Computational Optimization and Applications (2019) 73:201–235
https://doi.org/10.1007/s10589-019-00072-2

Inexact alternating direction methods of multipliers for separable convex optimization

William W. Hager[1] · Hongchao Zhang[2]

Received: 21 December 2017 / Published online: 7 February 2019

Abstract
Inexact alternating direction multiplier methods (ADMMs) are developed for solving general separable convex optimization problems with a linear constraint and with an objective that is the sum of smooth and nonsmooth terms. The approach involves linearized subproblems, a back substitution step, and either gradient or accelerated gradient techniques. Global convergence is established. The methods are particularly useful when the ADMM subproblems do not have closed form solution or when the solution of the subproblems is expensive. Numerical experiments based on image reconstruction problems show the effectiveness of the proposed methods.

Keywords Separable convex optimization · Alternating direction method of multipliers · Multiple blocks · Inexact ADMM · Global convergence

Mathematics Subject Classification 90C06 · 90C25 · 65Y20

December 19, 2017. The authors gratefully acknowledge support by the National Science Foundation under Grants 1522629, 1522654, 1819002, and 1819161, and by the Office of Naval Research under Grants N00014-15-1-2048 and N00014-18-1-2100.

✉ William W. Hager
hager@ufl.edu
http://people.clas.ufl.edu/hager/

Hongchao Zhang
hozhang@math.lsu.edu
http://www.math.lsu.edu/∼hozhang

1 Department of Mathematics, University of Florida, PO Box 118105, Gainesville, FL 32611-8105, USA

2 Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA

1 Introduction

We consider a convex separable linearly constrained optimization problem

$$\min \ \Phi(\mathbf{x}) \ \text{ subject to } \mathbf{Ax} = \mathbf{b} \tag{1.1}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\mathbf{A}$ is N by n. By a separable convex problem, we mean that the objective function is a sum of m independent components, and the matrix is partitioned compatibly as in

$$\Phi(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}_i) + h_i(\mathbf{x}_i) \quad \text{and} \quad \mathbf{Ax} = \sum_{i=1}^{m} \mathbf{A}_i \mathbf{x}_i. \tag{1.2}$$

Here f_i is convex and continuously differentiable with a Lipschitz continuous gradient, h_i is a proper closed convex function (possibly nonsmooth), $\mathbf{A}_i$ is N by n_i with $\sum_{i=1}^{m} n_i = n$, and the columns of $\mathbf{A}_i$ are linearly independent for $i \geq 2$. Constraints of the form $\mathbf{x}_i \in \mathcal{X}_i$, where $\mathcal{X}_i$ is a closed convex set, can be incorporated in the optimization problem by setting $h_i(\mathbf{x}_i) = \infty$ when $\mathbf{x}_i \notin \mathcal{X}_i$. The problem (1.1), (1.2) has attracted extensive research due to its importance in areas such as image processing, statistical learning and compressed sensing. See the recent survey [3] and its references.

Let $\mathcal{L}$ be the Lagrangian given by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \Phi(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{Ax} - \mathbf{b} \rangle,$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier for the linear constraint and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. It is assumed that there exists a solution $\mathbf{x}^*$ to (1.1), (1.2) and an associated Lagrange multiplier $\boldsymbol{\lambda}^* \in \mathbb{R}^N$ such that $\mathcal{L}(\cdot, \boldsymbol{\lambda}^*)$ attains a minimum at $\mathbf{x}^*$, or equivalently, the following first-order optimality conditions hold: $\mathbf{Ax}^* = \mathbf{b}$ and for $i = 1, 2, \ldots, m$ and for all $\mathbf{u} \in \mathbb{R}^{n_i}$, we have

$$\langle \nabla f_i(\mathbf{x}_i^*) + \mathbf{A}_i^{\mathsf{T}} \boldsymbol{\lambda}^*, \mathbf{u} - \mathbf{x}_i^* \rangle + h_i(\mathbf{u}) \geq h_i(\mathbf{x}_i^*), \tag{1.3}$$

where ∇ denotes the gradient.

A popular strategy for solving (1.1), (1.2) is the alternating direction multiplier method (ADMM) [16,17] given by

$$\begin{cases} \mathbf{x}_i^{k+1} = \arg \min\limits_{\mathbf{x}_i \in \mathbb{R}^{n_i}} \ L(\mathbf{x}_1^{k+1}, \ldots, \mathbf{x}_{i-1}^{k+1}, \mathbf{x}_i, \mathbf{x}_{i+1}^{k}, \ldots, \mathbf{x}_m^{k}, \boldsymbol{\lambda}^k), \\ \qquad\quad i = 1, \ldots, m, \\ \boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho(\mathbf{Ax}^{k+1} - \mathbf{b}), \end{cases} \tag{1.4}$$

where L, the augmented Lagrangian, is defined by

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) + \frac{\rho}{2} \|\mathbf{Ax} - \mathbf{b}\|^2. \tag{1.5}$$

Here $\rho > 0$ is the penalty parameter. Early ADMMs only consider problem (1.1), (1.2) with $m = 2$ corresponding to a 2-block structure. In this case, the global convergence and complexity can be found in [12,26]. When $m \geq 3$, the ADMM strategy (1.4), which is a natural extension of the 2-block ADMM, is not necessarily convergent [5], although its practical efficiency has been observed in many recent applications [34,35].

Recently, much research has focused on modifications to ADMM to ensure convergence when $m \geq 3$. References include [4,6,7,11,18,22,24,25,30,31]. One approach [4,7,11,22] assumes $m-2$ of the functions in the objective are strongly convex and the penalty parameter is sufficiently small. Linear convergence results under additional conditions are obtained in [31]. Analysis of a randomly permuted ADMM which allows for nonseparated variables is given in [6]. Another approach, first developed in [24,25], involves a back substitution step to complement the ADMM forward substitution step. The algorithms developed in our paper utilize this back substitution step.

The dominant computation in an iteration of ADMM is the solution of the subproblems in (1.4). Whenever an efficient closed form solution for the subproblems does not exist, the efficiency of ADMM depends on our ability to solve these subproblems inexactly while maintaining global convergence. One line of research is to solve the subproblems to an accuracy based on an absolute summable error criterion [9,12,19]. In [29], the authors combine an adaptive error criterion with the absolute summable error criterion for 2-block ADMM with logarithmic-quadratic proximal regularization and further correction steps to modify the solutions generated from the ADMM subproblems. In [14,15], the authors develop a 2-block ADMM with a relative error stopping condition for the subproblems, motivated by [13], based on the total subgradient error. Another line of research is to add proximal terms to make the subproblems strongly convex [8,23] and relatively easy to solve. However, this approach often requires accurate solution of the proximal subproblems. When $m = 1$, ADMM reduces to the standard augmented Lagrangian method (ALM), for which practical relative error criteria for solving the subproblems have been developed and encouraging numerical results have been obtained [13,33]. In this paper, motivated by our recent work on variable stepsize Bregman operator splitting methods (BOSVS), by recent complexity results for gradient and accelerated methods for convex optimization, and by the adaptive relative error strategy used in ALM, we develop new *inexact* approach for solving the ADMM subproblems. To the best of our knowledge, these are the first ADMMs for solving the general separable convex optimization problem (1.1), (1.2) based on an adaptive accuracy condition that does not employ an absolute summable error criterion and that guarantees global convergence, even when $m \geq 3$. As an alternative to inexact solutions of the ADMM subproblem, one could try to further split the variables and create additional subproblems which may be exactly solvable. However, further splitting the variables often decreases the convergence speed; moreover, in the general nonlinear setting, further splittings may not be possible.

To guarantee global convergence, a block Gaussian backward substitution strategy is used to make corrections to the approximate subproblem solutions. In the special case $m = 2$, the method will reduce to a 2-block ADMM without back substitution. This idea of using block Gaussian back substitution was first proposed in [24,25]. The method in this earlier work requires the exact solution of the subproblems to obtain global convergence, while our new approach allows an inexact solution. More recently, a linearly convergent ADMM was developed in [27]. This algorithm linearizes the subproblems to achieve an inexact solution, and requires that the functions f_i and h_i in the objective function satisfy certain "local error bound" conditions. In addition, to ensure linear convergence, the stepsize α_k in (1.4) must be sufficiently small, which could significantly deteriorate the practical performance.

In this paper, we focus on problems where the minimization of the augmented Lagrangian over one or more of the primal variables $\mathbf{x}_i$ is nontrivial, and the accuracy of an inexact minimizer needs to be taken into account. On the other hand, when these minimizations are simple enough, it is practical to minimize the augmented Lagrangian over $\mathbf{x}$ to obtain the dual function, which may be nonsmooth. The optimization of the dual function can be approached through smoothing techniques as in [28], or through active set techniques as in [21].

Our paper is organized as follows. In the Sect. 3, we first generalize the BOSVS algorithm [10,20] to handle multiple blocks. The original BOSVS algorithm was tailored to the two block case, but used an adaptive stepsize when solving the subproblem, and consequently, it achieved much better overall efficiency when compared to the Bregman operator splitting (BOS) type algorithms based on a fixed smaller stepsize. In Sects. 4 and 5 , more adaptive stopping criteria for the subproblems are proposed. The adaptive criteria for bounding the accuracy in the ADMM subproblems are based on both the current and accumulated iteration change in the subproblem. These novel stopping criteria are motivated by the complexity analysis of gradient methods for convex optimization, and by the relative accuracy strategy often used in an inexact augmented Lagrangian method for nonlinear programming. The basic idea in the methods of Sects. 4 and 5 is to introduce an inner loop in order to solve the ADMM subproblems with an adaptive accuracy which increases as iterates approach a solution. The method in Sect. 4 basically applies the gradient method to solve the ADMM subproblems, while the method in Sect. 5 applies an optimal (accelerated) gradient descent method. The goal is to ensure that the accumulated steps in the subproblems are asymptotically nondecreasing, which leads to a global convergence result. In our numerical experiments, the method based on accelerated gradient descent had the best performance. Although our analysis is carried out with vector variables, these results could be extended to matrix variables which could have more potential applications.

1.1 Notation

The set of solution/multiplier pairs for (1.1) is denoted $\mathcal{W}^*$, while $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{W}^*$ is a generic solution/multiplier pair. For $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^n$, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathsf{T}}\mathbf{y}$ is the standard inner product, where the superscript $^{\mathsf{T}}$ denotes transpose. The Euclidean norm, denoted $\|\cdot\|$, is defined by $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ and $\|\mathbf{x}\|_{\mathbf{G}} = \sqrt{\mathbf{x}^{\mathsf{T}}\mathbf{G}\mathbf{x}}$ for a positive definite matrix $\mathbf{G}$. $\mathbb{R}^+$ denotes the set of nonnegative real numbers, while $\mathbb{R}^{++}$ denotes the set of positive real numbers. We let $\partial f(\mathbf{x})$ denote the subdifferential at $\mathbf{x}$, when it exists. For a differentiable function, $\nabla f(\mathbf{x})$ is the gradient of f at $\mathbf{x}$, a column vector. If $\mathbf{x}$ is a vector, then $\mathbf{x}_+$ denotes the subvector obtained by dropping the first block of variables from $\mathbf{x}$. Thus if $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x}_i \in \mathbb{R}^{n_i}$ for $i \in [1, m]$, then $\mathbf{x}_+ = (\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_m)^\mathsf{T}$.

2 Algorithm structure

Three related inexact ADMMs are developed called generalized, multistep, and accelerated BOSVS. They differ in the details of the formula for the new iterate $\mathbf{x}^{k+1}$, but the overall structure of the algorithms is the same. Both multistep and accelerated BOSVS typically represent a more exact ADMM iteration when compared to generalized BOSVS, while the accelerated BOSVS subiterations often converge more rapidly than those of multistep BOSVS. The common elements of these three algorithms appear in Algorithm 2.1.

Parameters: $\rho, \sigma, \delta_{\min}, \theta_1, \theta_2, \theta_3 \in \mathbb{R}^{++}$, $\delta_{\min} < \delta_{\max}$, $\alpha \in (0, 1)$, $\sigma < 1 < \tau \le \eta$.
Starting guess: $\mathbf{x}^1$ and $\boldsymbol{\lambda}^1$.
Initialize: $\mathbf{y}^1 = \mathbf{x}^1$, $k = 1$, $\delta_{\min,i} = \delta_{\min}$ and $\Gamma_i^0 = 0$, $1 \le i \le m$, $e^0 = \infty$
Step 1: For $i = 1, \ldots, m$
  Generate $\mathbf{x}_i^{k+1}$ and $\mathbf{z}_i^k$, estimate $r_i^k \approx \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2$.
End
Step 2: If $e^k = \theta_1 \|\mathbf{z}_+^k - \mathbf{y}_+^k\| + \theta_2 \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\| + \theta_3 \sqrt{\sum_{i=1}^m r_i^k}$ is sufficiently small, terminate.
Step 3: Set $\mathbf{y}_1^{k+1} = \mathbf{z}_1^k$, $\mathbf{y}_+^{k+1} = \mathbf{y}_+^k + \alpha \mathbf{M}^{-\mathsf{T}} \mathbf{H}(\mathbf{z}_+^k - \mathbf{y}_+^k)$,
$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \alpha\rho(\mathbf{A}\mathbf{z}^k - \mathbf{b})$, $k := k + 1$, and go to Step 1.

Algorithm 2.1 Our ADMM structure.

The algorithms generate three sequences $\mathbf{x}^k$, $\mathbf{y}^k$, and $\mathbf{z}^k$. In Step 1 of Algorithm 2.1 there may be more than one ADMM subiteration, as determined by an adaptive stopping criterion. The iterate $\mathbf{x}^{k+1}$ is the final iterate generated in the ADMM (forward substitution) subproblems, $\mathbf{y}^k$ is generated by the back substitution process in Step 3, and $\mathbf{z}^k$ is an average of the iterates in the ADMM subproblems of Step 1. In generalized BOSVS, $\mathbf{z}^k = \mathbf{x}^{k+1}$ since there is only one ADMM subiteration, while multistep and accelerated BOSVS typically perform more than one subiteration and $\mathbf{z}^k$ is obtained by a nontrivial averaging process. The matrix $\mathbf{M}$ in Step 3 is the $m-1$ by $m-1$ block lower triangular matrix defined by

$$\mathbf{M}_{ij} = \begin{cases} \mathbf{A}_{i+1}^\mathsf{T} \mathbf{A}_{j+1} & \text{if } 1 \le j \le i < m, \\ \mathbf{0} & \text{if } 1 \le i < j < m. \end{cases} \tag{2.1}$$

The matrix $\mathbf{H}$ is the $m-1$ by $m-1$ block diagonal matrix whose diagonal blocks match those of $\mathbf{M}$. The matrices $\mathbf{M}$ and $\mathbf{H}$ are invertible since the columns of $\mathbf{A}_i$ are linearly independent for $i \ge 2$, which implies that $\mathbf{A}_i^\mathsf{T} \mathbf{A}_i$ is invertible for $i \ge 2$.

3 Generalized BOSVS

Our first algorithm is a generalization of the BOSVS algorithm developed in [10,20] for a two-block optimization problem. Let $\Phi_i^k : \mathbb{R}^{n_i} \times \mathbb{R}^{n_i} \times \mathbb{R} \to \mathbb{R}$ be defined by

$$\Phi_i^k(\mathbf{u}, \mathbf{v}, \delta) = f_i(\mathbf{v}) + \langle \nabla f_i(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{\delta}{2}\|\mathbf{u} - \mathbf{v}\|^2 + h_i(\mathbf{u}) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{u} - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho\|^2,$$

where

$$\mathbf{b}_i^k = \mathbf{b} - \sum_{j<i} \mathbf{A}_j \mathbf{z}_j^k - \sum_{j>i} \mathbf{A}_j \mathbf{y}_j^k. \tag{3.1}$$

The function Φ_i^k corresponds to the part of the augmented Lagrangian associated with the i-th component of $\mathbf{x}$, but with the smooth term f_i linearized around $\mathbf{v}$ and with a proximal term added to the objective. Algorithm 3.1 which follows is the Step 1 inner loop of Algorithm 2.1 for generalized BOSVS. Throughout the paper, the generalized BOSVS algorithm refers to Algorithm 2.1 with the Step 1 inner loop given by Algorithm 3.1.

Although $\mathbf{y}^k$ does not appear explicitly in Algorithm 3.1, it is hidden inside the $\mathbf{b}_i^k$ term of Φ_i^k. The iterate $\mathbf{x}_i^{k+1}$ is obtained by minimizing the Φ_i^k function and checking the line search condition of Step 1b. In Step 1a, mid denotes median and the initial stepsize $\delta_{i,0}^k$ of Step 1a is a safeguarded version of the Barzilai–Borwein formula [2].

Inner loop of Step 1 for generalized BOSVS:

1a. For $k > 1$, $\delta_{i,0}^k = \text{mid}\left\{\delta_{\min,i},\ s^{BB},\ \delta_{\max}\right\}$ where mid denotes median and $s^{BB} = \langle \nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\mathbf{x}_i^{k-1}),\ \mathbf{x}_i^k - \mathbf{x}_i^{k-1}\rangle / \|\mathbf{x}_i^k - \mathbf{x}_i^{k-1}\|^2$.
For $k = 1$, $\delta_{i,0}^k$ can be any scalar in $[\delta_{\min}, \delta_{\max}]$.

1b. Set $\delta_i^k = \eta^j \delta_{i,0}^k$, where $j \geq 0$ is the smallest integer such that
$f_i(\mathbf{x}_i^{k+1}) \leq f_i(\mathbf{x}_i^k) + \langle \nabla f_i(\mathbf{x}_i^k), \mathbf{x}_i^{k+1} - \mathbf{x}_i^k\rangle + \frac{(1-\sigma)\delta_i^k}{2}\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2$,
where $\mathbf{x}_i^{k+1} = \mathbf{z}_i^k = \arg\min\{\Phi_i^k(\mathbf{u}, \mathbf{x}_i^k, \delta_i^k) : \mathbf{u} \in \mathbb{R}^{n_i}\}$.

1c. Set $r_i^k = (1/\delta_i^k)\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2$.

1d. If $k > 1$ and $\delta_i^k > \max\{\delta_i^{k-1}, \delta_{\min,i}\}$, then $\delta_{\min,i} := \tau\delta_{\min,i}$.

Algorithm 3.1 Inner loop in Step 1 of Algorithm 2.1 for the generalized BOSVS scheme.

Let ζ_i denote the Lipschitz constant for ∇f_i. By a Taylor expansion of f_i around $\mathbf{x}_i^k$, we see that the line search condition of Step 1b is satisfied whenever $(1 - \sigma)\delta_i^k \geq \zeta_i$, or equivalently, when

$$\delta_i^k \geq \zeta_i/(1 - \sigma). \tag{3.2}$$

Since $\eta > 1, \delta_i^k$ increases as j increases, and consequently, (3.2) holds for j sufficiently large. Hence, if $j > 0$ at the termination of the line search, we have $\delta_i^k \leq \eta\zeta_i/(1-\sigma)$. If the line search terminates for $j = 0$, then $\delta_i^k \leq \delta_{\max}$. In summary, we have

$$\delta_{\min} \leq \delta_i^k \leq \max\{\eta\zeta_i/(1-\sigma), \delta_{\max}\} \quad \text{for all } k. \tag{3.3}$$

Since $s^{BB} \leq \zeta_i$ in Step 1a, it follows that $\delta_{i,0}^k = \delta_{\min,i}$ whenever $\delta_{\min,i} \geq \zeta_i$. In Step 1d, $\delta_{\min,i}$ is increased by the factor τ whenever $\delta_i^k > \delta_i^{k-1}$. Hence, after a finite number of iterations where $\delta_i^k > \delta_i^{k-1}$, we have $\delta_{\min,i} \geq \zeta_i/(1-\sigma)$, which implies that the line search terminates at $j = 0$ with $\delta_i^k = \delta_{i,0}^k = \delta_{\min,i}$. We conclude that

$$\delta_i^k \leq \delta_i^{k-1} \quad \text{for } k \text{ sufficiently large}, \tag{3.4}$$

where δ_i^k denotes the final accepted value in Step 1b. Note that the inequality in (3.4) cannot be replaced by equality since the number of iterations where $\delta_i^k > \delta_i^{k-1}$ may not be enough to yield $\delta_{\min,i} \geq \zeta_i/(1-\sigma)$.

In the BOS algorithm, the line search is essentially eliminated by taking δ_i^k larger than the Lipschitz constant ζ_i. Taking δ_i^k large, however, causes $\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|$ to be small due to the proximal term in the objective $\Phi_i^k(\,\cdot\,, \mathbf{x}_i^k, \delta_i^k)$ associated with $\mathbf{x}_i^{k+1}$. These small steps lead to slower convergence than what is achieved with BOSVS where δ_i^k is adjusted by the line search criterion in order to achieve a small, but acceptable, choice for δ_i^k.

In our analysis of generalized BOSVS, we first observe that when $e^k = 0$, we have reached a solution of (1.1), (1.2).

Lemma 3.1 *If $e^k = 0$ in the generalized BOSVS algorithm, then $\mathbf{x}^{k+1} = \mathbf{x}^k = \mathbf{y}^k$ solves (1.1), (1.2) and $(\mathbf{x}^k, \boldsymbol{\lambda}^k) \in \mathcal{W}^*$.*

Proof Let $\mathbf{x}^*$ denote $\mathbf{x}^k$. If $e^k = 0$, then $r_i = 0$ for each i, and by Step 1c of generalized BOSVS, $\mathbf{x}^{k+1} = \mathbf{x}^k = \mathbf{x}^*$. For generalized BOSVS, we set $\mathbf{z}^k = \mathbf{x}^{k+1}$ in Step 1b. Since $\mathbf{x}^{k+1} = \mathbf{x}^k$, it follows that $\mathbf{z}^k = \mathbf{x}^*$. The identity $\mathbf{z}^k = \mathbf{x}^{k+1}$ also implies that $\mathbf{z}^{k-1} = \mathbf{x}^k = \mathbf{x}^*$. In Step 3 of Algorithm 2.1, $\mathbf{y}_1^k = \mathbf{z}_1^{k-1} = \mathbf{x}_1^*$. Since $e^k = 0$, Step 2 of Algorithm 2.1 implies that $\mathbf{y}_+^k = \mathbf{z}_+^k = \mathbf{x}_+^*$. Hence, $\mathbf{y}^k = \mathbf{x}^* = \mathbf{z}^k$. Since $e^k = 0$, it also follows from Step 2 that $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Consequently, we have

$$\mathbf{b}_i^k = \mathbf{b} - \sum_{j<i} \mathbf{A}_j \mathbf{z}_j^k - \sum_{j>i} \mathbf{A}_j \mathbf{y}_j^k = \mathbf{b} - \sum_{j<i} \mathbf{A}_j \mathbf{x}_j^* - \sum_{j>i} \mathbf{A}_j \mathbf{x}_j^* = \mathbf{A}_i \mathbf{x}_i^*. \tag{3.5}$$

By Step 1b, $\mathbf{x}_i^{k+1}$ is the minimizer of $\Phi_i^k(\,\cdot\,, \mathbf{x}_i^k, \delta_i^k)$. Since $\mathbf{x}_i^{k+1} = \mathbf{x}_i^k = \mathbf{x}_i^*$, it follows that $\mathbf{x}_i^*$ is the minimizer of $\Phi_i^k(\,\cdot\,, \mathbf{x}_i^*, \delta_i^k)$. After taking into account (3.5), the first-order optimality condition associated with the minimizer $\mathbf{x}_i^*$ of $\Phi_i^k(\,\cdot\,, \mathbf{x}_i^*, \delta_i^k)$ is exactly the same as (1.3), but with $\boldsymbol{\lambda}^*$ replaced $\boldsymbol{\lambda}^k$. Hence, $(\mathbf{x}^*, \boldsymbol{\lambda}^k) \in \mathcal{W}^*$. □

Two lemmas are needed for the convergence of the generalized BOSVS algorithm.

Lemma 3.2 *Given* $\mathbf{v} \in \mathbb{R}^{n_i}$ *and* $\delta > 0$, *suppose that* $\mathbf{u}$ *minimizes* $\Phi_i^k(\cdot, \mathbf{v}, \delta)$ *and*

$$f_i(\mathbf{v}) + \langle \nabla f_i(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{(1-\sigma)\delta}{2} \|\mathbf{u} - \mathbf{v}\|^2 \geq f_i(\mathbf{u}) \tag{3.6}$$

for some $\sigma \in [0, 1)$. *Then, for any* $\mathbf{w} \in \mathbb{R}^{n_i}$ *we have*

$$\begin{aligned} L_i^k(\mathbf{w}) - L_i^k(\mathbf{u}) \geq\ & \frac{\delta}{2}(\|\mathbf{w} - \mathbf{u}\|^2 - \|\mathbf{w} - \mathbf{v}\|^2) \\ & + \frac{\rho}{2}\|\mathbf{A}_i(\mathbf{w} - \mathbf{u})\|^2 + \frac{\sigma\delta}{2}\|\mathbf{u} - \mathbf{v}\|^2, \end{aligned} \tag{3.7}$$

where L_i^k *is given by*

$$L_i^k(\mathbf{w}) = f_i(\mathbf{w}) + h_i(\mathbf{w}) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{w} - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho\|^2. \tag{3.8}$$

Proof Adding $h_i(\mathbf{u}) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{u} - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho\|^2$ to each side of the inequality (3.6) and rearranging, we obtain

$$\Phi_i^k(\mathbf{u}, \mathbf{v}, \delta) - \frac{\sigma\delta}{2}\|\mathbf{u} - \mathbf{v}\|^2 \geq L_i^k(\mathbf{u}).$$

Adding $L_i^k(\mathbf{w})$ to each side of this inequality gives

$$L_i^k(\mathbf{w}) - L_i^k(\mathbf{u}) \geq L_i^k(\mathbf{w}) - \Phi_i^k(\mathbf{u}, \mathbf{v}, \delta) + \frac{\sigma\delta}{2}\|\mathbf{u} - \mathbf{v}\|^2. \tag{3.9}$$

Utilizing the convexity inequality $f_i(\mathbf{w}) - f_i(\mathbf{v}) \geq \langle \nabla f_i(\mathbf{v}), \mathbf{w} - \mathbf{v} \rangle$, we have

$$\begin{aligned} L_i^k(\mathbf{w}) - \Phi_i^k(\mathbf{u}, \mathbf{v}, \delta) \geq\ & \frac{\rho}{2}\left(\|\mathbf{A}_i\mathbf{w} - \mathbf{b}_i^k - \boldsymbol{\lambda}^k/\rho\|^2 - \|\mathbf{A}_i\mathbf{u} - \mathbf{b}_i^k - \boldsymbol{\lambda}^k/\rho\|^2\right) \\ & + \langle \nabla f_i(\mathbf{v}), \mathbf{w} - \mathbf{u} \rangle - \frac{\delta}{2}\|\mathbf{u} - \mathbf{v}\|^2 + h_i(\mathbf{w}) - h_i(\mathbf{u}). \end{aligned}$$

Expand the smooth terms involving $\mathbf{w}$ on the right side in a Taylor series around $\mathbf{u}$ to obtain

$$L_i^k(\mathbf{w}) - \Phi_i^k(\mathbf{u}, \mathbf{v}, \delta) \geq \left\langle \mathbf{g}_i^k,\ \mathbf{w} - \mathbf{u} \right\rangle + h_i(\mathbf{w}) - h_i(\mathbf{u}) + \frac{\rho}{2}\|\mathbf{A}_i(\mathbf{w} - \mathbf{u})\|^2 - \frac{\delta}{2}\|\mathbf{u} - \mathbf{v}\|^2,$$

where $\mathbf{g}_i^k = \nabla f_i(\mathbf{v}) + \rho\mathbf{A}_i^{\mathsf{T}}(\mathbf{A}_i\mathbf{u} - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho)$. Since Φ_i^k is the sum of smooth and a nonsmooth term, the first-order optimality condition for the minimizer $\mathbf{u}$ of Φ_i^k can be expressed

$$\left\langle \mathbf{g}_i^k + \delta(\mathbf{u} - \mathbf{v}),\ \mathbf{w} - \mathbf{u} \right\rangle + h_i(\mathbf{w}) \geq h_i(\mathbf{u})$$

for all $\mathbf{w} \in \mathbb{R}^{n_i}$, which implies that

$$\left\langle \mathbf{g}_i^k,\ \mathbf{w} - \mathbf{u} \right\rangle + h_i(\mathbf{w}) - h_i(\mathbf{u}) \ge -\delta \langle \mathbf{u} - \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle$$

for all $\mathbf{w} \in \mathbb{R}^{n_i}$. Utilizing this inequality, we have

$$L_i^k(\mathbf{w}) - \Phi_i^k(\mathbf{u}, \mathbf{v}, \delta) \ge -\delta \langle \mathbf{u} - \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle + \frac{\rho}{2} \|\mathbf{A}_i(\mathbf{w} - \mathbf{u})\|^2 - \frac{\delta}{2} \|\mathbf{u} - \mathbf{v}\|^2.$$

Insert this in (3.9). Since

$$2\langle \mathbf{u} - \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle + \|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{w} - \mathbf{v}\|^2 - \|\mathbf{w} - \mathbf{u}\|^2,$$

the proof is complete. □

We use Lemma 3.2 to establish a decay property that is key to the convergence analysis. Recall that the ADMM parameter α lies in $(0, 1)$.

Lemma 3.3 *Let* $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{W}^*$ *be any solution/multiplier pair for* (1.1), (1.2), *let* $\mathbf{x}^k$, $\mathbf{y}^k$, $\mathbf{z}^k$, *and* $\boldsymbol{\lambda}^k$ *be the iterates of the generalized BOSVS algorithm, and define*

$$E_k = \rho \|\mathbf{y}_+^k - \mathbf{x}_+^*\|_{\mathbf{P}}^2 + \frac{1}{\rho} \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*\|^2 + \alpha \sum_{i=1}^m \delta_i^k \|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2,$$

where $\mathbf{P} = \mathbf{M}\mathbf{H}^{-1}\mathbf{M}^T$. *Then for k large enough that the monotonicity condition* (3.4) *holds for all* $i \in [1, m]$, *we have*

$$E_k \ge E_{k+1} + c_1 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + c_2 \rho (\|\mathbf{y}_+^k - \mathbf{z}_+^k\|_{\mathbf{H}}^2 + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\|^2),$$

where $c_1 = \sigma \alpha \delta_{\min}$ *and* $c_2 = \alpha(1 - \alpha)$.

Proof By the inequality (3.7) of Lemma 3.2 with $\mathbf{v} = \mathbf{x}_i^k$, $\mathbf{w} = \mathbf{x}_i^*$, and $\mathbf{u} = \mathbf{z}_i^k$, we have

$$\begin{aligned} L_i^k(\mathbf{x}_i^*) - L_i^k(\mathbf{z}_i^k) - \frac{\rho}{2} \|\mathbf{A}_i \mathbf{z}_{e,i}^k\|^2 &\ge \frac{\delta_i^k}{2} (\|\mathbf{z}_{e,i}^k\|^2 - \|\mathbf{x}_{e,i}^k\|^2) + \frac{\sigma \delta_i^k}{2} \|\mathbf{z}_i^k - \mathbf{x}_i^k\|^2 \\ &= \frac{\delta_i^k}{2} (\|\mathbf{x}_{e,i}^{k+1}\|^2 - \|\mathbf{x}_{e,i}^k\|^2) + \frac{\sigma \delta_i^k}{2} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2 \end{aligned} \tag{3.10}$$

where $\mathbf{x}_e^k = \mathbf{x}^k - \mathbf{x}^*$, $\mathbf{z}_e^k = \mathbf{z}^k - \mathbf{x}^*$, and $\mathbf{z}^k = \mathbf{x}^{k+1}$ by Step 1b of generalized BOSVS. Since $\mathbf{x}^*$ minimizes $L(\,\cdot\,, \boldsymbol{\lambda}^*)$ and the augmented Lagrangian is the sum of smooth and a nonsmooth term, the first-order optimality condition implies that for each $i \in [1, m]$,

$$\left\langle \mathbf{g}_i^*,\ \mathbf{w} - \mathbf{x}_i^* \right\rangle + h_i(\mathbf{w}) - h_i(\mathbf{x}_i^*) \ge 0 \tag{3.11}$$

for all $\mathbf{w} \in \mathbb{R}^{n_i}$, where $\mathbf{g}_i^*$ is the gradient of the smooth part of the objective evaluated at $\mathbf{x}_i^*$:

$$\mathbf{g}_i^* = \nabla f_i(\mathbf{x}_i^*) + \rho \mathbf{A}_i^{\mathsf{T}} \left(\sum_{j=1}^{m} \mathbf{A}_j \mathbf{x}_j^* - \mathbf{b} + \boldsymbol{\lambda}^*/\rho \right).$$

We add $L_i^k(\mathbf{x}_i^*) - L_i^k(\mathbf{z}_i^k) - \rho\|\mathbf{A}_i \mathbf{z}_{e,i}^k\|^2/2$ to both sides of (3.11) and take $\mathbf{w} = \mathbf{z}_i^k$. After much cancellation, we obtain the relation

$$\begin{aligned} & L_i^k(\mathbf{x}_i^*) - L_i^k(\mathbf{z}_i^k) - \frac{\rho}{2}\|\mathbf{A}_i \mathbf{z}_{e,i}^k\|^2 \\ & \quad \le f_i(\mathbf{x}_i^*) - f_i(\mathbf{z}_i^k) + \nabla f_i(\mathbf{x}_i^*)^{\mathsf{T}} \mathbf{z}_{e,i}^k - \rho \left\langle \sum_{j \le i} \mathbf{A}_j \mathbf{z}_{e,j}^k + \sum_{j>i} \mathbf{A}_j \mathbf{y}_{e,j}^k + \boldsymbol{\lambda}_e^k/\rho,\ \mathbf{A}_i \mathbf{z}_{e,i}^k \right\rangle \\ & \quad \le -\rho \left\langle \sum_{j \le i} \mathbf{A}_j \mathbf{z}_{e,j}^k + \sum_{j>i} \mathbf{A}_j \mathbf{y}_{e,j}^k + \boldsymbol{\lambda}_e^k/\rho,\ \mathbf{A}_i \mathbf{z}_{e,i}^k \right\rangle, \end{aligned} \tag{3.12}$$

where $\mathbf{y}_e^k = \mathbf{y}^k - \mathbf{x}^*$, $\boldsymbol{\lambda}_e^k = \boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*$, and the last inequality is due to the convexity of f_i. We combine this upper bound with the lower bound (3.10) to obtain

$$\begin{aligned} & -\rho \left\langle \mathbf{A}_i \mathbf{z}_{e,i}^k, \sum_{j \le i} \mathbf{A}_j \mathbf{z}_{e,j}^k + \sum_{j>i} \mathbf{A}_j \mathbf{y}_{e,j}^k + \boldsymbol{\lambda}_e^k/\rho \right\rangle \\ & \quad \ge \frac{\delta_i^k}{2}(\|\mathbf{x}_{e,i}^{k+1}\|^2 - \|\mathbf{x}_{e,i}^k\|^2) + \frac{\sigma \delta_i^k}{2}\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2. \end{aligned} \tag{3.13}$$

Focusing on the left side of (3.13), observe that

$$\begin{aligned} \sum_{j \le i} \mathbf{A}_j \mathbf{z}_{e,j}^k + \sum_{j>i} \mathbf{A}_j \mathbf{y}_{e,j}^k &= \sum_{j=1}^{m} \mathbf{A}_j (\mathbf{z}_j^k - \mathbf{x}_j^*) + \sum_{j>i} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k) \\ &= \mathbf{A}\mathbf{z}^k - \mathbf{b} + \sum_{j>i} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k) \end{aligned} \tag{3.14}$$

since $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Let τ_i^k denote the right side of (3.13):

$$\tau_i^k = \frac{\delta_i^k}{2}(\|\mathbf{x}_{e,i}^{k+1}\|^2 - \|\mathbf{x}_{e,i}^k\|^2) + \frac{\sigma \delta_i^k}{2}\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2.$$

Using this notation and the simplification (3.14), (3.13) becomes

$$-\rho \left\langle \mathbf{A}_i \mathbf{z}_{e,i}^k, \mathbf{A}\mathbf{z}^k - \mathbf{b} + \boldsymbol{\lambda}_e^k/\rho + \sum_{j>i} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k) \right\rangle \ge \tau_i^k. \tag{3.15}$$

We will sum the inequality (3.15) over i between 1 and m. Since

$$\sum_{i=1}^{m} \mathbf{A}_i \mathbf{z}_{e,i}^k = \sum_{i=1}^{m} \mathbf{A}_i (\mathbf{z}_i^k - \mathbf{x}_i^*) = \mathbf{A}\mathbf{z}^k - \mathbf{b} := \mathbf{r}^k,$$

it follows that in (3.15),

$$\sum_{i=1}^{m} \left\langle \mathbf{A}_i \mathbf{z}_{e,i}^k, \mathbf{r}^k + \boldsymbol{\lambda}_e^k/\rho \right\rangle = \left\langle \mathbf{r}^k, \mathbf{r}^k + \boldsymbol{\lambda}_e^k/\rho \right\rangle. \tag{3.16}$$

Also, observe that

$$\sum_{j>i} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k) = \sum_{j=2}^{m} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k) - \sum_{j=2}^{i} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k),$$

with the convention that the sum from $j = 2$ to $j = 1$ is 0. Take the inner product of this identity with $\mathbf{A}_i \mathbf{z}_{e,i}^k$ and sum over i to obtain

$$\begin{aligned} &\sum_{i=1}^{m} \left\langle \mathbf{A}_i \mathbf{z}_{e,i}^k, \sum_{j>i} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k) \right\rangle \\ &\quad = \left\langle \mathbf{r}^k, \sum_{j=2}^{m} \mathbf{A}_j (\mathbf{y}_j^k - \mathbf{z}_j^k) \right\rangle - (\mathbf{z}_+^k - \mathbf{x}_+^*)^{\mathsf{T}} \mathbf{M} (\mathbf{y}_+^k - \mathbf{z}_+^k), \end{aligned} \tag{3.17}$$

where $\mathbf{M}$ is defined in (2.1). We sum (3.15) over i between 1 and m and utilize (3.16) and (3.17) to obtain

$$\begin{aligned} &(\mathbf{y}_+^k - \mathbf{x}_+^*)^{\mathsf{T}} \mathbf{M} \mathbf{w} - \frac{1}{\rho} \left(\langle \mathbf{r}^k, \boldsymbol{\lambda}_e^k \rangle + \sum_{i=1}^{m} \tau_i^k \right) \\ &\quad \geq \mathbf{w}^{\mathsf{T}} \mathbf{M} \mathbf{w} + \left\langle \mathbf{r}^k, \mathbf{r}^k + \sum_{j=2}^{m} \mathbf{A}_j \mathbf{w}_{j-1} \right\rangle, \end{aligned} \tag{3.18}$$

where $\mathbf{w} = \mathbf{y}_+^k - \mathbf{z}_+^k$.

Observe that

$$\begin{aligned} \mathbf{w}^{\mathsf{T}} \mathbf{M} \mathbf{w} &= \frac{1}{2} \mathbf{w}^{\mathsf{T}} (\mathbf{M} + \mathbf{M}^{\mathsf{T}}) \mathbf{w} = \frac{1}{2} \mathbf{w}^{\mathsf{T}} (\mathbf{M} + \mathbf{M}^{\mathsf{T}} - \mathbf{H}) \mathbf{w} + \frac{1}{2} \mathbf{w}^{\mathsf{T}} \mathbf{H} \mathbf{w} \\ &= \frac{1}{2} \left\| \sum_{i=2}^{m} \mathbf{A}_i \mathbf{w}_{i-1} \right\|^2 + \frac{1}{2} \mathbf{w}^{\mathsf{T}} \mathbf{H} \mathbf{w} \end{aligned}$$

since $(\mathbf{M}+\mathbf{M}^{\mathsf{T}}-\mathbf{H})_{ij} = \mathbf{A}_{i+1}^{\mathsf{T}}\mathbf{A}_{j+1}$ by the definition of $\mathbf{M}$ and $\mathbf{H}$. With this substitution, the right side of (3.18) becomes a sum of squares:

$$\mathbf{w}^{\mathsf{T}}\mathbf{M}\mathbf{w} + \left\langle \mathbf{r}^k, \mathbf{r}^k + \sum_{j=2}^{m} \mathbf{A}_j \mathbf{w}_{j-1} \right\rangle = \frac{1}{2}\left(\mathbf{w}^{\mathsf{T}}\mathbf{H}\mathbf{w} + \|\mathbf{r}^k\|^2 + \left\| \mathbf{r}^k + \sum_{i=2}^{m} \mathbf{A}_i \mathbf{w}_{i-1} \right\|^2 \right).$$

Hence, it follows from (3.18) that

$$(\mathbf{y}_+^k - \mathbf{x}_+^*)^{\mathsf{T}}\mathbf{M}\mathbf{w} - \frac{1}{\rho}\left(\langle \mathbf{r}^k, \boldsymbol{\lambda}_e^k \rangle + \sum_{i=1}^{m} \tau_i^k \right) \geq \frac{1}{2}\left(\|\mathbf{w}\|_{\mathbf{H}}^2 + \|\mathbf{r}^k\|^2 \right). \tag{3.19}$$

Let $\mathbf{P} = \mathbf{M}\mathbf{H}^{-1}\mathbf{M}^{\mathsf{T}}$ and recall that $\mathbf{w} = \mathbf{y}_+^k - \mathbf{z}_+^k$. By the definition of $\mathbf{y}^{k+1}$ and $\boldsymbol{\lambda}^{k+1}$ in Step 3 of Algorithm 2.1, we have

$$\begin{aligned}
&\|\mathbf{y}_+^k - \mathbf{x}_+^*\|_{\mathbf{P}}^2 - \|\mathbf{y}_+^{k+1} - \mathbf{x}_+^*\|_{\mathbf{P}}^2 + \frac{1}{\rho^2}(\|\boldsymbol{\lambda}_e^k\|^2 - \|\boldsymbol{\lambda}_e^{k+1}\|^2) \\
&\quad = \|\mathbf{y}_+^k - \mathbf{x}_+^*\|_{\mathbf{P}}^2 - \|(\mathbf{y}_+^k - \mathbf{x}_+^*) - \alpha \mathbf{M}^{-\mathsf{T}}\mathbf{H}\mathbf{w}\|_{\mathbf{P}}^2 + \frac{1}{\rho^2}(\|\boldsymbol{\lambda}_e^k\|^2 - \|\boldsymbol{\lambda}_e^k + \alpha\rho\mathbf{r}^k\|^2) \\
&\quad = 2\alpha(\mathbf{y}_+^k - \mathbf{x}_+^*)^{\mathsf{T}}\mathbf{M}\mathbf{w} - \alpha^2\|\mathbf{w}\|_{\mathbf{H}}^2 - \frac{2\alpha}{\rho}\langle \mathbf{r}^k, \boldsymbol{\lambda}_e^k \rangle - \alpha^2\|\mathbf{r}^k\|^2.
\end{aligned}$$

On the right side of this inequality, we utilize (3.19) multiplied by 2α to conclude that

$$\begin{aligned}
&\|\mathbf{y}_+^k - \mathbf{x}_+^*\|_{\mathbf{P}}^2 - \|\mathbf{y}_+^{k+1} - \mathbf{x}_+^*\|_{\mathbf{P}}^2 + \frac{1}{\rho^2}(\|\boldsymbol{\lambda}_e^k\|^2 - \|\boldsymbol{\lambda}_e^{k+1}\|^2) - \frac{2\alpha}{\rho}\sum_{i=1}^{m} \tau_i^k \\
&\quad \geq c_2(\|\mathbf{y}_+^k - \mathbf{z}_+^k\|_{\mathbf{H}}^2 + \|\mathbf{r}^k\|^2)
\end{aligned} \tag{3.20}$$

where $c_2 = \alpha(1-\alpha) > 0$ since $\alpha \in (0, 1)$. By the definition of τ_i^k and the assumption that k is large enough that (3.4) holds for all i, it follows that

$$\begin{aligned}
-\tau_i^k &= \frac{\delta_i^k}{2}(\|\mathbf{x}_{e,i}^k\|^2 - \|\mathbf{x}_{e,i}^{k+1}\|^2) - \frac{\sigma\delta_i^k}{2}\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2 \\
&\leq \frac{\delta_i^k}{2}\|\mathbf{x}_{e,i}^k\|^2 - \frac{\delta_i^{k+1}}{2}\|\mathbf{x}_{e,i}^{k+1}\|^2 - \frac{\sigma\delta_i^k}{2}\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2.
\end{aligned}$$

This bound for $-\tau_i^k$ along with the inequality (3.20) and the definition of E^k complete the proof. □

The following theorem establishes the global convergence of generalized BOSVS.

Theorem 3.4 *If $\mathbf{x}^k$, $\mathbf{y}^k$, and $\boldsymbol{\lambda}^k$ are iterates of the generalized BOSVS algorithm, then the $\mathbf{x}^k$ and $\mathbf{y}^k$ sequences converge to a common limit denoted $\mathbf{x}^*$ and the $\boldsymbol{\lambda}^k$ converge to a limit denoted $\boldsymbol{\lambda}^*$ where $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{W}^*$.*

Proof Let $\bar{k}$ be chosen large enough that (3.4) holds for all $k \geq \bar{k}$. Since $\mathbf{x}^{k+1} = \mathbf{z}^k$ in generalized BOSVS, it follows from Lemma 3.3 that for $j \geq \bar{k}$ and $p > 0$, we have

$$E_j \geq E_{j+p} + c \sum_{k=j}^{j+p-1} (\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}_+^k - \mathbf{x}_+^{k+1}\|_{\mathbf{H}}^2 + \|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\|^2), \quad (3.21)$$

where $c = \min\{c_1, \rho c_2\} > 0$. Let p tend to $+\infty$ in (3.21). Since the columns of $\mathbf{A}_i$ are linearly independent for $i \geq 2$, $\mathbf{H}$ is positive definite and

$$\lim_{k\to\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| = \lim_{k\to\infty} \|\mathbf{y}_+^k - \mathbf{x}_+^{k+1}\| = \lim_{k\to\infty} \|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\| = 0. \quad (3.22)$$

By the definition of $\mathbf{b}_i^k$, we have

$$\mathbf{A}_i\mathbf{x}_i^{k+1} - \mathbf{b}_i^k = \sum_{j\leq i} \mathbf{A}_j\mathbf{x}_j^{k+1} + \sum_{j>i} \mathbf{A}_j\mathbf{y}_j^k - \mathbf{b}. \quad (3.23)$$

By (3.22), $\mathbf{y}_+^k$ approach $\mathbf{x}_+^{k+1}$, and by (3.23) and (3.22),

$$\lim_{k\to\infty} (\mathbf{A}_i\mathbf{x}_i^{k+1} - \mathbf{b}_i^k) = \lim_{k\to\infty} \mathbf{A}\mathbf{x}^{k+1} - \mathbf{b} = \mathbf{0} \quad (3.24)$$

for all $i \in [1, m]$.

By the definition of E_k in Lemma 3.3, we see that the iterates $\boldsymbol{\lambda}^k$ and $\mathbf{x}^k$ are uniformly bounded. Hence, there exist limits $\boldsymbol{\lambda}^*$ and $\mathbf{x}^*$, and an infinite sequence $\mathcal{K} \subset \{1, 2, \ldots\}$ such that $\boldsymbol{\lambda}^k$ and $\mathbf{x}^k$ for $k \in \mathcal{K}$ converge to $\boldsymbol{\lambda}^*$ and $\mathbf{x}^*$ respectively. By the first relation in (3.22), $\mathbf{x}^{k+1}$ also converges to $\mathbf{x}^*$ for $k \in \mathcal{K}$. In Step 1b of generalized BOSVS, we have

$$\mathbf{x}_i^{k+1} = \arg\min\{\Phi_i^k(\mathbf{u}, \mathbf{x}_i^k, \delta_i^k) : \mathbf{u} \in \mathbb{R}^{n_i}\}.$$

The first-order optimality conditions for $\mathbf{x}_i^{k+1}$ are

$$\left\langle \mathbf{g}_i^k,\ \mathbf{u} - \mathbf{x}_i^{k+1} \right\rangle + h_i(\mathbf{u}) \geq h_i(\mathbf{x}_i^{k+1}) \quad (3.25)$$

for all $\mathbf{u} \in \mathbb{R}^{n_i}$, where $\mathbf{g}_i^k$ is the gradient of the smooth part of the objective evaluated at $\mathbf{x}_i^{k+1}$:

$$\mathbf{g}_i^k = \nabla f_i(\mathbf{x}_i^k) + \rho\mathbf{A}_i^{\mathsf{T}}(\mathbf{A}_i\mathbf{x}_i^{k+1} - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho) + \delta_i^k(\mathbf{x}_i^{k+1} - \mathbf{x}_i^k).$$

As $k \in \mathcal{K}$ tends to infinity, $\nabla f_i(\mathbf{x}_i^k)$ approaches $\nabla f_i(\mathbf{x}^*)$ since ∇f_i is Lipschitz continuous, $\mathbf{A}_i\mathbf{x}_i^{k+1} - \mathbf{b}_i^k$ approaches $\mathbf{0}$ by (3.24), and $\delta_i^k(\mathbf{x}_i^{k+1} - \mathbf{x}_i^k)$ approaches $\mathbf{0}$ by (3.22) and the uniform bounded (3.4) for δ_i^k. Consequently, we have

$$\lim_{k\in\mathcal{K}} \mathbf{g}_i^k = \nabla f_i(\mathbf{x}^*) + \mathbf{A}_i^{\mathsf{T}}\boldsymbol{\lambda}^*. \quad (3.26)$$

Let $k \in \mathcal{K}$ tend to $+\infty$ in (3.25). By (3.26) and the lower semicontinuity of h_i, we deduce that

$$\langle \nabla f_i(\mathbf{x}_i^*) + \mathbf{A}_i^\mathsf{T}\boldsymbol{\lambda}^*, \mathbf{u} - \mathbf{x}_i^* \rangle + h_i(\mathbf{u}) \geq h_i(\mathbf{x}_i^*)$$

for all $\mathbf{u} \in \mathbb{R}^{n_i}$. Therefore, $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ satisfy the first-order optimality condition (1.3). By the last relation in (3.22), it follows that $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ and $\mathbf{x}^*$ is feasible in (1.1). By the convexity of f_i and h_i, $\mathbf{x}^*$ is a solution of (1.1), (1.2) and $\boldsymbol{\lambda}^*$ is an associated multiplier for the linear constraint.

Since $\mathbf{x}^{k+1}$ converges to $\mathbf{x}^*$ for $k \in \mathcal{K}$, the second relation in (3.22) implies that $\mathbf{y}_+^k$ converges to $\mathbf{x}_+^*$ for $k \in \mathcal{K}$. In Lemma 3.3, we use the specific limits $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ associated with $k \in \mathcal{K}$. Hence, E_k tends to 0 for $k \in \mathcal{K}$. It follows from (3.21) that the entire E_k sequence tends to 0. By the definition of E_k, we deduce that the entire $(\mathbf{x}^k, \mathbf{y}_+^k, \boldsymbol{\lambda}^k)$ sequence converges $(\mathbf{x}^*, \mathbf{x}_+^*, \boldsymbol{\lambda}^*)$. Since $\mathbf{y}_1^k = \mathbf{x}_1^k$ for each k, where $\mathbf{x}_1^k$ converges to $\mathbf{x}_1^*$, we conclude that $\mathbf{y}^k$ converges to $\mathbf{x}^*$. This completes the proof. □

4 Multistep BOSVS

For the template given by Algorithm 2.1, we only need to assume that the columns of $\mathbf{A}_i$ are linearly independent for $i \geq 2$ since only these columns enter into the matrix $\mathbf{M}$ which is inverted in Step 3. For generalized BOSVS, this assumption was sufficient for convergence. On the other hand, for both multistep and accelerated BOSVS, strong convexity of the augmented Lagrangian with respect to each of the variables $\mathbf{x}_i$ is needed in the analysis. Since it has already been assumed that the columns of $\mathbf{A}_i$ are linearly independent for $i \geq 2$, we will simply strengthen this assumption to require, henceforth, that the columns of $\mathbf{A}_i$ are linearly independent for every i. This ensures strong convexity of the augmented Lagrangian L with respect to each of the variables $\mathbf{x}_i$.

The inner loop for the multistep BOSVS algorithm appears in Algorithm 4.1.

Inner loop of Step 1 for multistep BOSVS:
Initialize: $\mathbf{u}_i^0 = \mathbf{x}_i^k$
For $l = 1, 2, \ldots$

1a. Choose $\delta_0^l \in [\delta_{\min}, \delta_{\max}]$.

1b. Set $\delta^l = \eta^j \delta_0^l$, where $j \geq 0$ is the smallest integer such that
$$f_i(\mathbf{u}_i^{l-1}) + \langle \nabla f_i(\mathbf{u}_i^{l-1}), \mathbf{u}_i^l - \mathbf{u}_i^{l-1} \rangle + \frac{(1-\sigma)\delta^l}{2} \|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 \geq f_i(\mathbf{u}_i^l),$$
where $\mathbf{u}_i^l = \arg\min\{\Phi_i^k(\mathbf{u}, \mathbf{u}_i^{l-1}, \delta_i^k) : \mathbf{u} \in \mathbb{R}^{n_i}\}$.

1c. If $\gamma^l := \sum_{j=1}^l 1/\delta^j \geq \Gamma_i^{k-1}$ and $\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|/\sqrt{\gamma^l} \leq \psi(e^{k-1})$, break, where ψ denotes any real-valued function, continuous at zero, such that $\psi(0) = 0$ and $\psi(s) > 0$ for $s > 0$ (e.g. $\psi(t) = t$).

Next

1d. Set $\mathbf{z}_i^k = \left(\sum_{j=1}^l \mathbf{u}_i^j/\delta^j\right)/\gamma^l$, $r_i^k = (1/\gamma^l)\sum_{j=1}^l \|\mathbf{u}_i^j - \mathbf{u}_i^{j-1}\|^2$, $\mathbf{x}_i^{k+1} = \mathbf{u}_i^l$, and $\Gamma_i^k = \gamma^l$.

Algorithm 4.1 Inner loop in Step 1 of Algorithm 2.1 for the multistep BOSVS scheme.

In generalized BOSVS, the iteration is given by $\mathbf{x}_i^{k+1} = \arg\min\{\Phi_i^k(\mathbf{u}, \mathbf{x}_i^k, \delta_i^k) : \mathbf{u} \in \mathbb{R}^{n_i}\}$ where δ_i^k is determined by a line search process. In the multistep BOSVS algorithm, this single minimization is replaced by the recurrence

$$\mathbf{u}_i^l = \arg\min\{\Phi_i^k(\mathbf{u}, \mathbf{u}_i^{l-1}, \delta_i^k) : \mathbf{u} \in \mathbb{R}^{n_i}\},$$

where $\mathbf{u}_i^0 = \mathbf{x}_i^k$. By converting the single minimization into a recurrence, we hope to achieve a better minimizer of the augmented Lagrangian. In generalized BOSVS, the convergence relies on a careful choice of δ_i^k based on safeguarding techniques. In multistep BOSVS, these restrictions on δ_i^k are replaced in Step 1c by a condition related to the accuracy of the iterates.

Since $\eta > 1$, the line search in Step 1b of multistep BOSVS terminates in a finite number of iterations and the final δ^l has exactly the same bounds (3.3) as that of generalized BOSVS. Since δ^l is uniformly bounded, it follows that the condition $\Gamma_i^k \geq \Gamma_i^{k-1}$ of Step 1c is fulfilled for l sufficiently large. In the numerical experiments for multistep BOSVS in Sect. 6, δ_0^l is given by the safeguarded BB choice of generalized BOSVS. Similar to Lemma 3.1, when $e^k = 0$, we have reached a solution of (1.1), (1.2).

The following inequality is based on Lemma 3.2.

Lemma 4.1 *In multistep BOSVS, we have*

$$\nu_i \rho \|\mathbf{z}_i^k - \bar{\mathbf{x}}_i^k\|^2 + \frac{\sigma}{\Gamma_i^k} \sum_{l=1}^{l_i^k} \|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 \leq \frac{\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\|^2}{\Gamma_i^k}, \tag{4.1}$$

for each $i \in [1, m]$, where l_i^k is the terminating value of l at iteration k, $\nu_i > 0$ is the smallest eigenvalue of $\mathbf{A}_i^T \mathbf{A}_i$, and

$$\bar{\mathbf{x}}_i^k = \arg\min\{L_i^k(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^{n_i}\} \tag{4.2}$$

with L_i^k defined in (3.8).

Proof By Lemma 3.2, we have

$$\begin{aligned} L_i^k(\mathbf{w}) - L_i^k(\mathbf{u}_i^l) \geq \frac{\delta^l}{2}(\|\mathbf{w} - \mathbf{u}_i^l\|^2 - \|\mathbf{w} - \mathbf{u}_i^{l-1}\|^2) + \frac{\rho}{2}\|\mathbf{A}_i(\mathbf{w} - \mathbf{u}_i^l)\|^2 \\ + \frac{\sigma\delta^l}{2}\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 \end{aligned} \tag{4.3}$$

for any $\mathbf{w} \in \mathbb{R}^{n_i}$. We take $\mathbf{w} = \bar{\mathbf{x}}_i^k$. Since $L_i^k(\mathbf{u}_i^l) - L_i^k(\bar{\mathbf{x}}_i^k) \geq 0$, we have

$$\frac{\rho}{\delta^l}\|\mathbf{A}_i(\bar{\mathbf{x}}_i^k - \mathbf{u}_i^l)\|^2 + \sigma\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 \leq \|\bar{\mathbf{x}}_i^k - \mathbf{u}_i^{l-1}\|^2 - \|\bar{\mathbf{x}}_i^k - \mathbf{u}_i^l\|^2. \tag{4.4}$$

Summing this inequality for l between 1 and l_i^k gives

$$\rho \sum_{l=1}^{l_i^k} \frac{1}{\delta^l} \|\mathbf{A}_i(\bar{\mathbf{x}}_i^k - \mathbf{u}_i^l)\|^2 + \sigma \sum_{l=1}^{l_i^k} \|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 \le \|\bar{\mathbf{x}}_i^k - \mathbf{x}_i^k\|^2. \tag{4.5}$$

Since the quadratic $\|\mathbf{A}_i(\bar{\mathbf{x}}_i^k - \mathbf{u})\|^2$ is a convex function of $\mathbf{u}$, it follows from Jensen's inequality that

$$\sum_{l=1}^{l_i^k} \frac{1}{\delta^l} \|\mathbf{A}_i(\bar{\mathbf{x}}_i^k - \mathbf{u}_i^l)\|^2 \ge \Gamma_i^k \|\mathbf{A}_i(\bar{\mathbf{x}}_i^k - \mathbf{z}_i^k)\|^2 \ge \Gamma_i^k \nu_i \|\mathbf{z}_i^k - \bar{\mathbf{x}}_i^k\|^2,$$

where $\nu_i > 0$ is the smallest eigenvalue of $\mathbf{A}_i^\mathsf{T}\mathbf{A}_i$. Combine this with (4.5) to obtain (4.1). □

Remark 4.1 L_i^k is strongly convex since it is the sum of convex functions and a strongly convex quadratic $\langle \mathbf{A}_i\mathbf{u}, \mathbf{A}_i\mathbf{u}\rangle$; consequently, the minimizer $\bar{\mathbf{x}}_i^k$ exists. Due to the upper bound (3.3) for δ^l in multistep BOSVS, γ^l grows linearly in l. Hence, for the inner loop of multistep BOSVS, (4.1) implies that $\|\mathbf{z}_i^k - \bar{\mathbf{x}}_i^k\| = O(1/\sqrt{l_i^k})$. By (4.3), the objective values satisfy $L_i^k(\mathbf{z}_i^k) - L_i^k(\bar{\mathbf{x}}_i^k) = O(1/l_i^k)$; to see this, divide (4.3) by δ^l, sum over l between 1 and l_i^k, and apply Jensen's inequality twice, to the terms involving $L(\mathbf{u}^l)$ and to the terms involving $\mathbf{A}_i$. As a consequence of Lemma 4.1, the stopping conditions in Step 1c of multistep BOSVS are satisfied for a finite l.

Similar to generalized BOSVS, the key to the convergence of multistep BOSVS is a decay property for the iterates. The analogue of Lemma 3.3 for multistep BOSVS is the following result.

Lemma 4.2 *Let $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{W}^*$ be any solution/multiplier pair for* (1.1), (1.2)*, let $\mathbf{x}^k$, $\mathbf{y}^k$, $\mathbf{z}^k$, $\mathbf{u}_k^l$, and $\boldsymbol{\lambda}^k$ be the iterates of the multistep BOSVS algorithm, let l_i^k be the terminating value of l at iteration k, and define*

$$E_k = \rho\|\mathbf{y}_+^k - \mathbf{x}_+^*\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*\|^2 + \alpha \sum_{i=1}^m \frac{\|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2}{\Gamma_i^k},$$

where $\mathbf{P} = \mathbf{M}\mathbf{H}^{-1}\mathbf{M}^T$. Then for all k, we have

$$E_k \ge E_{k+1} + c_1 \sum_{i=1}^m \sum_{l=1}^{l_i^k} \frac{\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2}{\Gamma_i^k} + c_2\rho(\|\mathbf{y}_+^k - \mathbf{z}_+^k\|_{\mathbf{H}}^2 + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\|^2),$$

where $c_1 = \sigma\alpha$ and $c_2 = \alpha(1-\alpha)$.

Proof We put $\mathbf{w} = \mathbf{x}_i^*$ in (4.3) to obtain

$$\frac{L_i^k(\mathbf{x}_i^*) - F_i^k(\mathbf{u}_{i,k}^l)}{\delta^l} \geq \frac{1}{2}(\|\mathbf{x}_i^* - \mathbf{u}_{i,k}^l\|^2 - \|\mathbf{x}_i^* - \mathbf{u}_{i,k}^{l-1}\|^2) + \frac{\sigma}{2}\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2,$$

where $F_i^k(\mathbf{u}_{i,k}^l) = L_i^k(\mathbf{u}_{i,k}^l) + (\rho/2)\|\mathbf{A}_i(\mathbf{u}_{i,k}^l - \mathbf{x}_i^*)\|^2$. Summing this inequality over l yields

$$\sum_{l=1}^{l_i^k}\left(\frac{L_i^k(\mathbf{x}_i^*) - F_i^k(\mathbf{u}_{i,k}^l)}{\delta^l}\right)$$
$$\geq \frac{1}{2}(\|\mathbf{x}_i^* - \mathbf{u}_{i,k}^{l_i^k}\|^2 - \|\mathbf{x}_i^* - \mathbf{u}_{i,k}^0\|^2) + \frac{\sigma}{2}\sum_{l=1}^{l_i^k}\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2. \tag{4.6}$$

Since F_i^k is convex, it follows from Jensen's inequality and the definition of Γ_i^k and $\mathbf{z}_i^k$ in Step 1c of multistep BOSVS that

$$\frac{1}{\Gamma_i^k}\sum_{l=1}^{l_i^k}\frac{1}{\delta^l}F_i^k(\mathbf{u}_{i,k}^l) \geq F_i^k\left(\frac{1}{\Gamma_i^k}\sum_{l=1}^{l_i^k}\frac{1}{\delta^l}\mathbf{u}_{i,k}^l\right) = F_i^k(\mathbf{z}_i^k). \tag{4.7}$$

Substitute $\mathbf{x}_i^{k+1} = \mathbf{u}_{i,k}^{l_i^k}$ and $\mathbf{x}_i^k = \mathbf{u}_{i,k}^0$ in (4.6) and use (4.7) to obtain

$$L_i^k(\mathbf{x}_i^*) - F_i^k(\mathbf{z}_i^k) \geq \frac{1}{2\Gamma_i^k}(\|\mathbf{x}_{e,i}^{k+1}\|^2 - \|\mathbf{x}_{e,i}^k\|^2) + \frac{\sigma}{2\Gamma_i^k}\sum_{l=1}^{l_i^k}\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2, \tag{4.8}$$

where $\mathbf{x}_{e,i}^k = \mathbf{x}_i^k - \mathbf{x}_i^*$. By (3.12), we have the upper bound

$$L_i^k(\mathbf{x}_i^*) - F_i^k(\mathbf{z}_i^k) \leq -\rho\left\langle\sum_{j\leq i}\mathbf{A}_j\mathbf{z}_{e,j}^k + \sum_{j>i}\mathbf{A}_j\mathbf{y}_{e,j}^k + \boldsymbol{\lambda}_e^k/\rho,\ \mathbf{A}_i\mathbf{z}_{e,i}^k\right\rangle.$$

Combining lower and upper bounds gives

$$-\rho\left\langle\sum_{j\leq i}\mathbf{A}_j\mathbf{z}_{e,j}^k + \sum_{j>i}\mathbf{A}_j\mathbf{y}_{e,j}^k + \boldsymbol{\lambda}_e^k/\rho,\ \mathbf{A}_i\mathbf{z}_{e,i}^k\right\rangle$$
$$\geq \frac{1}{2\Gamma_i^k}(\|\mathbf{x}_{e,i}^{k+1}\|^2 - \|\mathbf{x}_{e,i}^k\|^2) + \frac{\sigma}{2\Gamma_i^k}\sum_{l=1}^{l_i^k}\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2, \tag{4.9}$$

which is the same as (3.13) but with the following exchanges:

$$\delta_i^k \longleftrightarrow 1/\Gamma_i^k \quad \text{and} \quad \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|^2 \longleftrightarrow \sum_{l=1}^{l_i^k} \|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2.$$

Except for these adjustments, the remainder of the proof is the same as the proof of Lemma 3.3, starting with equation (3.14). □

Using Lemma 4.2, we can now prove the convergence of multistep BOSVS. The analysis parallels that of Theorem 3.4. To facilitate the analysis, we recall the definition and some properties of the proximal mapping. For any closed convex extended real-valued function h,

$$\text{prox}_h(\mathbf{v}) = \arg\min \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{v} - \mathbf{u}\|^2 : \mathbf{u} \in \text{dom}(h) \right\}.$$

As shown in [32, p. 340], the proximal mapping is nonexpansive:

$$\|\text{prox}_h(\mathbf{v}_1) - \text{prox}_h(\mathbf{v}_2)\| \le \|\mathbf{v}_1 - \mathbf{v}_2\|.$$

Moreover, if g is a differentiable convex function and

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} g(\mathbf{u}) + h(\mathbf{u}), \tag{4.10}$$

then it follows from the first-order optimality conditions for $\mathbf{u}^*$ that

$$\mathbf{u}^* = \text{prox}_h(\mathbf{u}^* - \nabla g(\mathbf{u}^*)). \tag{4.11}$$

Conversely, if (4.11) holds, then so does (4.10). Hence, these relations are equivalent. These properties will be used in the convergence analysis of multistep BOSVS.

Theorem 4.3 *If multistep BOSVS performs an infinite number of iterations generating iterates $\mathbf{y}^k$, $\mathbf{z}^k$, and $\boldsymbol{\lambda}^k$, then the sequences $\mathbf{y}^k$ and $\mathbf{z}^k$ both approach a common limit $\mathbf{x}^*$ and $\boldsymbol{\lambda}^k$ approaches a limit $\boldsymbol{\lambda}^*$ where $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{W}^*$.*

Proof For any $p > 0$, we sum the decay property of Lemma 4.2 to obtain

$$E_j \ge E_{j+p} + c \sum_{k=j}^{j+p-1} \left(\|\mathbf{y}_+^k - \mathbf{z}_+^k\|_{\mathbf{H}}^2 + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\|^2 + \sum_{i=1}^{m} \sum_{l=1}^{l_i^k} \frac{\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2}{\Gamma_i^k} \right), \tag{4.12}$$

where $c = \min\{c_1, \rho c_2\} > 0$. Let p tend to $+\infty$. Since $\mathbf{H}$ is positive definite, and the Γ_i^k are monotone nondecreasing as a function of k, it follows from (4.12) that

$$\lim_{k\to\infty} \|\mathbf{y}_+^k - \mathbf{z}_+^k\| = 0 = \lim_{k\to\infty} \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\|. \tag{4.13}$$

Moreover, by the definition of E_k in Lemma 4.2, $\mathbf{y}_+^k$ and $\boldsymbol{\lambda}^k$ are bounded sequences, and by the first equation in (4.13), $\mathbf{z}_+^k$ is also a bounded sequence. The second equation in (4.13) is equivalent to

$$\lim_{k\to\infty} \left\| \mathbf{A}_1 \mathbf{z}_1^k - \left(\mathbf{b} - \sum_{i=2}^{m} \mathbf{A}_i \mathbf{z}_i^k \right) \right\| = 0.$$

Since $\mathbf{z}_+^k$ is bounded and the columns of $\mathbf{A}_1$ are linearly independent, $\mathbf{z}_1$ is bounded. Hence, both $\mathbf{z}^k$ and $\boldsymbol{\lambda}^k$ are bounded sequences, and there exist an infinite sequence $\mathcal{K} \subset \{1, 2, \ldots\}$ and limits $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ such that

$$\lim_{k\in\mathcal{K}} \mathbf{z}^k = \mathbf{x}^* \quad \text{and} \quad \lim_{k\in\mathcal{K}} \boldsymbol{\lambda}^k = \boldsymbol{\lambda}^*. \tag{4.14}$$

By the first equation in (4.13), we have

$$\lim_{k\in\mathcal{K}} \mathbf{y}_+^k = \mathbf{x}_+^*. \tag{4.15}$$

By the second equation in (4.13), $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Consequently, by (4.14) and (4.15),

$$\lim_{k\in\mathcal{K}} \left(\mathbf{A}_i \mathbf{z}_i^k - \mathbf{b}_i^k \right) = \lim_{k\in\mathcal{K}} \left(\sum_{j\le i} \mathbf{A}_j \mathbf{z}_j^k + \sum_{j>i} \mathbf{A}_j \mathbf{y}_j^k - \mathbf{b} \right) = \mathbf{A}\mathbf{x}^* - \mathbf{b} = \mathbf{0} \tag{4.16}$$

for all $i \in [1, m]$.

The decay property (4.12) also implies that for each i,

$$\lim_{k\to\infty} r_i^k = \lim_{k\to\infty} \frac{1}{\Gamma_i^k} \sum_{l=1}^{l_i^k} \|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2 = 0. \tag{4.17}$$

Combine this with (4.13) to conclude that

$$\lim_{k\to\infty} e^k = \lim_{k\to\infty} \psi(e^k) = 0. \tag{4.18}$$

Next, we will show that

$$\mathbf{x}_i^* = \arg\min \left\{ f_i(\mathbf{u}) + h_i(\mathbf{u}) + \langle \boldsymbol{\lambda}^*, \mathbf{A}_i \mathbf{u} \rangle : \mathbf{u} \in \mathbb{R}^{n_i} \right\}. \tag{4.19}$$

If this were to hold for all $i = 1, \ldots, m$, then it would follow that

$$\mathbf{x}^* = \arg\min\{\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) : \mathbf{x} \in \mathbb{R}^n\}. \tag{4.20}$$

Since $\mathbf{A}\mathbf{x}^* = \mathbf{b}$, we conclude that $\mathbf{x}^*$ is an optimal solution of (1.1), (1.2) and $\boldsymbol{\lambda}^*$ is an associated multiplier. The remainder of the proof is partitioned into two cases depending on whether the monotone nondecreasing sequence Γ_i^k either approaches a finite limit, or tends to infinity.

Case 1 For some i, Γ_i^k approaches a finite limit. Due to the upper bound (3.3) for δ^l in Step 1b of multistep BOSVS, we conclude that l_i^k is uniformly bounded. By (4.17), $\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|$ approaches zero, where the convergence is uniform in k and $l \in [1, l_i^k]$. Since $\mathbf{u}_{i,k}^0 = \mathbf{x}_i^k$, the triangle inequality and the uniform upper bound for l_i^k imply that $\|\mathbf{x}_i^k - \mathbf{u}_{i,k}^l\|$ approaches zero, where the convergence is uniform in k and $l \in [1, l_i^k]$. Since $\mathbf{z}_i^k$ is a convex combination of $\mathbf{u}_{i,k}^l$ for $0 \le l \le l_i^k$ with l_i^k uniformly bounded and $\|\mathbf{x}_i^k - \mathbf{u}_{i,k}^l\|$ approaching zero, it follows that $\|\mathbf{z}_i^k - \mathbf{x}_i^k\|$ approaches zero. We summarize these observations in the relation

$$\lim_{k\to\infty} \|\mathbf{z}_i^k - \mathbf{x}_i^k\| = \lim_{k\to\infty} \|\mathbf{z}_i^k - \mathbf{u}_{i,k}^0\| = \lim_{k\to\infty} \|\mathbf{z}_i^k - \mathbf{u}_{i,k}^1\| = 0. \tag{4.21}$$

In multistep BOSVS, $\mathbf{u}_{i,k}^1$ minimizes $\Phi_i(\cdot, \mathbf{u}_i^0, \delta_i^k)$. Identify g in (4.10) with the smooth terms in Φ_i. By (4.11), we have

$$\mathbf{u}_{i,k}^1 = \text{prox}_{h_i}\left(\mathbf{u}_{i,k}^1 - \nabla f_i(\mathbf{u}_{i,k}^0) - \delta_i^k(\mathbf{u}_{i,k}^1 - \mathbf{u}_{i,k}^0) - \rho\mathbf{A}_i^\mathsf{T}(\mathbf{A}_i\mathbf{u}_{i,k}^1 - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho)\right).$$

Let us now take the limit as k tends to infinity with $k \in \mathcal{K}$. By (4.14), $\mathbf{z}_i^k$ approaches $\mathbf{x}_i^*$. By (4.21) both $\mathbf{u}_{i,k}^0$ and $\mathbf{u}_{i,k}^1$ approach $\mathbf{z}_i^k$, and by (4.16) $\mathbf{A}_i\mathbf{u}_{i,k}^1 - \mathbf{b}_i^k$ approaches zero. Since the prox function and ∇f_i are both Lipschitz continuous, we deduce that in the limit, as k tends to infinity with $k \in \mathcal{K}$,

$$\mathbf{x}_i^* = \text{prox}_{h_i}\left(\mathbf{x}_i^* - \nabla f_i(\mathbf{x}_i^*) - \mathbf{A}_i^\mathsf{T}\boldsymbol{\lambda}^*\right).$$

Again, by (4.10), (4.19) holds. And if this were to hold for all $i \in [1, m]$, it follows that $\mathbf{x}^*$ is an optimal solution of (1.1), (1.2), and $\boldsymbol{\lambda}^*$ is an associated multiplier. To show that (4.19) holds for all i, we need to also consider the situation where Γ_i^k tends to infinity.

Case 2 Suppose that Γ_i^k approaches infinity. Let $\bar{\mathbf{x}}_i^k$ be the minimizer of L_i^k defined in (3.8). Observe that minimizing $L_i^k(\mathbf{u})$ over $\mathbf{u} \in \mathbb{R}^{n_i}$ is equivalent to minimizing a sum of the form $g(\mathbf{u}) + h(\mathbf{u}) + \langle \mathbf{u}, \mathbf{c}^k \rangle$ where h corresponds to h_i, $\mathbf{c}^k = \mathbf{A}_i^\mathsf{T}(\boldsymbol{\lambda}^k - \rho\mathbf{b}_i^k)$, and $g(\mathbf{u}) = f_i(\mathbf{u}) + 0.5\rho\|\mathbf{A}_i\mathbf{u}\|^2$. Note that g is smooth and satisfies a strong convexity condition

$$(\mathbf{u} - \mathbf{v})^\mathsf{T}(\nabla g(\mathbf{u}) - \nabla g(\mathbf{v})) \ge \rho\nu_i\|\mathbf{u} - \mathbf{v}\|^2, \tag{4.22}$$

where $\nu_i > 0$ is the smallest eigenvalue of $\mathbf{A}_i^\mathsf{T}\mathbf{A}_i$. By the strong convexity of L_i^k, it has a unique minimizer, and from the first-order optimality conditions and the strong convexity condition (4.22), we obtain the bound

$$\|\bar{\mathbf{x}}_i^j - \bar{\mathbf{x}}_i^k\| \le \|\mathbf{c}^j - \mathbf{c}^k\|/(\rho\nu_i). \tag{4.23}$$

Since $\mathbf{z}^k$, $\mathbf{y}_+^k$, and $\boldsymbol{\lambda}^k$ are bounded sequences, it follows that $\bar{\mathbf{x}}_i^k$ is a bounded sequence. For $k \in \mathcal{K}$, the sequences $\mathbf{z}^k$, $\mathbf{y}_+^k$, and $\boldsymbol{\lambda}^k$ converge to $\mathbf{x}^*$, $\mathbf{x}_+^*$, and $\boldsymbol{\lambda}^*$ respectively, which implies that

$$\mathbf{c}^* = \lim_{k \in \mathcal{K}} \mathbf{c}^k = \mathbf{A}_i^\mathsf{T} \left[\boldsymbol{\lambda}^* - \rho \left(\mathbf{b} - \sum_{j \neq i} \mathbf{A}_j \mathbf{x}_j^* \right) \right] = \mathbf{A}_i^\mathsf{T} \left[\boldsymbol{\lambda}^* - \rho \mathbf{A}_i \mathbf{x}_i^* \right], \quad (4.24)$$

where the last equality is due to the identity $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Consequently, by (4.23), $\bar{\mathbf{x}}_i^k$ for $k \in \mathcal{K}$ forms a Cauchy sequence which approaches a limit. We use the stopping condition to determine the limit.

Let us insert $l = l_i^k$ and $\mathbf{u}_i^l = \mathbf{x}_i^{k+1}$ in the inequality (4.4). By the linear independence of the columns of $\mathbf{A}_i$ and the upper bound (3.3) for δ^l, there exists $\beta > 0$ such that

$$\begin{aligned} \beta \|\bar{\mathbf{x}}_i^k - \mathbf{x}_i^{k+1}\|^2 &\le \frac{\rho}{\delta^l} \|\mathbf{A}_i (\bar{\mathbf{x}}_i^k - \mathbf{u}_{i,k}^l)\|^2 \le \|\bar{\mathbf{x}}_i^k - \mathbf{u}_{i,k}^{l-1}\|^2 - \|\bar{\mathbf{x}}_i^k - \mathbf{u}_{i,k}^l\|^2 \\ &= 2\langle \bar{\mathbf{x}}_i^k - \mathbf{x}_i^{k+1}, \mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1} \rangle + \|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2 \\ &\le 2\|\bar{\mathbf{x}}_i^k - \mathbf{x}_i^{k+1}\| \|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\| + \|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2. \end{aligned}$$

We complete the square on the right side to obtain the relation

$$\|\bar{\mathbf{x}}_i^k - \mathbf{x}_i^{k+1}\| \le \frac{\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|}{\beta} \left(1 + \sqrt{\beta + 1} \right).$$

Square this inequality and divide by Γ_i^k to get

$$\frac{\|\bar{\mathbf{x}}_i^k - \mathbf{x}_i^{k+1}\|^2}{\Gamma_i^k} \le \frac{\|\mathbf{u}_{i,k}^l - \mathbf{u}_{i,k}^{l-1}\|^2}{\beta^2 \Gamma_i^k} \left(1 + \sqrt{\beta + 1} \right)^2.$$

Since $l = l_i^k$, it follows from the stopping condition of Step 1c and from (4.18) that the right of this inequality approaches zero as k tends to infinity. Earlier we showed that $\bar{\mathbf{x}}_i^k$ is a bounded sequence. Since Γ_i^k tends to infinity in Case 2, and $\bar{\mathbf{x}}_i^k / \sqrt{\Gamma_i^k}$ approaches zero, we conclude that $\mathbf{x}_i^{k+1} / \sqrt{\Gamma_i^k}$ approaches zero. Due to the inequality $\Gamma_i^{k+1} \ge \Gamma_i^k$, $\mathbf{x}_i^{k+1} / \sqrt{\Gamma_i^{k+1}}$ also approaches zero as k tends to infinity. Since $\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\| \le \|\mathbf{x}_i^k\| + \|\bar{\mathbf{x}}_i^k\|$, the right side of (4.1) approaches zero. Hence, (4.1) implies that $\mathbf{z}_i^k$ approaches $\bar{\mathbf{x}}_i^k$ as k tends to infinity. And since $\mathbf{z}_i^k$ also approaches $\mathbf{x}_i^*$ for $k \in \mathcal{K}$, we conclude that $\bar{\mathbf{x}}_i^k$ approaches $\mathbf{x}_i^*$ as $k \in \mathcal{K}$ tends to infinity. Let $\bar{\mathbf{x}}_i^*$ be defined by

$$\bar{\mathbf{x}}_i^* = \arg \min_{\mathbf{u}} \{ g(\mathbf{u}) + h(\mathbf{u}) + \langle \mathbf{u}, \mathbf{c}^* \rangle \}.$$

By (4.23) and the fact that $\bar{\mathbf{x}}_i^k$ approaches $\mathbf{x}_i^*$ as $k \in \mathcal{K}$ tends to infinity, we conclude that $\bar{\mathbf{x}}_i^* = \mathbf{x}_i^*$. In summary, we have

$$\lim_{k \in \mathcal{K}} \bar{\mathbf{x}}_i^k = \mathbf{x}_i^* = \bar{\mathbf{x}}_i^* = \arg\min_{\mathbf{u}} \{g(\mathbf{u}) + h(\mathbf{u}) + \langle \mathbf{u}, \mathbf{c}^* \rangle\}.$$
$$= \arg\min_{\mathbf{u}} \{f_i(\mathbf{u}) + 0.5\rho\|\mathbf{A}_i\mathbf{u}\|^2 + h_i(\mathbf{u}) + \langle \boldsymbol{\lambda}^* - \rho\mathbf{A}_i\mathbf{x}_i^*, \mathbf{u}\rangle\}. \quad (4.25)$$

The first-order optimality conditions for (4.25) are exactly the same as the first-order optimality conditions for (4.19). This shows that (4.19) holds in either Case 1 or Case 2. Hence, (4.20) holds and $\mathbf{x}^*$ is an optimal solution of (1.1), (1.2) with associated multiplier $\boldsymbol{\lambda}^*$.

Finally, we need to show that the entire sequence converges. If Γ_i^k is uniformly bounded as in Case 1, then by (4.21), $\mathbf{x}_i^k$ approaches $\mathbf{x}_i^*$ and $\|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2/\Gamma_i^k$ approaches zero as k tends to infinity with $k \in \mathcal{K}$. On the other hand, when Γ_i^k tends to infinity as in Case 2, we showed that $\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\|^2/\Gamma_i^k$ approaches zero. Since $\bar{\mathbf{x}}_i^k$ for $k \in \mathcal{K}$ approaches $\mathbf{x}_i^*$ by (4.25) and Γ_i^k tends to infinity, it follows that $\|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2/\Gamma_i^k$ approaches zero for $k \in \mathcal{K}$. Thus in either Case 1 or Case 2, $\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\|^2/\Gamma_i^k$ approaches zero as k tends to infinity with $k \in \mathcal{K}$. Letting j tend to infinity in (4.12) with $j \in \mathcal{K}$, it follows that E_j approaches zero. Moreover, (4.12) implies that along the entire sequence, $\mathbf{y}_+^k$ approaches $\mathbf{x}_+^*$ and $\boldsymbol{\lambda}^k$ approaches $\boldsymbol{\lambda}^*$. By (4.13), the entire sequence of iterates $\mathbf{z}_+^k$ approaches $\mathbf{x}_+^*$. Since $\mathbf{Az}^k$ approaches $\mathbf{b}$ (see (4.13)), $\mathbf{Ax}^* = \mathbf{b}$, and $\mathbf{A}_1^\mathsf{T}\mathbf{A}_1$ is invertible, the entire sequence $\mathbf{z}_1^k$ approaches $\mathbf{x}_1^*$. Finally, since $\mathbf{y}_1^{k+1} = \mathbf{z}_1^k$, we deduce that the entire $\mathbf{y}^k$ sequence approaches $\mathbf{x}^*$. This completes the proof. □

5 Accelerated BOSVS

The inner loop for the accelerated BOSVS algorithm appears in Algorithm 5.1. As we will see, the inner loop (Step 1) of accelerated BOSVS converges to the minimizer of L_i^k, exactly as in multistep BOSVS; however, the convergence speed of the multistep BOSVS inner loop is $O(1/\sqrt{l})$ for the $\mathbf{z}_i^k$ iterates and $O(1/l)$ for the objective (see Remark 4.1), while the convergence speed in accelerated BOSVS is $O(1/l)$ for the $\mathbf{z}_i^k$ iterates and $O(1/l^2)$ for the objective, which is optimal for first-order methods applied to general convex, possibly nonsmooth optimization problems.

Two parameter sequences appear in the accelerated BOSVS scheme, the δ^l and α^l sequences. They must be chosen so that the line search condition of Step 1a is satisfied for each value of l, and the stopping condition of Step 1b is satisfied for l sufficiently large. If the Lipschitz constant ζ_i of f_i is known, then we could take

$$\delta^l = \frac{1}{(1-\sigma)} \frac{2\zeta_i}{l} \quad \text{and} \quad \alpha^l = \frac{2}{l+1} \in (0, 1], \quad (5.1)$$

Inner loop of Step 1 for accelerated BOSVS:
Initialize: $\mathbf{a}_i^0 = \mathbf{u}_i^0 = \mathbf{x}_i^k$ and $\alpha^1 = 1$.
For $l = 1, 2, \ldots$

1a. Choose $\delta^l \geq \delta_{\min}$ and when $l > 1$, choose $\alpha^l \in (0, 1)$ such that
$$f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^l - \bar{\mathbf{a}}_i^l \rangle + \frac{(1-\sigma)\delta^l}{2\alpha^l} \|\mathbf{a}_i^l - \bar{\mathbf{a}}_i^l\|^2 \geq f_i(\mathbf{a}_i^l),$$
where $\mathbf{a}_i^l = (1-\alpha^l)\mathbf{a}_i^{l-1} + \alpha^l \mathbf{u}_i^l$, $\bar{\mathbf{a}}_i^l = (1-\alpha^l)\mathbf{a}_i^{l-1} + \alpha^l \mathbf{u}_i^{l-1}$, and
$\mathbf{u}_i^l = \arg\min\{Q(\mathbf{u}) + h_i(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^{n_i}\}$ with
$$Q(\mathbf{u}) = \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u} \rangle + \frac{\delta^l}{2} \|\mathbf{u} - \mathbf{u}_i^{l-1}\|^2 + \frac{\rho}{2} \|\mathbf{A}_i \mathbf{u} - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho\|^2.$$

1b. If $\gamma^l = (1/\delta^1) \prod_{j=2}^{l} (1-\alpha^j)^{-1} \geq \Gamma_i^{k-1}$, where $\gamma^1 = 1/\delta^1$,
and $\|\mathbf{a}_i^l - \mathbf{a}_i^{l-1}\| \leq \psi(e^{k-1})$, then break.

Next
1c. Set $\mathbf{x}_i^{k+1} = \mathbf{u}_i^l$, $\mathbf{z}_i^k = \mathbf{a}_i^l$, $\Gamma_i^k = \gamma^l$, and $r_i^k = (1/\Gamma_i^k) \sum_{j=1}^{l} \|\mathbf{u}_i^j - \mathbf{u}_i^{j-1}\|^2$.

Algorithm 5.1 Inner loop in Step 1 of Algorithm 2.1 for the accelerated BOSVS scheme.

in which case, we have

$$\frac{(1-\sigma)\delta^l}{\alpha^l} = \frac{(l+1)\zeta_i}{l} > \zeta_i.$$

This relation along with a Taylor series expansion of f_i around $\mathbf{u}_i^{l-1}$ implies that the line search condition in Step 1a of accelerated BOSVS is satisfied for each l. Moreover, we show (after Lemma 5.1) that with these choices for δ^l and α^l, the stopping condition of Step 1b is also satisfied eventually.

A different, adaptive way to choose the parameters, that does not require knowledge of the Lipschitz constant for f_i, is the following: Choose $\delta_0^l \in [\delta_{\min}, \delta_{\max}]$, where $0 < \delta_{\min} < \delta_{\max} < \infty$ are safeguard parameters, and set

$$\delta^l = \frac{2}{\theta^l + \sqrt{(\theta^l)^2 + 4\theta^l \Lambda^{l-1}}} \quad \text{and} \quad \alpha^l = \frac{1}{1 + \delta^l \Lambda^{l-1}}, \quad \text{where}$$
$$\Lambda^l = \sum_{i=1}^{l} 1/\delta^i, \quad \Lambda^0 = 0, \quad \text{and} \quad \theta^l = 1/(\delta_0^l \eta^j) \text{ with } \eta > 1. \tag{5.2}$$

After some algebra, it can be shown that

$$\frac{\delta^l}{\alpha^l} = \frac{1}{\theta^l} = \delta_0^l \eta^j. \tag{5.3}$$

Hence, the ratio δ^l/α^l appearing in the line search condition of Step 1a tends to infinity as j tends to infinity since $\eta > 1$. We take $j \geq 0$ to be the smallest integer for which the line search condition is satisfied. Based on the identity (5.3), the expression δ^l/α^l has exactly the same effect as δ_i^k in generalized BOSVS. Consequently, it satisfies exactly the same inequality (3.3). Moreover, similar to Lemma 3.1, when $e^k = 0$, we have reached a solution of (1.1), (1.2).

We now establish the following analogue of Lemma 4.1.

Lemma 5.1 *If the inner loop sequence $\xi^l := \delta^l \alpha^l \gamma^l$ associated with accelerated BOSVS is nonincreasing as a function of l, then for each $i \in [1, m]$, we have*

$$\nu_i \rho \|\mathbf{z}_i^k - \bar{\mathbf{x}}_i^k\|^2 + \frac{\sigma}{\Gamma_i^k} \sum_{l=1}^{l_i^k} \xi^l \|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 \leq \frac{\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\|^2}{\Gamma_i^k}, \tag{5.4}$$

where l_i^k is the terminating value of l at iteration k, $\bar{\mathbf{x}}_i^k$ is the minimizer of the function L_i^k defined in (3.8), and $\nu_i > 0$ is the smallest eigenvalue of $\mathbf{A}_i^T \mathbf{A}_i$.

Proof By the definition $\mathbf{a}_i^l = (1 - \alpha^l)\mathbf{a}_i^{l-1} + \alpha^l \mathbf{u}_i^l$, we have

$$\langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^l - \bar{\mathbf{a}}_i^l \rangle = (1 - \alpha^l)\langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^{l-1} - \bar{\mathbf{a}}_i^l \rangle + \alpha^l \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u}_i^l - \bar{\mathbf{a}}_i^l \rangle.$$

Add to this the identity $f_i(\bar{\mathbf{a}}_i^l) = (1 - \alpha^l) f_i(\bar{\mathbf{a}}_i^l) + \alpha^l f_i(\bar{\mathbf{a}}_i^l)$ to obtain

$$\begin{aligned} &f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^l - \bar{\mathbf{a}}_i^l \rangle \\ &\quad = (1 - \alpha^l) \left[f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^{l-1} - \bar{\mathbf{a}}_i^l \rangle \right] + \alpha^l \left[f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u}_i^l - \bar{\mathbf{a}}_i^l \rangle \right]. \end{aligned}$$

By the convexity of f_i, it follows that $f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^{l-1} - \bar{\mathbf{a}}_i^l \rangle \leq f_i(\mathbf{a}_i^{l-1})$. Hence,

$$f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^l - \bar{\mathbf{a}}_i^l \rangle \leq (1 - \alpha^l) f_i(\mathbf{a}_i^{l-1}) + \alpha^l \left[f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u}_i^l - \bar{\mathbf{a}}_i^l \rangle \right].$$

Adding and subtracting any $\mathbf{u} \in \mathbb{R}^{n_i}$ in the last term, and then exploiting the convexity of f_i gives

$$\begin{aligned} f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u}_i^l - \bar{\mathbf{a}}_i^l \rangle &= \left[f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u} - \bar{\mathbf{a}}_i^l \rangle \right] + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u}_i^l - \mathbf{u} \rangle \\ &\leq f_i(\mathbf{u}) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u}_i^l - \mathbf{u} \rangle. \end{aligned}$$

Therefore,

$$f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^l - \bar{\mathbf{a}}_i^l \rangle \leq (1 - \alpha^l) f_i(\mathbf{a}_i^{l-1}) + \alpha^l [f_i(\mathbf{u}) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{u}_i^l - \mathbf{u} \rangle]. \tag{5.5}$$

Now by the line search condition in Step 1a of accelerated BOSVS and then by (5.5), we have

$$\begin{aligned} L_i^k(\mathbf{a}_i^l) &= f_i(\mathbf{a}_i^l) + \frac{\rho}{2} \|\mathbf{A}_i \mathbf{a}_i^l - \mathbf{b}_i^k + \boldsymbol{\lambda}^k / \rho\|^2 + h_i(\mathbf{a}_i^l) \\ &\leq f_i(\bar{\mathbf{a}}_i^l) + \langle \nabla f_i(\bar{\mathbf{a}}_i^l), \mathbf{a}_i^l - \bar{\mathbf{a}}_i^l \rangle + \frac{(1 - \sigma)\delta^l}{2\alpha^l} \|\mathbf{a}_i^l - \bar{\mathbf{a}}_i^l\|^2 \end{aligned}$$

$$+\frac{\rho}{2}\|\mathbf{A}_i\mathbf{a}_i^l-\mathbf{b}_i^k+\boldsymbol{\lambda}^k/\rho\|^2+h_i(\mathbf{a}_i^l)$$
$$\le(1-\alpha^l)f_i(\mathbf{a}_i^{l-1})+\alpha^l f_i(\mathbf{u})+\alpha^l\langle\nabla f_i(\bar{\mathbf{a}}_i^l),\mathbf{u}_i^l-\mathbf{u}\rangle+\frac{(1-\sigma)\delta^l}{2\alpha^l}\|\mathbf{a}_i^l-\bar{\mathbf{a}}_i^l\|^2$$
$$+\frac{\rho}{2}\|\mathbf{A}_i\mathbf{a}_i^l-\mathbf{b}_i^k+\boldsymbol{\lambda}^k/\rho\|^2+h_i(\mathbf{a}_i^l).$$

Next, we utilize the definitions of $\mathbf{a}_i^l$ and $\bar{\mathbf{a}}_i^l$ and the convexity of both h_i and the norm term to obtain

$$\begin{aligned}L_i^k(\mathbf{a}_i^l)&\le(1-\alpha^l)f_i(\mathbf{a}_i^{l-1})+\alpha^l[f_i(\mathbf{u})+\langle\nabla f_i(\bar{\mathbf{a}}_i^l),\mathbf{u}_i^l-\mathbf{u}\rangle]+\frac{(1-\sigma)\delta^l}{2\alpha^l}\|\mathbf{a}_i^l-\bar{\mathbf{a}}_i^l\|^2\\
&\quad+(1-\alpha^l)\left(\frac{\rho}{2}\|\mathbf{A}_i\mathbf{a}_i^{l-1}-\mathbf{b}_i^k+\boldsymbol{\lambda}^k/\rho\|^2+h_i(\mathbf{a}_i^{l-1})\right)\\
&\quad+\alpha^l\left(\frac{\rho}{2}\|\mathbf{A}_i\mathbf{u}_i^l-\mathbf{b}_i^k+\boldsymbol{\lambda}^k/\rho\|^2+h_i(\mathbf{u}_i^l)\right)\\
&=(1-\alpha^l)\left(f_i(\mathbf{a}_i^{l-1})+\frac{\rho}{2}\|\mathbf{A}_i\mathbf{a}_i^{l-1}-\mathbf{b}_i^k+\boldsymbol{\lambda}^k/\rho\|^2+h_i(\mathbf{a}_i^{l-1})\right)\\
&\quad+\alpha^l[f_i(\mathbf{u})+\langle\nabla f_i(\bar{\mathbf{a}}_i^l),\mathbf{u}_i^l-\mathbf{u}\rangle]+\frac{(1-\sigma)\delta^l\alpha^l}{2}\|\mathbf{u}_i^l-\mathbf{u}_i^{l-1}\|^2\\
&\quad+\alpha^l\left(\frac{\rho}{2}\|\mathbf{A}_i\mathbf{u}_i^l-\mathbf{b}_i^k+\boldsymbol{\lambda}^k/\rho\|^2+h_i(\mathbf{u}_i^l)\right)\\
&=(1-\alpha^l)L_i^k(\mathbf{a}_i^{l-1})+\alpha^l[f_i(\mathbf{u})+\langle\nabla f_i(\bar{\mathbf{a}}_i^l),\mathbf{u}_i^l-\mathbf{u}\rangle]\\
&\quad+\frac{(1-\sigma)\delta^l\alpha^l}{2}\|\mathbf{u}_i^l-\mathbf{u}_i^{l-1}\|^2+\alpha^l\left(\frac{\rho}{2}\|\mathbf{A}_i\mathbf{u}_i^l-\mathbf{b}_i^k+\boldsymbol{\lambda}^k/\rho\|^2+h_i(\mathbf{u}_i^l)\right).\end{aligned}\tag{5.6}$$

Since h_i is convex, we have

$$h_i(\mathbf{u}^l)+\langle\mathbf{p},\mathbf{u}-\mathbf{u}^l\rangle\le h_i(\mathbf{u})\tag{5.7}$$

for any $\mathbf{p}\in\partial h_i(\mathbf{u}^l)$. The expansion of the quadratic Q in Step 1a of accelerated BOSVS around $\mathbf{u}^l$ can be written

$$Q(\mathbf{u}^l)+\nabla Q(\mathbf{u}^l)(\mathbf{u}-\mathbf{u}^l)+\frac{1}{2}(\mathbf{u}-\mathbf{u}^l)^{\mathsf{T}}(\delta^l\mathbf{I}+\rho\mathbf{A}_i^{\mathsf{T}}\mathbf{A}_i)(\mathbf{u}-\mathbf{u}^l)=Q(\mathbf{u}).\tag{5.8}$$

Since $\mathbf{u}^l$ minimizes $Q+h_i$ in Step 1a, the first-order optimality conditions imply that $\mathbf{p}+\nabla Q(\mathbf{u}^l)=\mathbf{0}$ for some $\mathbf{p}\in\partial h_i(\mathbf{u}^l)$. We choose $\mathbf{p}=-\nabla Q(\mathbf{u}^l)$, and then multiply (5.7) and (5.8) by α^l and add to (5.6) to obtain

$$\begin{aligned}L_i^k(\mathbf{a}_i^l)&\le(1-\alpha^l)L_i^k(\mathbf{a}_i^{l-1})+\alpha^l L_i^k(\mathbf{u})+\frac{\delta^l\alpha^l}{2}(\|\mathbf{u}-\mathbf{u}_i^{l-1}\|^2-\|\mathbf{u}-\mathbf{u}_i^l\|^2)\\
&\quad-\frac{\sigma\delta^l\alpha^l}{2}\|\mathbf{u}_i^l-\mathbf{u}_i^{l-1}\|^2-\frac{\alpha^l\rho}{2}\|\mathbf{A}_i(\mathbf{u}-\mathbf{u}_i^l)\|^2.\end{aligned}$$

Hence, for any $\mathbf{u} \in \mathbb{R}^{n_i}$ we have

$$L_i^k(\mathbf{a}_i^l) - L_i^k(\mathbf{u}) \le (1-\alpha^l)(L_i^k(\mathbf{a}_i^{l-1}) - L_i^k(\mathbf{u})) + \frac{\delta^l\alpha^l}{2}(\|\mathbf{u}-\mathbf{u}_i^{l-1}\|^2 - \|\mathbf{u}-\mathbf{u}_i^l\|^2) - \frac{\sigma\delta^l\alpha^l}{2}\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 - \frac{\alpha^l\rho}{2}\|\mathbf{A}_i(\mathbf{u}-\mathbf{u}_i^l)\|^2. \tag{5.9}$$

From the definition of γ^l in accelerated BOSVS, it follows that $(1-\alpha^l)\gamma^l = \gamma^{l-1}$ with the convention that $\gamma^0 = 0$ (since $\alpha^1 = 1$). Hence, for any sequence d^l, $l \ge 0$, we have

$$\sum_{l=1}^{j}\left(\gamma^l d^l - (1-\alpha^l)\gamma^l d^{l-1}\right) = \sum_{l=1}^{j}\left(\gamma^l d^l - \gamma^{l-1}d^{l-1}\right) = \gamma^j d^j. \tag{5.10}$$

Suppose that $d^l \ge 0$ for each l. By assumption, $\xi^l = \gamma^l\delta^l\alpha^l$ is nonincreasing; since $\alpha^1 = 1$ and $\gamma^1 = 1/\delta^1$, it follows that $\xi^1 = 1$, and we have

$$\sum_{l=1}^{j}\xi^l\left(d^l - d^{l-1}\right) = d^1 - d^0 + \sum_{l=2}^{j}\xi^l\left(d^l - d^{l-1}\right) \ge d^1 - d^0 + \sum_{l=2}^{j}\left(\xi^l d^l - \xi^{l-1}d^{l-1}\right) = \xi^j d^j - d^0. \tag{5.11}$$

We now multiply (5.9) by γ^l and sum over l between 1 and l_i^k. Exploiting the identity (5.10) with $d^l = L_i^k(\mathbf{a}_i^l) - L_i^k(\mathbf{u})$ and (5.11) with $d^l = \|\mathbf{u}_i^l - \mathbf{u}\|^2$, we obtain

$$L_i^k(\mathbf{u}) - L_i^k(\mathbf{a}_i^{l_i^k}) \ge \frac{1}{2\Gamma_i^k}(\xi^{l_i^k}\|\mathbf{u}-\mathbf{u}_i^{l_i^k}\|^2 - \|\mathbf{u}-\mathbf{u}_i^0\|^2) + \frac{\sigma}{2\Gamma_i^k}\sum_{l=1}^{l_i^k}\xi^l\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 + \frac{\rho}{2\Gamma_i^k}\sum_{l=1}^{l_i^k}(\gamma^l\alpha^l)\|\mathbf{A}_i(\mathbf{u}-\mathbf{u}_i^l)\|^2, \tag{5.12}$$

where Γ_i^k denotes the final γ^l in accelerated BOSVS.

Next, we multiply the definition $\mathbf{a}_i^j = (1-\alpha^j)\mathbf{a}_i^{j-1} + \alpha^l\mathbf{u}_i^j$ by γ^j and sum over j between 1 and l. Again, exploiting the identity $(1-\alpha^j)\gamma^j = \gamma^{j-1}$ yields

$$\mathbf{a}_i^l = \frac{1}{\gamma^l}\sum_{j=1}^{l}(\gamma^j\alpha^j)\mathbf{u}_i^j. \tag{5.13}$$

Since $\alpha^j \gamma^j = \gamma^j - \gamma^{j-1}$, it follows that

$$\gamma^l = \sum_{j=1}^{l} \alpha^j \gamma^j. \tag{5.14}$$

Consequently, $\mathbf{a}_i^l$ is a convex combination of $\mathbf{u}_i^1$ through $\mathbf{u}_i^l$. Since $\|\mathbf{A}_i(\mathbf{u}-\mathbf{w})\|^2$ is a convex function of $\mathbf{w}$, Jensen's inequality yields

$$\frac{1}{\Gamma_i^k}\sum_{l=1}^{l_i^k}(\gamma^l\alpha^l)\|\mathbf{A}_i(\mathbf{u}-\mathbf{u}_i^l)\|^2 \ge \|\mathbf{A}_i(\mathbf{u}-\mathbf{a}_i^{l_i^k})\|^2 = \|\mathbf{A}_i(\mathbf{u}-\mathbf{z}_i^k)\|^2.$$

We apply this inequality to the last term in (5.12) and substitute $\mathbf{z}_i^k = \mathbf{a}_i^{l_i^k}$, $\mathbf{x}_i^{k+1} = \mathbf{u}_i^{l_i^k}$, and $\mathbf{x}_i^k = \mathbf{u}_i^0$ to obtain

$$\begin{aligned} L_i^k(\mathbf{u}) - L_i^k(\mathbf{z}_i^k) \ge \frac{1}{2\Gamma_i^k}(\xi^{l_i^k}\|\mathbf{u}-\mathbf{x}_i^{k+1}\|^2 - \|\mathbf{u}-\mathbf{x}_i^k\|^2) + \frac{\sigma}{2\Gamma_i^k}\sum_{l=1}^{l_i^k}\xi^l\|\mathbf{u}_i^l-\mathbf{u}_i^{l-1}\|^2 \\ +\frac{\rho}{2}\|\mathbf{A}_i(\mathbf{u}-\mathbf{z}_i^k)\|^2. \end{aligned} \tag{5.15}$$

Finally, take $\mathbf{u} = \bar{\mathbf{x}}_i^k$. Since the left side of (5.15) is nonpositive for this choice of $\mathbf{u}$, the proof is complete. □

Let us now examine the assumptions and consequences of Lemma 5.1 in the context of the choices (5.1) and (5.2) for the parameters δ^l and α^l. For the choice (5.1) and for $l \ge 2$, we have

$$\gamma^l = \frac{1}{\delta^1}\prod_{j=2}^{l}(1-\alpha^j)^{-1} = \frac{1}{\delta^1}\prod_{j=2}^{l}\frac{j+1}{j-1} = \frac{1}{\delta^1}\frac{l(l+1)}{2}. \tag{5.16}$$

Hence, γ^l is $O(l^2)$. Since $\delta^l = \delta^1/l$, it follows that for $l \ge 2$,

$$\xi^l := \delta^l\alpha^l\gamma^l = \left(\frac{\delta^1}{l}\right)\left(\frac{2}{l+1}\right)\left(\frac{l(l+1)}{2\delta^1}\right) = 1.$$

In the special case $l = 1$, $\xi^1 = \delta^1/\delta^1 = 1$. Since the sequence ξ^l is identically one, it is nonincreasing and the assumption of Lemma 5.1 is satisfied. Since Γ_i^k is the final value for γ^l in Step 1 of accelerated BOSVS, it follows from (5.4) that $\|\mathbf{z}_i^k - \bar{\mathbf{x}}_i^k\| = O(1/l_i^k)$.

For the choice (5.2) and for $l \ge 2$, we have $\Lambda^l = (1/\delta^l) + \Lambda^{l-1}$ and $\alpha^l = (1/\delta^l)/\Lambda^l$. It follows that $1-\alpha^l = \Lambda^{l-1}/\Lambda^l$ and for $l \ge 2$, we have

$$\gamma^l = \frac{1}{\delta^1}\prod_{j=2}^{l}(1-\alpha^j)^{-1} = \frac{1}{\delta^1}\prod_{j=2}^{l}(\Lambda^j/\Lambda^{j-1}) = \frac{1}{\delta^1}\frac{\Lambda^l}{\Lambda^1} = \Lambda^l.$$

Hence,

$$\xi^l := \delta^l \alpha^l \gamma^l = \delta^l \left(\frac{1/\delta^l}{\Lambda^l} \right) \Lambda^l = 1.$$

In the special case $l = 1$, we also have $\xi^1 = 1$. Again, the sequence ξ^l is identically one, which satisfies the requirement of Lemma 5.1; consequently, the speed with which $\mathbf{z}_i^k$ converges to $\bar{\mathbf{x}}_i^k$ depends on the growth rate of γ^l. By the definition of γ^l in accelerated BOSVS,

$$\sqrt{\gamma^l} - \sqrt{\gamma^{l-1}} = \sqrt{\gamma^l} - \sqrt{(1-\alpha^l)\gamma^l} = \left(1 - \sqrt{1-\alpha^l}\right)\sqrt{\gamma^l} \geq \frac{\alpha^l \sqrt{\gamma^l}}{2}. \quad (5.17)$$

Since $\xi^l := \delta^l \alpha^l \gamma^l = 1$, it follows from (5.3) that $(\alpha^l/\theta^l)\alpha^l\gamma^l = (\alpha^l)^2\gamma^l/\theta^l = 1$, which implies that

$$\alpha^l \sqrt{\gamma^l} = \sqrt{\theta^l}. \quad (5.18)$$

By (5.17), we have

$$\sqrt{\gamma^l} - \sqrt{\gamma^{l-1}} \geq \frac{\sqrt{\theta^l}}{2}. \quad (5.19)$$

As noted beneath (5.3), $1/\theta^l$ satisfies the inequality (3.3) for δ_i^k, which implies that

$$\theta^l \geq \Theta := \frac{1-\sigma}{\eta\zeta_i + (1-\sigma)\delta_{\max}}. \quad (5.20)$$

Hence, (5.19) yields $\sqrt{\gamma^l} - \sqrt{\gamma^{l-1}} \geq \sqrt{\Theta}/2$. Since $\gamma^1 = 1/\delta^1 = \theta^1$, it follows that

$$\sqrt{\gamma^l} \geq \sqrt{\Theta} + \left(\frac{l-1}{2}\right)\sqrt{\Theta} \geq \left(\frac{l}{2}\right)\sqrt{\Theta} \quad \text{or} \quad \gamma^l \geq \left(\frac{l^2}{4}\right)\Theta.$$

In summary, for either of the choices (5.1) or (5.2), we have $\xi^l = 1$ for each l, and $\|\mathbf{z}_i^k - \bar{\mathbf{x}}_i^k\| = O(1/l_i^k)$. Moreover, by the inequality (5.15) with $\mathbf{u} = \bar{\mathbf{x}}_i^k$, the objective value satisfies $L_i^k(\mathbf{z}_i^k) - L_i^k(\bar{\mathbf{x}}_i^k) = O(1/(l_i^k)^2)$.

Although Lemma 5.1 was stated in terms of the terminating iteration l_i^k of the inner iteration, it applies to any of the inner iterations; that is, for each i and l, we have

$$\nu_i \rho \|\mathbf{a}^l - \bar{\mathbf{x}}_i^k\|^2 + \frac{\sigma}{\gamma^l} \sum_{j=1}^{l} \xi^j \|\mathbf{u}_i^j - \mathbf{u}_i^{j-1}\|^2 \leq \frac{\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\|^2}{\gamma^l}.$$

Whenever γ^l approaches infinity, as it does with the choices (5.1) and (5.2), the right side approach zero and $\mathbf{a}^l$ converges to $\bar{\mathbf{x}}_i^k$. Hence, the stopping conditions in Step 1b of accelerated BOSVS are satisfied for l sufficiently large when $e^{k-1} \neq 0$.

The convergence of accelerated BOSVS, like the other algorithms, relies on a decay property for the iterates, which we now give.

Lemma 5.2 *If the accelerated BOSVS parameters γ^l tend infinity as l grows and $\xi^l := \delta^l \alpha^l \gamma^l = 1$ for each l, then Lemma 4.2 holds for the accelerated scheme.*

Proof We substitute $\mathbf{u} = \mathbf{x}_i^*$ and $\xi^l = 1$ in (5.15) to obtain

$$L_i^k(\mathbf{x}_i^*) - F_i^k(\mathbf{z}_i^k) \geq \frac{1}{2\Gamma_i^k}\left(\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 - \|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2\right) + \frac{\sigma}{2\Gamma_i^k}\sum_{l=1}^{l_i^k}\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2,$$

where $F_i^k(\mathbf{w}) = L_i^k(\mathbf{w}) + (\rho/2)\|\mathbf{A}_i(\mathbf{w} - \mathbf{x}_i^*)\|^2$. This is exactly the same as (4.8) in the proof of Lemma 4.2. The remainder of the proof is exactly as in the proof of Lemma 4.2. □

Using the decay property of Lemmas 4.2 and 5.2 , we now obtain the convergence of accelerated BOSVS.

Theorem 5.3 *Suppose that for the inner loop sequence $\xi^l := \delta^l \alpha^l \gamma^l$ associated with accelerated BOSVS we have $\xi^l = 1$ for each l, γ^l tends to infinity as l grows, and there exists a constant $\kappa > 0$ such that $\gamma^l(\alpha^l)^2 \geq \kappa$ for all l. If accelerated BOSVS performs an infinite number of iterations generating iterates $\mathbf{y}^k$, $\mathbf{z}^k$, and $\boldsymbol{\lambda}^k$, then the sequences $\mathbf{y}^k$ and $\mathbf{z}^k$ both approach a common limit $\mathbf{x}^*$ and $\boldsymbol{\lambda}^k$ approaches a limit $\boldsymbol{\lambda}^*$ where $(\mathbf{x}^*, \boldsymbol{\lambda}^*) \in \mathcal{W}^*$.*

Proof The proof is identical to that of Theorem 4.3 through the end of Case 1. For accelerated BOSVS, the fact that $\mathbf{z}_i^k$ is a convex combination of $\mathbf{u}_{i,k}^l$ is shown in (5.13), (5.14). The treatment of accelerated BOSVS first differs from that of multistep BOSVS in the second paragraph of Case 2 (Γ_i^k tends to $+\infty$) where the multistep BOSVS stopping condition $\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|/\sqrt{\gamma^l} \leq \psi(e^{k-1})$, is used to show that $\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\|^2/\Gamma_i^k$ approaches zero. Since accelerated BOSVS uses the new stopping condition $\|\mathbf{a}_i^l - \mathbf{a}_i^{l-1}\| \leq \psi(e^{k-1})$, a new analysis is needed in Case 2.

By the definition of $\mathbf{a}^l$, we have

$$\|\mathbf{a}^l - \mathbf{a}^{l-1}\| = \alpha^l\|\mathbf{u}^l - \mathbf{a}^{l-1}\| \geq \alpha^l(\|\mathbf{u}^l - \mathbf{a}^l\| - \|\mathbf{a}^l - \mathbf{a}^{l-1}\|).$$

If ψ_k denotes $\psi(e^{k-1})$ and $l = l_i^k$ so that $\mathbf{a}^l$ satisfies the stopping criterion $\|\mathbf{a}_i^l - \mathbf{a}_i^{l-1}\| \leq \psi_k$, then

$$\alpha^l\|\mathbf{u}^l - \mathbf{a}^l\| \leq (1+\alpha^l)\|\mathbf{a}^l - \mathbf{a}^{l-1}\| \leq 2\psi_k \quad \text{or} \quad \|\mathbf{x}_i^{k+1} - \mathbf{z}_i^k\| \leq \frac{2\psi_k}{\alpha^l}$$

since $\mathbf{u}^l = \mathbf{x}_i^{k+1}$ and $\mathbf{a}^l = \mathbf{z}_i^k$ when $l = l_i^k$. Squaring this, dividing by $\gamma^l = \Gamma_i^k$, and utilizing the assumption that $\gamma^l(\alpha^l)^2 \geq \kappa$ for all l, we deduce that

$$\frac{\|\mathbf{x}_i^{k+1} - \mathbf{z}_i^k\|^2}{\Gamma_i^k} \leq \frac{4\psi_k^2}{\kappa}. \tag{5.21}$$

Since ψ_k approach zero by (4.18), it follows that $\|\mathbf{x}_i^{k+1} - \mathbf{z}_i^k\|^2/\Gamma_i^k$ approaches zero as k tends to infinity. Since Γ_i^k is nondecreasing, $\|\mathbf{x}_i^{k+1} - \mathbf{z}_i^k\|^2/\Gamma_i^{k+1}$ also approaches zero as k tends to infinity. Since $\mathbf{z}_i^k$ is a bounded sequence and Γ_i^k tends to infinity in Case 2, we can replace $\mathbf{z}_i^k$ by any other bounded sequence and reach the same conclusion. In particular, since the sequence $\bar{\mathbf{x}}_i^k$ is bounded we conclude that $\|\mathbf{x}_i^k - \bar{\mathbf{x}}_i^k\|^2/\Gamma_i^k$ approaches zero as k tends to infinity, the same conclusion we reached in multistep BOSVS scheme. The rest of the proof is exactly as in Theorem 4.3. This completes the proof. □

Remark 5.1 The parameter choices given in both (5.1) and (5.2) satisfy the assumption of Theorem 5.3 that $\gamma^l(\alpha^l)^2 \geq \kappa > 0$ for some constant κ. In particular, for (5.1), we show in (5.16) that $\gamma^l = l(l+1)/(2\delta^1)$. This is combined with the definition of α^l in (5.1) to obtain

$$\gamma^l(\alpha^l)^2 = \frac{2l}{\delta^1(l+1)} \geq \frac{1}{\delta^1}$$

for $l \geq 1$. For the choice (5.2), it follows from (5.18) and (5.20) that

$$\gamma^l(\alpha^l)^2 \geq \Theta := \frac{1-\sigma}{\eta\zeta_i + (1-\sigma)\delta_{\max}}.$$

Remark 5.2 In this paper, we have focused on algorithms based on an inexact minimization of L_i^k in Step 1 of Algorithm 2.1. In cases where f_i and h_i are simple enough that the exact minimizer $\bar{\mathbf{x}}_i^k$ of L_i^k can be quickly evaluated, we could simply set $\mathbf{x}_i^{k+1} = \mathbf{z}_i^k = \bar{\mathbf{x}}_i^k$ and $r_i^k = 0$ in Step 1. The analysis of this exact algorithm is very similar to the analysis in Theorems 4.3 and 5.3 .

6 Numerical experiments

In this section, we investigate the performance of the algorithms for an image reconstruction problem that can be formulated as

$$\min_{\mathbf{u}} \ \frac{1}{2}\|\mathbf{Fu} - \mathbf{f}\|^2 + \alpha\|\mathbf{u}\|_{TV} + \beta\|\boldsymbol{\Psi}^{\mathsf{T}}\mathbf{u}\|_1, \tag{6.1}$$

where $\mathbf{f}$ is the given image data, $\mathbf{F}$ is a matrix describing the imaging device, $\|\cdot\|_{TV}$ is the total variation norm, $\|\cdot\|_1$ is the ℓ_1 norm, $\boldsymbol{\Psi}$ is a wavelet transform, and $\alpha > 0$ and $\beta > 0$ are weights. The first term in the objective is the data fidelity term, while the next two terms are for regularization; they are designed to enhance edges and increase image sparsity. In our experiments, $\boldsymbol{\Psi}$ is a normalized Haar wavelet with four levels and $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{T}} = I$. The problem (6.1) is equivalent to

$$\min_{(\mathbf{u},\mathbf{w},\mathbf{z})} \ \frac{1}{2}\|\mathbf{Fu} - \mathbf{f}\|^2 + \alpha\|\mathbf{w}\|_{1,2} + \beta\|\mathbf{z}\|_1 \text{ subject to } \mathbf{Bu} = \mathbf{w}, \ \boldsymbol{\Psi}^{\mathsf{T}}\mathbf{u} = \mathbf{z}, \tag{6.2}$$

where $\mathbf{Bu} = \nabla\mathbf{u}$ and $(\nabla\mathbf{u})_i$ is the vector of finite differences in the image along the coordinate directions at the i-th pixel in the image, $\|\mathbf{w}\|_{1,2} = \sum_{i=1}^{N} \|(\nabla\mathbf{u})_i\|_2$, and N is the total number of pixels in the image.

The problem (6.2) has the structure appearing in (1.1), (1.2) with $h_1 := 0$, $f_1(\mathbf{u}) = 1/2\|\mathbf{Fu} - \mathbf{f}\|^2$, $h_2(\mathbf{w}) = \|\mathbf{w}\|_{1,2}$, $f_2 := 0$, $h_3(\mathbf{z}) = \|\mathbf{z}\|_1$, $f_3 := 0$,

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{B} \\ \boldsymbol{\Psi}^\mathsf{T} \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} -\mathbf{I} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{A}_3 = \begin{pmatrix} \mathbf{0} \\ -\mathbf{I} \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

When solving the test problems using accelerated BOSVS, we use choose α^l and δ^l as in (5.2). Since $f_2 = f_3 = 0$, the line search condition holds automatically, and the second and third subproblems are solved in closed form, due to the simple structure of h_2 and h_3. Only the first subproblem is solved inexactly. At iteration k, the solution of this subproblem approximates the solution of

$$\min_{\mathbf{u}} \; L_1^k(\mathbf{u}) := \frac{1}{2}\|\mathbf{Fu} - \mathbf{f}\|^2 + \frac{\rho}{2}\|\mathbf{Bu} - \mathbf{w}^k + \rho^{-1}\boldsymbol{\lambda}^k\|^2 + \frac{\rho}{2}\|\boldsymbol{\Psi}^\mathsf{T}\mathbf{u} - \mathbf{z}^k + \rho^{-1}\boldsymbol{\mu}^k\|^2, \tag{6.3}$$

where $\boldsymbol{\lambda}^k$ and $\boldsymbol{\mu}^k$ are the Lagrange multipliers at iteration k for the constraints $\mathbf{Bu} = \mathbf{w}$ and $\boldsymbol{\Psi}^\mathsf{T}\mathbf{u} = \mathbf{z}$ respectively.

The stopping condition for the inner loop of either multistep or accelerated BOSVS required that $\Gamma_i^k \geq \Gamma_i^{k-1}$. To improve efficiency, we replaced this condition by $l_i^k \geq l_i^{k-1}$ or $\Gamma_i^k \geq \Gamma_i^{k-1}$, where l_i^k is the number of iterations performed by the inner loop for block i at iteration k. For all the algorithm, we chose the initial δ_0^l in the line search using the BB approximation, which is given in Step 1a of generalized BOSVS. Moreover, when $\Gamma_i^k < \Gamma_i^{k-1}$, we increase $\delta_{\min,i}$ by setting $\delta_{\min,i} := \tau\delta_{\min,i}$, where $\tau = 1.1$ in our numerical experiments. When $\delta_{\min,i}$ is sufficiently large, we have $\delta_0^l = \delta_{\min,i}$ and the line search condition in the algorithms is satisfied by δ_0^l; that is, $\delta^l = \delta_0^l = \delta_{\min,i}$. Consequently, when $\delta_{\min,i}$ is sufficiently large, we have

$$\Gamma_i^k = \sum_{l=1}^{l_i^k} \frac{1}{\delta^l} = \frac{l_i^k}{\delta_{\min,i}},$$

and the relaxed stopping condition $l_i^k \geq l_i^{k-1}$ implies that $\Gamma_i^k \geq \Gamma_i^{k-1}$, the original stopping condition. Since $\tau > 1$, it follows that $\Gamma_i^k < \Gamma_i^{k-1}$ for only a finite number of iterations, and hence, $\Gamma_i^k \geq \Gamma_i^{k-1}$ for k sufficiently large. This ensures the global convergence of the algorithms.

Another improvement to efficiency was achieved by further relaxing the line search criterion. In particular, for the line search in generalized BOSVS (Step 1b), we replaced the right side $f_i(\mathbf{x}_i^{k+1})$ by $f_i(\mathbf{x}_i^{k+1}) - \epsilon^k$ where $\epsilon^k \geq 0$ is a summable sequence. In the line search of multistep BOSVS (Step 1b), $f_i(\mathbf{u}_i^l)$ was replaced by $f_i(\mathbf{u}_i^l) - \pi^l$, where $\pi^l = \epsilon^k\delta^l\omega^l$ with ω^l a summable sequence. In the line search of accelerated

BOSVS (Step 1a), we replaced $f_i(\mathbf{a}_i^l)$ by $f_i(\mathbf{a}_i^l) - \pi^l$, where $\pi^l = \epsilon^k \omega^l / \gamma^l$. It can be proved that when the line search is relaxed in this way using summable sequences, there is no effect on the global convergence theory; these ϵ^k and π^l terms need to be inserted in each inequality in the analysis, but in the end, the steps and the conclusions are unchanged. On the other hand, when the line search is relaxed, it can terminate sooner, and the algorithms can be more efficient. For the numerical experiments, we took $\epsilon^k = 10/k^{1.1}$. For multistep BOSVS, $\omega^l = 1/(\gamma^l)^{1.2}$, while for accelerated BOSVS, $\omega^l = 1/(\gamma^l)^{0.6}$. Since γ^l grows in proportion to l for multistep BOSVS and in proportion to l^2 for accelerated BOSVS, the ω^l sequences are summable. Hence, for these choices of ϵ^k and ω^l, global convergence is guaranteed. The specific exponents 1.1, 1.2, and 0.6 in the formulas for ϵ^k and ω^l seemed to work reasonably well in our experiments.

In all the algorithms, we use the following parameters:

$$\delta_{\min} = 10^{-10}, \ \delta_{\max} = 10^{10}, \ \alpha = 0.999, \ \sigma = 10^{-5}, \ \eta = 3, \ \text{and} \ \tau = 1.1.$$

For the inner loop stopping condition, we took $\psi(t) = \min\{0.1t, t^{1.1}\}$ in multistep BOSVS, and $\psi(t) = 0.5t$ in accelerated BOSVS, while in Step 2 of the ADMM template Algorithm 2.1, we took $\theta_1 = 10^{-6}\sqrt{\rho}, \theta_2 = \sqrt{\rho}$, and $\theta_3 = 10^{-6}\sqrt{\sigma/(1-\alpha)}$. For comparison, we provide numerical results based on the algorithm in [24] where we use MATLAB's conjugate gradient routine CGS with starting point u^{k-1}, the solution of the subproblem at the previous iteration, to solve the subproblem (6.3) almost exactly, stopping when $\|\nabla L_1^k(\mathbf{u})\| \le 10^{-6}$. The algorithm in [24] was guaranteed to converge due to a back substitution step. We also implemented ADMM without the back substitution step; in this case, there is no convergence guarantee. MATLAB's conjugate gradient routine was utilized for the subproblem (6.3) since the objective is quadratic with a positive definite Hessian, and the conjugate gradient method works reasonably well in this case. All the codes were implemented in MATLAB (version R2014a). The following figures show the relative objective error $(\Phi(\mathbf{u}^k) - \Phi^*)/\Phi^*$ versus CPU time, where Φ^* is the optimal function value of (6.1) obtained by applying accelerated BOSVS until the eighth digit of the relative objective value did not change in four consecutive iterations.

The first experiment employs an image deblurring problem from [1]. The original image is the well-known Cameraman image of size 256×256 and the observed data $\mathbf{f}$ in (6.1) is a blurred image obtained by imposing a uniform blur of size 9×9 with Gaussian noise and SNR of 40dB. The weights in (6.1) are $\alpha = 0.005$ and $\beta = 0.001$, and the penalty parameter $\rho = 5 \times 10^{-4}$. Figure 1a shows the base-10 logarithm of the relative objective error versus CPU time. In this problem where the subproblems are relatively easy, generalized BOSVS is significantly slower than the other algorithms, while both multistep and accelerated BOSVS were faster than the exact ADMM schemes.

The second set of test problems, which arise in partially parallel imaging (PPI), are found in [10]. The observed data, corresponding to 3 different images, are denoted data 1, data 2, and data 3. For these test problems, the weights in (6.1) are $\alpha = 10^{-5}$ and $\beta = 10^{-6}$, and the penalty parameter $\rho = 10^{-3}$. The performance of the algorithms is shown in Fig. 1b–d. These test problems are much more difficult than the first problem since $\mathbf{F}$ is large, relatively dense, and ill conditioned. In this case, all the inexact

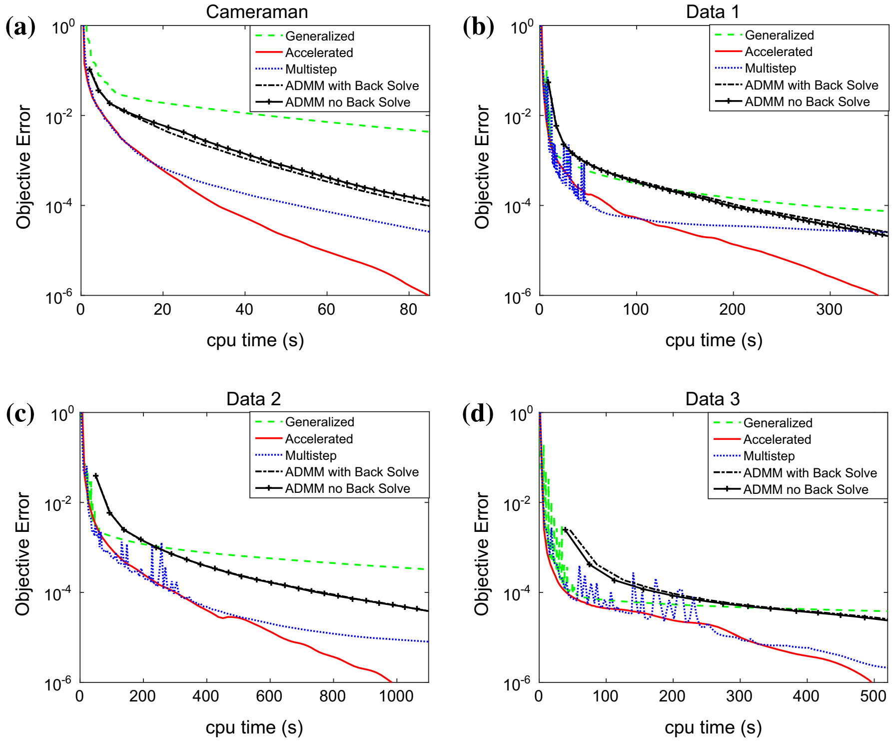

Fig. 1 Base-10 logarithm of the relative objective error versus CPU time for the test problems

algorithms are faster than the exact ADMM algorithms initially. The exact algorithms becomes faster than generalized BOSVS when the relative error is around 10^{-3} or 10^{-4}. Accelerated BOSVS is always significantly faster than the exact algorithms.

7 Conclusion

Three inexact alternating direction multiplier methods were presented for solving separable convex linearly constrained optimization problems, where the objective function is the sum of smooth and relatively simple nonsmooth terms. The nonsmooth terms could be infinite, so the algorithms and analysis included problems with additional convex constraints. These algorithms all originate from the 2-block variable stepsize BOSVS scheme of [10,20] which employs indefinite proximal terms and linearized subproblems. The 2-block scheme was generalized to a multiblock scheme using a back substitution process to generate an auxiliary sequence $\mathbf{y}^k$ that played the role of $\mathbf{x}^k$ in the original, potentially divergent [5], multiblock ADMM (1.4). The three new methods, called generalized, multistep, and accelerated BOSVS, correspond to different accuracy levels when solving the ADMM subproblems. Generalized BOSVS employed only one iteration in the subproblems, while multistep and accelerated BOSVS performed multiple iterations until the iteration change was sufficiently small. The multistep and accelerated schemes differed in the rate with which they solved the the subproblems. If l was the number of iterations in the subproblem, then multistep BOSVS had a convergence rate of $O(1/l)$, while accelerated BOSVS had a convergence rate of $O(1/l^2)$. Global convergence was established for all the methods. Numerical experiments were performed using image reconstruction problems. The accelerated BOSVS algorithm had the best performance when compared with either the other inexact algorithms, or the exact algorithm of [24]. This paper established global convergence of the proposed inexact ADMM methods. The overall iteration complexities of these methods will be developed in a separate paper.

References

1. Afonso, M., Bioucas-Dias, J., Figueiredo, M.: Fast image recovery using variable splitting and constrained optimization. IEEE Trans. Image Process. **19**, 2345–2356 (2010)
2. Barzilai, J., Borwein, J.M.: Two point step size gradient methods. IMA J. Numer. Anal. **8**, 141–148 (1988)
3. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Mach. Learn. **3**, 1–122 (2010)
4. Cai, X., Han, D., Yuan, X.: 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**, 39–73 (2017)
5. Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math. Program. **155**, 57–79 (2016)
6. Chen, C., Li, M., Liu, X., Ye, Y.: On the convergence of multi-block alternating direction method of multipliers and block coordinate descent method (2015). arXiv:1508.00193
7. Chen, C., Shen, Y., You, Y.: On the convergence analysis of the alternating direction method of multipliers with three blocks. Abstr. Appl. Anal. (2013). https://doi.org/10.1155/2013/183961
8. Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program. **64**, 81–101 (1994)
9. Chen, L., Sun, D., Toh, K.: An efficient inexact symmetric Gauss–Seidel based majorized ADMM for high-dimensional convex composite conic programming. Math. Program. **161**, 237–270 (2017)
10. Chen, Y., Hager, W.W., Yashtini, M., Ye, X., Zhang, H.: Bregman operator splitting with variable stepsize for total variation image reconstruction. Comput. Optim. Appl. **54**, 317–342 (2013)
11. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set Valued Var. Anal. **25**, 829–858 (2017)
12. Eckstein, J., Bertsekas, D.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
13. Eckstein, J., Silva, P.J.S.: A practical relative error criterion for augmented Lagrangians. Math. Program. **141**, 319–348 (2013)
14. Eckstein, J., Yao, W.: Approximate ADMM algorithms derived from Lagrangian splitting. Comput. Optim. Appl. **68**, 363–405 (2017)
15. Eckstein, J., Yao, W.: Relative-error approximate versions of Douglas–Rachford splitting and special cases of the ADMM. Math. Program. **170**, 417–444 (2018)
16. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
17. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, New York (1984)
18. Goldfarb, D., Ma, S.: Fast multiple-splitting algorithms for convex optimization. SIAM J. Optim. **22**, 533–556 (2012)
19. Golshtein, E., Tretyakov, N.: Modified Lagrangians in convex programming and their generalizations, in point-to-set maps and mathematical programming. In: Huard, P. (ed.) Mathematical Programming Studies, vol. 10, pp. 86–97. Springer, Berlin (1979)

20. Hager, W.W., Yashtini, M., Zhang, H.: An O(1/k) convergence rate for the variable stepsize Bregman operator splitting algorithm. SIAM J. Numer. Anal. **54**, 1535–1556 (2016)
21. Hager, W.W., Zhang, H.: Projection onto a polyhedron that exploits sparsity. SIAM J. Optim. **29**, 1773–1798 (2016)
22. Han, D., Yuan, X.: A note on the alternating direction method of multipliers. J. Optim. Theory Appl. **155**, 227–238 (2012)
23. He, B., Liao, L., Han, D., Yan, H.: A new inexact alternating directions method for monotone variational inequalities. Math. Program. **92**, 103–118 (2002)
24. He, B., Tao, M., Xu, M., Yuan, X.: An alternating direction-based contraction method for linearly constrained separable convex programming problems. Optimization **62**, 573–596 (2013)
25. He, B., Tao, M., Yuan, X.: Alternating direction method with Gaussian back substitution for separable convex programming. SIAM J. Optim. **22**, 313–340 (2012)
26. He, B., Yuan, X.: On the $\mathcal{O}(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
27. Hong, M., Luo, Z.-Q.: On the linear convergence of the alternating direction method of multipliers. Math. Program. **162**, 165–199 (2017)
28. Li, J., Chen, G., Dong, Z., Wu, Z.: A fast dual proximal-gradient method for separable convex optimization with linear coupled constraints. Comput. Optim. Appl. **64**, 671–697 (2016)
29. Li, M., Liao, L., Yuan, X.: Inexact alternating direction methods of multipliers with logarithmic–quadratic proximal regularization. J. Optim. Theory Appl. **159**, 412–436 (2013)
30. Li, M., Sun, D., Toh, K.C.: A convergent 3-block semi-proximal ADMM for for convex minimization problems with one strongly convex block. Asia Pac. J. Oper. Res. **32**, 1–19 (2015)
31. Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the ADMM with multiblock variables. SIAM J. Optim. **25**, 1478–1497 (2015)
32. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
33. Solodov, M.V., Svaiter, B.F.: An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. Math. Oper. Res. **25**, 214–230 (2000)
34. Tao, M., Yuan, X.M.: Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM J. Optim. **21**, 57–81 (2011)
35. Wen, Z., Goldfarb, D., Yin, W.: Alternating direction augmented Lagrangian methods for semidefinite programming. Math. Program. Comput. **2**, 203–230 (2010)