 \leq \sigma^2\) for some \(\sigma > 0\), independent of \(t\) and the iteration number \(k\). Then, for any integer \(T > 0\), there exist \(\omega_2, \omega_3 \geq 0\) such that

\[
\mathbb{E}

\left

\left\{ F(w_T) - F(w) + (w_T - w)^T J(w) \right\}

\leq \frac{1}{2T} \left\{ \sigma^2 \sum_{k=1}^T \eta_k m_k + \frac{4}{m_1 (m_1 + 1) \eta_1} \|x - \hat{x}\|^2_{\hat{Q}_1} + \|w - w^1\|^2_{\hat{Q}_1} + \omega_2 \|Ax^1 + By^1 - b\|^2 + \omega_3 \|B(y^1 - y^0)\|^2 \right\} \tag{3.12}
\]

for any \(w \in \Omega\), where

\[
w_T := \frac{1}{T} \sum_{k=1}^T \hat{w}^k. \tag{3.13}
\]

Proof. By the assumption, \(\mathcal{D}_k \supseteq \mathcal{D}_{k+1} \supseteq 0\) which implies that the matrix \(\hat{Q}_k\) in (3.2) satisfies \(\hat{Q}_k \supseteq \hat{Q}_{k+1} \supseteq 0\). Substituting (3.7) into (3.1) and utilizing the relation \(\hat{Q}_k \supseteq \hat{Q}_{k+1}\) are to achieve

\[
F(\hat{w}^k) - F(w) + (\hat{w}^k - w)^T J(w) \leq \\
- \hat{c}_k^1 \left\{ \|w - w^k\|^2_{\hat{Q}_k} - \|w - w^{k+1}\|^2_{\hat{Q}_{k+1}} \right\} \\
+ \frac{\omega_2}{2} \left( \|Ax^k + By^k - b\|^2 - \|Ax^{k+1} + By^{k+1} - b\|^2 \right) \\
+ \frac{\omega_3}{2} \left( \|B(y^{k+1} - y^k)\|^2 - \|B(y^k - y^{k+1})\|^2 \right).
\]

where \(\omega_2, \omega_3\) are given by Theorem 3.3. Summing the above inequality over \(k\) between 1 and \(T\), we deduce

\[
\sum_{k=1}^T F(\hat{w}^k) - T \left\{ F(w) + (w_T - w)^T J(w) \right\} \\
\leq - \sum_{k=1}^T \hat{c}_k^1 \left\{ \|w - w^1\|^2_{\hat{Q}_1} + \omega_2 \|Ax^1 + By^1 - b\|^2 + \omega_3 \|B(y^1 - y^0)\|^2 \right\}. \tag{3.14}
\]
By the convexity of $F$ and the definition of $w_T$ in (3.13), it follows that

\begin{equation}
(3.15) \quad F(w_T) \leq \frac{1}{T} \sum_{k=1}^{T} F(\overline{w}_k).
\end{equation}

Then, dividing (3.14) by $T$ and using (3.15), we achieve from (3.14) that

\begin{equation}
(3.16) \quad F(w_T) - F(w) + (w_T - w)^T \mathcal{J}(w) \leq \frac{1}{T} \sum_{k=1}^{T} -\zeta_k + \frac{1}{2} \left\{ \|w - w^k\|^2_{\hat{Q}_k} + 2\|A\hat{x}^k + B\hat{y} - b\|^2 + \omega_1 \|B(y^1 - y^0)\|^2 \right\}.
\end{equation}

Let us now focus on the terms involving $\zeta_k$. By assumption, the sequence $\{m_k(m_k + 1)\eta_k\}$ is nondecreasing and $H > 0$, thus we have

\begin{equation}
(3.17) \quad \sum_{k=1}^{T} \frac{2}{m_k(m_k + 1)\eta_k} \left( \|x - \hat{x}^k\|^2_{\hat{H}} - \|x - \hat{x}^{k+1}\|^2_{\hat{H}} \right) \leq \frac{2}{m_1(m_1 + 1)\eta_1} \|x - \hat{x}^1\|^2_{\hat{H}}.
\end{equation}

Also, since $\delta_t$ only depends on the index $\xi_t$ while $\hat{x}_t$ depends on $\xi_{t-1}$, $\xi_{t-2}$, ..., it follows from Lemma 3.4 that

$$
E[(\delta_t, \hat{x}_t - x)] = 0.
$$

Finally,

$$
E \left[ \sum_{t=1}^{T} \|\delta_t\|^2_{\hat{H}} \right] \leq \frac{\sigma^2 m(m + 1)(2m + 1)}{6} \left\{ \|x - \hat{x}^1\|^2_{\hat{H}} + \frac{\sigma^2}{2} \sum_{k=1}^{T} \eta_k m_k \right\}.
$$

Finally, applying the expectation operator to (3.16) and substituting this bound into the $\zeta_k$ term complete the proof. \(Q.E.D\)

**Corollary 3.6.** Suppose the conditions in Theorem 3.5 hold. Let

\begin{equation}
(3.18) \quad \eta_k = \min \left\{ \frac{c_1}{m_k(m_k + 1)\nu}, \frac{1}{2\nu} \right\} \quad \text{and} \quad m_k = \max \left\{ \left\lceil c_2(1 + k)^{\varrho} \right\rceil, m \right\},
\end{equation}

where $c_1, c_2 > 0$, $\varrho \geq 1$ are constants and $m > 0$ is a given integer. Then, for any $w \in \Omega$, we have

\begin{equation}
(3.19) \quad E \left[ F(w_T) - F(w) + (w_T - w)^T \mathcal{J}(w) \right] \leq \mathcal{O} \left( \frac{1}{T} \left( 1 + \sum_{k=1}^{T} \frac{1}{(1 + k)^{\varrho}} \right) \right).
\end{equation}

**Proof.** By the choices of $\eta_k$ and $m_k$ in (3.18), $\eta_k m_k(m_k + 1)$ is a constant for $k$ sufficiently large and $\sum_{k=1}^{\infty} \eta_k m_k < \infty$. The bound (3.19) follows from Theorem 3.5. \(Q.E.D\)

According to Corollary 3.6, if $\varrho > 1$, then $\sum_{k=1}^{T} \frac{1}{(1 + k)^{\varrho}}$ is bounded by a constant, independent of $T$, and the right-hand side of (3.19) goes to zero like $\mathcal{O}(1/T)$. If $\varrho = 1$, then $\sum_{k=1}^{T} \frac{1}{(1 + k)^{\varrho}} = \mathcal{O}(\log T)$ and the right-hand side of (3.19) goes to zero like $\mathcal{O}(\log T/T)$. 

4. Further discussions. In this section, we discuss a degradation and three-block extensions of Algorithm 1.1, and show their convergence concisely.

4.1. Stochastic ALM. Consider the following special convex problem

\[
\min \{ f(x) \mid Ax = b, \ x \in X \},
\]

where \( X \subset \mathbb{R}^n \) is a closed convex subset, and \( f \) is an average of component convex functions as mentioned in Section 1, that is \( f(x) = \frac{1}{N} \sum_{j=1}^{N} f_j(x) \) with \( f_j : \mathcal{X} \to \mathbb{R} \) being convex (not necessarily strongly convex) and Lipschitz continuously differentiable function, and \( A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n \) are given.

Define

\[
L_\beta(x, \lambda) := L(x, \lambda) + \frac{\beta}{2} \|Ax - b\|^2,
\]

where \( L(x, \lambda) = f(x) - \lambda^T(Ax - b) \). Then, motivated by Algorithm 1.1 we have an Accelerated Stochastic Augmented Lagrange Method (AS-ALM, see Algorithm 4.1).

**Lemma 4.1.** Let \( \{x^k\} \) be generated by Algorithm 4.1. Then, the inequality (2.8) holds with

\[
\tilde{\lambda}^k = \lambda^k - \beta (Ax^{k+1} - b).
\]

**Theorem 4.2.** Let \( \{w^k := (x^k, \lambda^k)\} \) be generated by Algorithm 4.1. Then, we have \( w^k := (x^k, \lambda^k) \in \Omega \) and

\[
f(x) - f(x^{k+1}) + \langle w - \tilde{w}^k, \mathcal{J}(w) \rangle \geq (w - \tilde{w}^k)^T Q_k(w^k - \tilde{w}^k) + \zeta^k
\]

for all \( w \in \Omega \), where \( \zeta^k \) is given by (2.10) and

\[
\mathcal{J}(w) = \begin{pmatrix}
-A^T A \\
-A^T b
\end{pmatrix}, \quad Q_k = \begin{bmatrix} D_k & -\frac{1}{\beta}I \\
\frac{1}{\beta}I & \frac{1}{\beta}I
\end{bmatrix}.
\]

**Proof.** Combining the inequality (2.8) and the relation \( Ax^{k+1} - b = \frac{1}{\beta} (\lambda^k - \tilde{\lambda}^k) \) gives the results. \( \square \)
**Theorem 4.3.** Let \( \{w^k\} \) be generated by Algorithm 4.1. Then, for any \( w \in \Omega \), we have

\[
    f(x) - f(x^{k+1}) + (w - \bar{w}^k)^T J(w) \\
    \geq \frac{1}{2} \left\{ \|w - w_k+1\|^2_{\mathcal{Q}_k} - \|w - w_k\|^2_{\mathcal{Q}_k} + \|w - \bar{w}_k\|^2_{\mathcal{G}_k} \right\} + \zeta_k,
\]

where \( \zeta_k \) is given by (2.10) and

\[
    \mathcal{Q}_k = \begin{bmatrix} \mathcal{D}_k & s_{\mathcal{Q}} \mathbf{I} \end{bmatrix}, \quad \mathcal{G}_k = \begin{bmatrix} \mathcal{D}_k & \frac{s_{\mathcal{G}}}{s_{\rho}} \mathbf{I} \end{bmatrix}.
\]

**Proof.** The proof is similar to that of Corollary 3.1 and is omitted here. \( \Box \)

Finally, according to Theorem 4.3 and the proof of Corollary 3.6, we can deduce for any \( w_T := \frac{1}{1+T} \sum_{k=1}^{T} w_k \) that

\[
    E \left[ f(x_T) - f(x) + (w_T - w)^T J(w) \right] \leq O \left( \frac{1}{T} \left( 1 + \sum_{k=1}^{T} \frac{1}{(1+k)^2} \right) \right).
\]

**4.2. Three-block extensions of Las-ADMM.** Consider the following 3-block convex optimization model

\[
    \min_{x \in \mathcal{X}, \ y \in \mathcal{Y}, \ z \in \mathcal{Z}} \quad F(u) := f(x) + g(y) + l(z)
    \]

subject to

\[
    Ku := Ax + By + Cz = b,
\]

where \( u = (z; x; y) \) is variable, \( l \) is a simple closed convex function, \( C \in \mathbb{R}^{n \times n_3} \) is a given matrix, \( \mathcal{Z} \subset \mathbb{R}^{n_3} \) is a convex subset, the functions \( f \) and \( g \) as well as the remaining notations are the same meanings as in Section 1. In applications, the function \( l \) could be used to promote a data structure different from the structure promoted by \( g \) at the solution.

**4.2.1. Extension in Gauss-Seidel update.** Assume \( C^TA = 0 \), then our Algorithm 1.1 can be directly extended to Algorithm 4.2 for solving the above 3-block convex problem (4.2), where the updating order is \( z^{k+1} \rightarrow x^{k+1} \rightarrow y^{k+1} \rightarrow \lambda^{k+1} \) in a Gauss-Seidel scheme. In this subsection, we denote

\[
    \tilde{\lambda}^k = \lambda^k - \beta \left( Cz^{k+1} + Ax^{k+1} + By^k - b \right), \quad w = (u; \lambda),
\]

and

\[
    \bar{w}^k := \begin{pmatrix} \bar{z}^k \\ \bar{x}^k \\ \bar{y}^k \\ \tilde{\lambda}^k \end{pmatrix} = \begin{pmatrix} z_{k+1}^k \\ x_{k+1}^k \\ y_{k+1}^k \\ \lambda^k \end{pmatrix}, \quad v^{k+1} = \begin{pmatrix} x_{k+1}^k \\ y_{k+1}^k \\ \bar{\lambda}^k \end{pmatrix}, \quad J(w) = \begin{pmatrix} -C^T \lambda \\ -A^T \lambda \\ -B^T \lambda \\ Ku - b \end{pmatrix}.
\]

Now, we only provide a key convergence theorem of Algorithm 4.2.

**Theorem 4.4.** Assume \( C^TA = 0 \). Then the iterates generated by Algorithm 4.2 satisfy \( \bar{w}^k \in \Omega \) and

\[
    F(u) - F(\bar{u}^k) + (w - \bar{w}^k)^T J(w) \\
    \geq \frac{1}{2} \left\{ \|v - v^{k+1}\|^2_{\mathcal{Q}_k} - \|v - v^k\|^2_{\mathcal{Q}_k} + \|v^k - \bar{v}^{k}\|^2_{\mathcal{G}_k} \right\} + \zeta^k
\]
and the assumption (4.8)

Combining the above inequalities (4.5), (2.8), (4.6) and (4.7), we get (4.7)

Besides, it follows from the updating of (4.6)

Similarly, we have by the updates of both $h^k$ and $\bar{x}^k$ in Algorithm 4.2, it is easy to obtain the same result (2.8) as before. From the first-order optimality condition of $z$-subproblem and the assumption $CTA = 0$, we have

(4.5) \[ z^{k+1} \in Z, \quad l(z) - l(z^{k+1}) + \langle z - z^{k+1}, p^k_z \rangle \geq 0, \quad \forall z \in Z, \]

where

\[ p^k_z = -CT\bar{x}^k + \beta CT(Cz^{k+1} + Ax^k + By^k - b) \]
\[ = -CT\bar{x}^k - \beta CT(A(x^{k+1} - x^k)) \]
\[ = -CT\bar{x}^k. \]

Similarly, we have by the $y$-update that

(4.6) \[ y^{k+1} \in Y, \quad g(y) - g(y^{k+1}) + \langle y - y^{k+1}, p^k_y \rangle \geq 0, \quad \forall y \in Y, \]

where

\[ p^k_y = -B^T\bar{x}^{k+1} + \beta B^T(Ku^{k+1} - b) + L(y^{k+1} - y^k) \]
\[ = -B^T\bar{x}^{k+1} + B^T(\bar{x}^k - \bar{x}^k) + [L + \beta BTB](y^{k+1} - y^k) \]
\[ = -B^T\bar{x}^k + \tau B^T(\bar{x}^k - \bar{x}^k) + [L + \beta BTB](y^{k+1} - y^k), \]

Besides, it follows from the updating of $\bar{x}^k$ that

(4.7) \[ \langle \lambda - \bar{x}^k, Ku^k - b + \frac{1}{\beta} (\bar{x}^k - x^k) - B(\bar{y}^k - y^k) \rangle = 0, \quad \forall \lambda \in \mathbb{R}^n. \]

Combining the above inequalities (4.5), (2.8), (4.6) and (4.7), we get

(4.8) \[ F(u) - F(\bar{u}^k) + (w - \bar{w}^k)^T J(\bar{y}^k) \geq (v - \bar{v}^k)^T Q_k(v^k - \bar{v}^k) + \zeta^k, \]
Analogous to the notations in Section 2, we denote \( y \) the \( z \)-subproblem and \( \tilde{x} \)-subproblem by linearizing their quadratic terms in order to get closed-form solutions of these subproblems. According to Theorem 4.4 and the remaining convergence analysis in Section 3.2, Algorithm 4.4 converges in expectation with a sublinear convergence rate. Here, we omit the detailed proof for the sake of conciseness.

4.2.2. Extension in partially Jacobi update. When both the matrices \( B \) and \( C \) are complicated, closed-form solutions of the \( y \)-subproblem and \( z \)-subproblem may not exist. In this case, we could again modify the \( y \)-subproblem and \( z \)-subproblem by linearizing their quadratic terms in order to get closed-form solutions of these subproblems. However, to ensure convergence of 3-block extensions, after solving the \( x \)-subproblem by a stochastic gradient method as before, we solve the modified \( y \)-subproblem and \( z \)-subproblem in a Jacobian scheme as stated in Algorithm 4.3. Analogous to the notations in Section 2, we denote

\[
\begin{align*}
\mathbf{w}^k &= \begin{pmatrix} x^k \\ y^k \\ z^k \\ \lambda^k \end{pmatrix}, \quad \tilde{\mathbf{u}}^k := \mathbf{u}^{k+1} = \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \\ \tilde{z}^k \\ \tilde{\lambda}^k \end{pmatrix}, \\
\tilde{\mathbf{w}}^k := \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \\ \tilde{z}^k \\ \tilde{\lambda}^k \end{pmatrix}, \quad \mathcal{J}(\mathbf{w}^k) &= \begin{pmatrix} -A^T \lambda^k \\ -B^T \lambda^k \\ -C^T \lambda^k \\ Ku^k - b \end{pmatrix}.
\end{align*}
\]

Since the basic results before Theorem 2.2 still hold, to show convergence of Algorithm 4.3, we start analyzing the following Theorem 4.5.

**Theorem 4.5.** The iterates generated by Algorithm 4.3 satisfy

\[
F(\mathbf{u}) - F(\tilde{\mathbf{u}}^k) + \langle \mathbf{w} - \tilde{\mathbf{w}}^k, \mathcal{J}(\mathbf{w}) \rangle \geq (\mathbf{w} - \tilde{\mathbf{w}}^k)^T Q_k (\mathbf{w}^k - \tilde{\mathbf{w}}^k) + \zeta^k
\]
where
\begin{align}
\lambda^k &= \lambda^k - \beta \left( A\lambda^{k+1} + B\lambda^k + C\lambda^k - b \right) \quad \text{and} \\
Q_k &= \begin{bmatrix}
D_k & L_1 + \beta B^T B & -\tau B^T \\
L_1 + \beta B^T B & L_2 + \beta C^T C & -\tau C^T \\
-\beta B & -\tau C & \frac{1}{\beta} I
\end{bmatrix}.
\end{align}

Proof. By the first-order optimality condition of $y$-subproblem, we have
\begin{align}
y^{k+1} &\in Y, \quad g(y) - g(y^{k+1}) + \langle y - y^{k+1}, p_y^k \rangle \geq 0, \quad \forall y \in Y,
\end{align}
with
\begin{align}
p_y^k &= -B^T \lambda^{k+\frac{1}{2}} + \beta B^T \left( A\lambda^{k+1} + B\lambda^k + C\lambda^k - b \right) + L_1 \left( y^{k+1} - y^k \right) \\
&= -B^T \lambda^{k+\frac{1}{2}} + \beta B^T \left( A\lambda^{k+1} + B\lambda^k + C\lambda^k - b \right) + (L_1 + \beta B^T B) \left( y^{k+1} - y^k \right) \\
&= -B^T \lambda^k + \tau B^T (\lambda^{k} - \lambda^k) + (L_1 + \beta B^T B) \left( y^{k+1} - y^k \right),
\end{align}
where we use the relationship
\begin{align}
\lambda^{k+\frac{1}{2}} = \lambda^k - \tau (\lambda^k - \bar{\lambda}^k).
\end{align}

Similarly, we can get
\begin{align}
z^{k+1} &\in Z, \quad h(z) - h(z^{k+1}) + \langle z - z^{k+1}, p_z^k \rangle \geq 0, \quad \forall z \in Z,
\end{align}
where
\begin{align}
p_z^k &= -C^T \lambda^{k+\frac{1}{2}} + \beta C^T \left( A\lambda^{k+1} + B\lambda^k + C\lambda^k - b \right) + L_2 \left( z^{k+1} - z^k \right) \\
&= -C^T \lambda^{k+\frac{1}{2}} + \beta C^T \left( A\lambda^{k+1} + B\lambda^k + C\lambda^k - b \right) + (L_2 + \beta C^T C) \left( z^{k+1} - z^k \right).
\end{align}

Besides, by the update of $\bar{\lambda}^k$ in (4.11) we have
\begin{align}
\lambda^k - \bar{\lambda}^k &= \beta \left( K\bar{\lambda}^k - b \right) + \beta B \left( y^k - \bar{y}^k \right) + \beta C \left( z^k - \bar{z}^k \right),
\end{align}
which implies
\begin{align}
\left\langle \lambda - \bar{\lambda}^k, K\bar{\lambda}^k - b - B (\bar{y}^k - y^k) - C (\bar{z}^k - z^k) + \frac{1}{\beta} (\bar{\lambda}^k - \lambda^k) \right\rangle = 0.
\end{align}

Combining the inequalities (2.8), (4.13), (4.15), (4.17) and the matrix $Q$ in (4.12), the proof is completed. \(\Box\)

Analogous to convergence analysis in Section 3, we just need to prove that Algorithm 4.3 is convergent for any $(\tau, s) \in \Delta$.

**Lemma 4.6.** Let $L_1 = \gamma_1 \sigma_1 I - \beta B^T B$ with $\sigma_1 \geq \beta B^T B$, and $L_2 = \gamma_2 \sigma_2 I - \beta C^T C$ with $\sigma_2 \geq \beta C^T C$. Then, for any $(\tau, s) \in \Delta$, the iterates generated by Algorithm 4.3
By adding (4.22) with \( y \) (4.23)

\[
2(Ku^{k+1} - b)^T [B(y^k - y^{k+1}) + C(z^k - z^{k+1})] \\
\geq 2 \left( \frac{1}{1 + s} \right) (Ku^k - b)^T [B(y^k - y^{k+1}) + C(z^k - z^{k+1})] \\
+ \frac{1}{1 + \tau} \left[ \gamma - 4 - 2\tau \right] \|B(y^k - y^{k+1})\|^2 - \gamma \|B(y^{k-1} - y^k)\|^2 \\
+ \frac{1}{1 + \tau} \left[ \gamma - 4 - 2\tau \right] \|C(z^k - z^{k+1})\|^2 - \gamma \|C(z^{k-1} - z^k)\|^2 \\
(4.18) \\
+ \frac{1}{1 + \tau} \left[ \|B(y^k - y^{k+1})\|_{H}^2 - \|B(y^k - y^{k-1})\|_{H}^2 \right]. \\
\]

where

\[
\begin{bmatrix}
I & -I \\
-I & I
\end{bmatrix}
\]

Proof. First of all, similar to proof of (3.10) we have

\[
\frac{1}{\beta} (y^k - y^{k+1})^T L_1 [(y^k - y^{k+1}) - (y^{k-1} - y^k)] \\
\geq (\gamma - 1) \|B(y^k - y^{k+1})\|^2 - \frac{\gamma}{2} \left( \|B(y^k - y^{k+1})\|^2 + \|B(y^{k-1} - y^k)\|^2 \right) \\
+ (y^k - y^{k+1})^T B (y^{k-1} - y^k) \\
= \frac{\gamma - 4}{2} \|B(y^k - y^{k+1})\|^2 - \frac{\gamma}{2} \|B(y^{k-1} - y^k)\|^2 \\
+ \|B(y^k - y^{k+1})\|^2 + (y^k - y^{k+1})^T B (y^{k-1} - y^k).
\]

Analogously, we have

\[
\frac{1}{\beta} (z^k - z^{k+1})^T L_2 [(z^k - z^{k+1}) - (z^{k-1} - z^k)] \\
\geq \frac{\gamma - 4}{2} \|C(z^k - z^{k+1})\|^2 - \frac{\gamma}{2} \|C(z^{k-1} - z^k)\|^2 \\
+ \|C(z^k - z^{k+1})\|^2 + (z^k - z^{k+1})^T C (z^{k-1} - z^k).
\]

Secondly, noticing by (4.13) that

\[
g(y) - g(y^{k+1}) + \left( y - y^{k+1}, -B^T \lambda^{k+\frac{1}{2}} + \beta B^T (Ku^{k+1} - b) - \right. \\
\beta B^T C(z^{k+1} - z^k) + L_1(y^{k+1} - y^k) \right) \geq 0,
\]

whose \( k \)-th iteration reads

\[
g(y) - g(y^{k+1}) + \left( y - y^{k}, -B^T \lambda^{k+\frac{1}{2}} + \beta B^T (Ku^{k} - b) - \right. \\
\beta B^T C(z^{k} - z^{k-1}) + L_1(y^{k} - y^{k-1}) \right) \geq 0.
\]

By adding (4.22) with \( y = y^k \) to (4.23) with \( y = y^{k+1} \) and noticing \( \lambda^k = \lambda^{k-1} - s\beta(Kw^k - b) \), we get

\[
(Ku^{k+1} - b)^T B(y^k - y^{k+1})
\]
\[
\begin{align*}
\geq & \frac{1-s}{1+s} (Ku^k - b)^T B (y^k - y^{k+1}) - \frac{\tau}{1+\tau} \|B(y^k - y^{k+1})\|^2 \\
+ & \frac{1}{1+\tau} (y^k - y^{k+1})^T B^T C \left[ z^{k+1} - z^k - (z^k - z^{k-1}) \right] \\
+ & \frac{1}{\beta(1+\tau)} \left( y^k - y^{k+1} \right)^T L_1 \left[ y^k - y^{k+1} - (y^{k-1} - y^k) \right] \\
\geq & \frac{1-s}{1+s} (Kw^k - b)^T B (y^k - y^{k+1}) \\
+ & \frac{1}{2(1+\tau)} \left[ (\gamma - 4 + 2\tau) \|B(y^k - y^{k+1})\|^2 - \gamma \|B(y^{k-1} - y^k)\|^2 \right] \\
+ & \frac{1}{1+\tau} \left\{ (y^k - y^{k+1})^T B^T C \left[ z^{k-1} - z^k + z^{k+1} - z^k \right] \\
& + \|B(y^k - y^{k+1})\|^2 + (y^k - y^{k+1})^T B^T B (y^k - y^{k-1}) \right\},
\end{align*}
\]

where the last inequality uses the inequality (4.20). Analogously, we deduce by (4.15) that

\[
(4.25) \quad (Ku^{k+1} - b)^T C (z^k - z^{k+1})
\geq \frac{1-s}{1+s} (Ku^k - b)^T C (z^k - z^{k+1}) \\
+ \frac{1}{2(1+\tau)} \left[ (\gamma - 4 + 2\tau) \|C(z^k - z^{k+1})\|^2 - \gamma \|C(z^{k-1} - z^k)\|^2 \right] \\
+ \frac{1}{1+\tau} \left\{ (z^k - z^{k+1})^T C^T B \left[ y^{k-1} - y^k + y^{k+1} - y^k \right] \\
& + \|C(z^k - z^{k+1})\|^2 + (z^k - z^{k+1})^T C^T C (z^k - z^{k-1}) \right\}.
\]

Adding the above (4.24) and (4.25) together, the proof is completed by the fact

\[
(y^k - y^{k+1})^T B^T C \left[ z^{k-1} - z^k + z^{k+1} - z^k \right] \\
+ \|B(y^k - y^{k+1})\|^2 + (y^k - y^{k+1})^T B^T B (y^k - y^{k-1}) \\
+ (z^k - z^{k+1})^T C^T B \left[ y^{k-1} - y^k + y^{k+1} - y^k \right] \\
+ \|C(z^k - z^{k+1})\|^2 + (z^k - z^{k+1})^T C^T C (z^k - z^{k-1}) \\
= \left[ B(y^k - y^{k+1}) \right]^T H \left[ B(y^k - y^{k+1}) \right] - \left[ B(y^k - y^{k+1}) \right]^T H \left[ B(y^k - y^{k+1}) \right] \\
\geq \frac{1}{2} \left\{ \|B(y^k - y^{k+1})\|^2 - \|B(y^k - y^{k+1})\|^2 \right\}.
\]

\[\blacksquare\]

**Theorem 4.7.** Let the conditions of Lemma 4.6 hold. Then the iterates generated by Algorithm 4.3 satisfy

\[
(4.26) \quad F(u) - F(\tilde{u})^T + (w - \tilde{w})^T J(w) \\
\geq \frac{1}{2} \left\{ \|w - w^{k+1}\|^2_{Q_u} - \|w - w^k\|^2_{Q_u} + \|w^k - \tilde{w}\|^2_{\tilde{Q}_u} \right\} + \zeta.
\]

Moreover, for any \((\tau, s) \in \Delta\) and \(\gamma \geq \frac{2(\tau^2 - 2\tau + 3)}{1+\tau}\), there exist \(\omega_0 > 0\) and \(\omega_1(i =
}
1, . . . , 4) ≥ 0 as shown in (4.34) such that
\[
\frac{1}{\beta} \left\| \mathbf{w}^k - \tilde{\mathbf{w}}^k \right\|_D^2 \geq \omega_0 \left\| \mathbf{Ku}^{k+1} - \mathbf{b} \right\|^2 + \omega_1 \left[ \left\| \mathbf{B}(\mathbf{y}^k - \mathbf{y}^{k+1}) \right\|^2 + \left\| \mathbf{C}(\mathbf{z}^k - \mathbf{z}^{k+1}) \right\|^2 \right]
\]
\[
+ \frac{1}{\beta} \left\| \mathbf{x}^k - \mathbf{x}^{k+1} \right\|_D^2 + \omega_2 \left[ \left\| \mathbf{Ku}^{k+1} - \mathbf{b} \right\|^2 - \left\| \mathbf{Ku}^k - \mathbf{b} \right\|^2 \right]
\]
\[
+ \omega_3 \left[ \left\| \mathbf{B}(\mathbf{y}^k - \mathbf{y}^{k+1}) \right\|^2 - \left\| \mathbf{B}(\mathbf{y}^{k-1} - \mathbf{y}^k) \right\|^2 + \left\| \mathbf{C}(\mathbf{z}^k - \mathbf{z}^{k+1}) \right\|^2 - \left\| \mathbf{C}(\mathbf{z}^{k-1} - \mathbf{z}^k) \right\|^2 \right]
\]
\[
+ \omega_4 \left[ \frac{\left\| \mathbf{B}(\mathbf{y}^k - \mathbf{y}^{k+1}) \right\|^2}{\mathbf{C}(\mathbf{z}^k - \mathbf{z}^{k+1})} - \frac{\left\| \mathbf{B}(\mathbf{y}^{k-1} - \mathbf{y}^k) \right\|^2}{\mathbf{C}(\mathbf{z}^{k-1} - \mathbf{z}^k)} \right],
\]
where ζ^k, ̃Q_k, ̃G_k and H are defined in (2.10), (4.28), (4.29) and (4.19), respectively.

Proof. Followed by (4.14) and the update of λ^{k+1}, it holds
\[
\lambda^k - \lambda^{k+1} = (\tau + s) \left( \lambda^k - \tilde{\lambda}^k \right) - s\beta \mathbf{B} (\mathbf{y}^k - \tilde{\mathbf{y}}^k) - s\beta \mathbf{C} (\mathbf{z}^k - \tilde{\mathbf{z}}^k).
\]
Therefore, (4.26) holds with the following matrices
\[
(4.27) \quad P = \begin{bmatrix}
I & I \\
-\beta \mathbf{B} & -s\beta \mathbf{C} & (\tau + s)I
\end{bmatrix},
\]
\[
(4.28) \quad ̃Q_k = Q_k P^{-1}
\]
\[
= \begin{bmatrix}
\mathcal{D}_k & L_1 + (1 - \frac{\tau + s}{\tau})\beta \mathbf{B}^T \mathbf{B} & -\frac{\tau - s}{\tau + s} \beta \mathbf{B}^T \mathbf{C} \\
-\frac{s}{\tau + s} \beta \mathbf{C}^T \mathbf{B} & L_2 + (1 - \frac{\tau + s}{\tau})\beta \mathbf{C}^T \mathbf{C} & -\frac{\tau - s}{\tau + s} \beta \mathbf{C}^T \\
-\frac{\tau}{\tau + s} \beta \mathbf{C}^T & -\frac{s}{\tau + s} \beta \mathbf{C}^T & (\tau + s)I
\end{bmatrix},
\]
and
\[
(4.29) \quad ̃G_k = ̃Q_k - (I - P)\mathbf{C} (I - P)
\]
\[
= \begin{bmatrix}
\mathcal{D}_k & L_1 + (1 - s)\beta \mathbf{B}^T \mathbf{B} & -s\beta \mathbf{B}^T \mathbf{C} \\
-\frac{s}{\tau + s} \beta \mathbf{C}^T \mathbf{B} & L_2 + (1 - s)\beta \mathbf{C}^T \mathbf{C} & (s - 1)\beta \mathbf{C}^T \\
-\frac{s}{\tau + s} \beta \mathbf{C}^T & -\frac{s}{\tau + s} \beta \mathbf{C}^T & (s - 1)\beta \mathbf{C}^T
\end{bmatrix}.
\]
By the conditions of Lemma 4.6, we have
\[
(4.30) \quad \frac{1}{\beta} \left\| \mathbf{w}^k - \tilde{\mathbf{w}}^k \right\|_D^2 \geq \frac{1}{\beta} \left\| \mathbf{x}^k - \mathbf{x}^{k+1} \right\|_D^2
\]
\[
+ (\gamma - s) \left\| \mathbf{B}(\mathbf{y}^k - \mathbf{y}^{k+1}) \right\|^2 + (\gamma - s) \left\| \mathbf{C}(\mathbf{z}^k - \mathbf{z}^{k+1}) \right\|^2
\]
\[
- 2s \left( \mathbf{y}^k - \mathbf{y}^{k+1} \right)^T \mathbf{B}^T \mathbf{C} (\mathbf{z}^k - \mathbf{z}^{k+1}) + 2(s - 1) \frac{1}{\beta} \left( \mathbf{y}^k - \mathbf{y}^{k+1} \right)^T \mathbf{B}^T \left( \lambda^k - \tilde{\lambda}^k \right)
\]
\[
+ (2 - \tau - s) \frac{1}{\beta^2} \left\| \lambda^k - \tilde{\lambda}^k \right\|^2 + 2(s - 1) \frac{1}{\beta} \left( \mathbf{z}^k - \mathbf{z}^{k+1} \right)^T \mathbf{C}^T \left( \lambda^k - \tilde{\lambda}^k \right).
\]
Now, we simplify the last three terms involved in the right-hand side of (4.30). Using (4.16) we have

$$\begin{align*}
(4.31) & \quad R_{II} = \frac{1}{\beta} \| \beta (Ku^{k+1} - b) + \beta B(y^k - y^{k+1}) + \beta C(z^k - z^{k+1}) \|^2 \\
& = \|Ku^{k+1} - b\|^2 + \|B(y^k - y^{k+1})\|^2 + \|C(z^k - z^{k+1})\|^2 \\
& + 2(Ku^{k+1} - b)^T B(y^k - y^{k+1}) + 2(Ku^{k+1} - b)^T C(z^k - z^{k+1}) \\
& + 2(y^k - y^{k+1})^T B^T C(z^k - z^{k+1}),
\end{align*}$$

and

$$\begin{align*}
(4.32) & \quad R_I = \left[Ku^{k+1} - b + C(z^k - z^{k+1})\right]^T B(y^k - y^{k+1}) + \|B(y^k - y^{k+1})\|^2, \\
\end{align*}$$

and

$$\begin{align*}
(4.33) & \quad R_{III} = \left[Ku^{k+1} - b + B(y^k - y^{k+1})\right]^T C(z^k - z^{k+1}) + \|C(z^k - z^{k+1})\|^2.
\end{align*}$$

Combining the above inequalities (4.30)-(4.33), it holds

$$\begin{align*}
\frac{1}{\beta} \left\| w^k - \tilde{w}^k \right\|^2 G_k \\
& \geq \frac{1}{\beta} \left\| x^k - x^{k+1} \right\|^2 D_k + (\gamma - \tau) \left\| B(y^k - y^{k+1}) \right\|^2 + (\gamma - \tau) \left\| C(z^k - z^{k+1}) \right\|^2 \\
& + (2 - \tau - s) \left\| Ku^{k+1} - b \right\|^2 - \tau \left( \left\| B(y^k - y^{k+1}) \right\|^2 + \left\| C(z^k - z^{k+1}) \right\|^2 \right) \\
& + \frac{1}{1 + \tau} \left\{ - (1 - s)^2 \left\| Ku^k - b \right\|^2 - 2 (1 - \tau)^2 \left[ \left\| B(y^k - y^{k+1}) \right\|^2 + \left\| C(z^k - z^{k+1}) \right\|^2 \right] \\
& + (\gamma - 4 - 2 \tau) \left( \left\| B(y^k - y^{k+1}) \right\|^2 + \left\| C(z^k - z^{k+1}) \right\|^2 \right) \\
& + \left\langle B(y^k - y^{k+1}) \right\| H - \left\| C(z^k - z^{k+1}) \right\| H \\
& - \gamma \left[ \left\| B(y^k - y^{k+1}) \right\|^2 + \left\| C(z^k - z^{k+1}) \right\|^2 \right] \right\} \\
& = \omega_0 \left\| Ku^{k+1} - b \right\|^2 + \omega_1 \left[ \left\| B(y^k - y^{k+1}) \right\|^2 + \left\| C(z^k - z^{k+1}) \right\|^2 \right] \\
& + \frac{1}{\beta} \left\| x^k - x^{k+1} \right\|^2 D_k + \omega_2 \left[ \left\| Ku^{k+1} - b \right\|^2 - \left\| Ku^k - b \right\|^2 \right] \\
& + \omega_3 \left[ \left\| B(y^k - y^{k+1}) \right\|^2 - \left\| B(y^k - y^{k-1}) \right\|^2 + \left\| C(z^k - z^{k+1}) \right\|^2 - \left\| C(z^k - z^{k-1}) \right\|^2 \right] \\
& + \omega_4 \left( \left\| B(y^k - y^{k+1}) \right\|^2 \right) _H - \left\| B(y^k - y^{k-1}) \right\|^2 \right)_H.
\end{align*}$$
Here,

\begin{equation}
\begin{cases}
\omega_0 = 2 - \tau - s - \frac{(1-s)^2}{1+\tau}, & \omega_1 = \gamma - 2\tau - \frac{6(1-\tau)}{1+\tau}, \\
\omega_2 = \frac{(1-s)^2}{1+\tau}, & \omega_3 = \frac{3(1-s)^2}{1+\tau}, & \omega_4 = \frac{1-\tau}{1+\tau},
\end{cases}
\end{equation}

It is not difficult to verify that the above parameters \(\omega_i \geq 0\) for any \((\tau, s) \in \Delta\) and \(\gamma \geq \frac{2(\tau^2-2\tau+3)}{1+\tau}\). So the proof is completed. \(\square\)

For any integer \(T > 0\), let

\[
w_T := \frac{1}{T} \sum_{k=1}^{T} \tilde{w}^k \quad \text{and} \quad u_T := \frac{1}{T} \sum_{k=1}^{T} \tilde{u}^k.
\]

Then, by using the above Theorem 4.7 and the similar analysis of Theorem 3.5, there exist \(\tilde{\omega}_i = \beta \omega_i (i = 2, \ldots, 4)\) such that

\[
\mathbb{E} \left[ F(u_T) - F(u) + (w_T - w)^T J(w) \right] \\
\leq \frac{1}{2T} \left\{ \sigma^2 \sum_{k=1}^{T} \eta_k m_k + \frac{4}{m_1(m_1+1)\eta_1} \| x - \bar{x} \|_2^2 + \| w - w^1 \|_{Q_1}^2 \\
+ \tilde{\omega}_2 \| Ku^1 - b \|^2 + \tilde{\omega}_3 \left( \| B(y^1 - y^0) \|_2^2 + \| C(z^1 - z^0) \|_2^2 \right) + \tilde{\omega}_4 \left( \frac{B(y^1 - y^0)}{C(z^1 - z^0)} \right)^2 \right\}.
\]

So, by the choice of the parameters \((\eta_k, m_k)\) in Corollary 3.6, we can obtain

\[
\mathbb{E} \left[ F(u_T) - F(u) + (w_T - w)^T J(w) \right] \leq O \left( \frac{1}{T} \left( 1 + \sum_{k=1}^{T} \frac{1}{(1+k)^6} \right) \right).
\]

5. Numerical experiments. Consider the following graph-guided fused lasso model in machine learning:

\[
\min_x \frac{1}{N} \sum_{j=1}^{N} f_j(x) + \mu \| Ax \|_1,
\]

where \(f_j(x) = \log(1 + \exp(-b_j a_j^T x))\) denotes the logistic loss function on the feature-label pair \((a_j, b_j) \in \mathbb{R}^d \times \{-1, 1\}\), \(N\) is the data size, \(\mu > 0\) is a given regularization parameter, and \(A = [G; I]\) is a matrix encoding the feature sparsity pattern with \(G\) being the sparsity pattern of the graph that is obtained by sparse inverse covariance estimation [5]. Introducing an auxiliary variable \(y\) transforms the above problem into a special case of Problem (1.1):

\begin{equation}
\min_{x, y} F(x, y) := \frac{1}{N} \sum_{j=1}^{N} f_j(x) + \mu \| y \|_1 \\
\text{s.t.} \quad Ax - y = 0.
\end{equation}

It is easy to verify that Assumptions 2.1-2.2 hold. Since here \(B = -I\), thus we apply Algorithm 1.1 with \(L = 0\) (when \(B \neq I\), we could use the technique in Lemma 3.2 to reduce the \(y\)-subproblem to a proximal mapping) to solve (5.1) in which the closed-form solutions of the involved subproblems are

\[
\begin{aligned}
x_{t+1} & = \left[ \gamma_t H + M_k \right]^{-1} \left[ \gamma_t H \bar{x}_t + M_k x^k - d_t - h^k \right], \\
y_{k+1} & = \text{Shrink} \left( \frac{\mu}{\beta}, Ax^{k+1} - \frac{X^{k+1/2}}{\beta} \right).
\end{aligned}
\]
Here, Shrink(·) denotes the so-called soft shrinkage operator and can be coded by the MATLAB inner function “wthresh”.

The penalty parameter in las-ADMM uses the tuned value 1e-3. The other parameters and the random vector \( e_t \) in las-ADMM (i.e. Algorithm 1.1) are chosen the same as that in [2, Section 7.1]. For convenience, we use

\[
(5.3) \quad \text{Obj} \_\text{err} = \frac{|F(x, y) - F^*|}{\max\{F^*, 1\}} \quad \text{and} \quad \text{Equ} \_\text{err} = \|Ax - y\|,
\]

to denote the average relative objective error and the average constraint error. Then we use \( \text{Opt} \_\text{err} = \max(\text{Obj} \_\text{err}, \text{Equ} \_\text{err}) \) to measure the algorithmic performance. Here \((x, y)\) represents the ergodic mean of the iterations generated over the last 2/3 of the total CPU time budget, and \( F^* \) is the approximate optimal objective function value obtained by running Algorithm 1.1 for more than 10 minutes. All experiments are implemented in MATLAB R2018a (64-bit) with the same starting point \((x^0, y^0, \lambda^0) = (0, 0, 0)\) and performed on a PC with Windows 10 operating system, with an Intel i7-8700K CPU and 16GB RAM.

![Fig. 5.1. Comparison of Opt_err vs CPU time for Problem (5.1) on the a9a dataset](image)

We test numerical performance of our proposed algorithm las-ADMM with tuned stepsizes \((\tau, s) = (0.9, 1.09)\) as in [1, Section 5.2.2] and AS-ADMM [2] for solving Problem (5.1) on the dataset \textit{mnist} (including 11,791 samples and 784 features) that is downloaded from LIBSVM website. The regularization parameter \( \mu \) is fixed as \( 10^{-5} \). We make 20 successive runs of the compared algorithms under different CPU time budgets, and comparison results are shown in Figure 5.1. Here, we only compare las-ADMM with AS-ADMM since in [2] AS-ADMM was proved better than other state-of-the-art methods. From Figure 5.1, we can see that las-ADMM performs worse than AS-ADMM at the beginning iteration while runs faster than AS-ADMM after the last 2/3 of the total CPU time budget, which shows that the proposed algorithm is a feasible method and performs very promising.

6. Conclusion. In this article, we have studied the convergence complexity of an accelerated stochastic alternating direction method of multipliers, whose dual variable uses larger stepsizes, for solving the “smooth + nonsmooth” separable convex optimization problem. Our stochastic algorithm is proved convergent in expectation
with the worst-case $O(1/T)$ convergence rate, where $T$ represents the number of outer iterations. We also analyzed the convergence of two extended 3-block algorithms in either fully Gauss-Seidel update or partially Jacobi update, while still maintaining the same stepsize region shown in (1.3). A degradation case, that is accelerated stochastic augmented Lagrange method, with stepsize region $(0, 2)$ is also investigated in a concise way. Preliminary experiments show that the new stochastic algorithm is feasible and efficient for solving the so-called graph-guided fused lasso problem arising in machine learning.

REFERENCES