arXiv:2102.05433v1 [math.OC] 10 Feb 2021

A FRAMEWORK OF INERTIAL ALTERNATING DIRECTION METHOD OF MULTIPLIERS FOR NON-CONVEX NON-SMOOTH OPTIMIZATION*

LE THI KHANH HIEN†, DUY NHAT PHAN‡, AND NICOLAS GILLIS†

Abstract. In this paper, we propose an algorithmic framework dubbed inertial alternating direction methods of multipliers (iADMM), for solving a class of nonconvex nonsmooth multiblock composite optimization problems with linear constraints. Our framework employs the general minimization-majorization (MM) principle to update each block of variables so as to not only unify the convergence analysis of previous ADMM that use specific surrogate functions in the MM step, but also lead to new efficient ADMM schemes. To the best of our knowledge, in the *nonconvex nonsmooth* setting, ADMM used in combination with the MM principle to update each block of variables, and ADMM combined with inertial terms for the primal variables have not been studied in the literature. Under standard assumptions, we prove the subsequential convergence and global convergence for the generated sequence of iterates. We illustrate the effectiveness of iADMM on a class of nonconvex low-rank representation problems.

1. Introduction.

In this paper, we consider the following nonconvex minimization problem with linear constraint

$$\begin{aligned} \min_{x,y} \quad & F(x_1,\ldots,x_s)+h(y) \\ \text{such that} \quad & \sum_{i=1}^{s}\mathcal{A}_i x_i+\mathcal{B}y=b, \end{aligned} \tag{1}$$

where $y\in\mathbb{R}^{\mathbf{q}}$, $x_i\in\mathbb{R}^{\mathbf{n}_i}$, $x:=[x_1;\ldots;x_s]\in\mathbb{R}^{\mathbf{n}}$, $\mathbf{n}=\sum_{i=1}^{s}\mathbf{n}_i$, $\mathcal{A}_i$ is a linear map from $\mathbb{R}^{\mathbf{n}_i}$ to $\mathbb{R}^{\mathbf{m}}$, $\mathcal{B}$ is a linear map from $\mathbb{R}^{\mathbf{q}}$ to $\mathbb{R}^{\mathbf{m}}$, $b\in\mathbb{R}^{\mathbf{m}}$, $h:\mathbb{R}^{\mathbf{q}}\to\mathbb{R}$ is a differentiable function, and $F(x)=f(x)+\sum_{i=1}^{s}g_i(x_i)$, where $f:\mathbb{R}^{\mathbf{n}}\to\mathbb{R}$ is a nonconvex nonsmooth function and $g_i:\mathbb{R}^{\mathbf{n}_i}\to\mathbb{R}\cup\{+\infty\}$ are proper lower semi-continuous functions for $i=1,2,\ldots,s$. We assume that F satisfies[1] $\partial F(x)=\partial_{x_1}F(x)\times\ldots\times\partial_{x_s}F(x)$, where ∂F denote the limiting subdifferential of F (see the definition in the supplementary document).

Notation. We denote $[s]:=\{1,\ldots,s\}$. For the $\mathbf{p}$-dimensional Euclidean space $\mathbb{R}^{\mathbf{p}}$, we use $\langle\cdot,\cdot\rangle$ to denote the inner product and $\|\cdot\|$ to denote the corresponding induced norm. For a linear map $\mathcal{M}$, $\mathcal{M}^*$ denotes the adjoint linear map with respect to the inner product and $\|\mathcal{M}\|$ is the induced operator norm of $\mathcal{M}$. We use $\mathcal{I}$ to denote the identity map. For a positive definite self-adjoint operator $\mathcal{Q}$, we denote $\|x\|_{\mathcal{Q}}^2:=\langle x,\mathcal{Q}x\rangle$. We denote the smallest eigenvalue of a symmetric linear self-map (that is, $\mathcal{M}=\mathcal{M}^*$) by $\lambda_{\min}(\mathcal{M})$. We use $Im(\mathcal{B})$ to denote the image of $\mathcal{B}$.

1.1. Nonconvex low-rank representation problem.

One of the important applications of Problem (1) is the following generalized nonconvex low-rank represen-

*L.T.K. Hien and D.N. Phan contributed equally to this work

Funding: L. T. K. Hien and N. Gillis are supported by the Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO) under EOS project no 30468160 (SeLMA), and by the European Research Council (ERC starting grant 679515).

†Department of Mathematics and Operational Research, University of Mons, Belgium (thikhanhhien.le@umons.ac.be, nicolas.gillis@umons.ac.be).

‡Department of Mathematics and Informatics, HCMC University of Education, Vietnam (nhatpd@hcmue.edu.vn).

[1]This condition is satisfied when f is a sum of a continuously differentiable function and a block separable function that has limiting subdifferential, see [2, Proposition 2.1].

tation problem: given a data matrix $D \in \mathbb{R}^{d\times n}$, solve

$$\begin{aligned} \min_{X,Y,Z} \quad & \sum_{i=1}^{\min(m,n)} r_1(\sigma_i(X)) + r_2(Y) + r_3(Z) \\ \text{subject to} \quad & D = A_1X + YA_2 + Z, \end{aligned} \tag{2}$$

where $X \in \mathbb{R}^{m\times n}, Y \in \mathbb{R}^{d\times q}, Z \in \mathbb{R}^{d\times n}$, $A_1 \in \mathbb{R}^{d\times m}$, $A_2 \in \mathbb{R}^{q\times n}$, $r_1(\cdot)$ is an increasing concave function to promote X to be of low rank, $r_2(\cdot)$ is regularization function, and $r_3(\cdot)$ is a function that models some noise (for example, if we take $r_3(Z) = \frac{1}{2}\|Z\|_F^2$ then Z represents a Gaussian noise). Problem (2) generalizes several important problems in machine learning. Let us mention some examples:

(i) When A_1 and A_2 are identity matrices, $r_1(t) = t^\chi$ with $0 < \chi \leq 1$, $r_2(Y) = \sum_{i=1}^{q-1} \|Y_i - Y_{i+1}\|$, where Y_i is the i-th column of Y, Problem (2) decomposes the data matrix D into three components, X, Y and Z. For example, in video surveillance, each column of D is a vectorized image of a video frame, X is a low-rank matrix that plays the role of the background, Y is the foreground that has small variations betweet its columns (such as slowly moving objectives), and Z represents some noise [42].

(ii) When A_1 and A_2 are identity matrices, $r_1(t) = t$, $r_2(Y) = \lambda\|Y\|_1$, where λ is some constant, Problem (2) recovers the robust principal component analysis model, see, e.g., [10], where X is a low-rank matrix, Y represents a sparse noise, and Z represents additional noise. It is also used for foreground-background separation in video surveillance.

(iii) When $r_1(t) = t$ and $r_2(Y) = \|Y\|_*$, Problem (2) is the latent low-rank representation problem [27]. The authors [27] used $A_1 = DP_1$ and $A_2 = P_2^*D$, where P_1 and P_2 are computed by orthogonalizing the columns of D^* and D, respectively. We will use this application to illustrate the effectiveness of our proposed framework, iADMM, in Section 3.

Other applications of Problem (1) include statistical learning, see, e.g., [3, 43], and minimization on compact manifolds, see, e.g., [22, 44].

1.2. Motivation and related works. Let

$$\mathcal{A} := [\mathcal{A}_1 \dots \mathcal{A}_s], \qquad \mathcal{A}x := \sum_{i=1}^{s} \mathcal{A}_i x_i \in \mathbb{R}^{\mathbf{m}}.$$

The augmented Lagrangian for Problem (1) is given by

$$\mathcal{L}(x, y, \omega) := F(x) + h(y) + \langle \omega, \mathcal{A}x + \mathcal{B}y - b\rangle + \frac{\beta}{2}\|\mathcal{A}x + \mathcal{B}y - b\|^2, \tag{3}$$

where $\beta > 0$ is a penalty parameter. ADMM was first introduced by [15] and [14]. It has recently become popular because of its efficacy in solving emerging large-scale problems in machine learning and computer vision [8, 39, 48, 49, 50]. For simplicity, let us describe the iteration scheme of a classical ADMM for solving Problem (1) with 2 blocks x and y:

$$x^{k+1} \in \operatorname*{argmin}_x \mathcal{L}(x, y^k, \omega^k), \tag{4a}$$

$$y^{k+1} \in \operatorname*{argmin}_y \mathcal{L}(x^{k+1}, y, \omega^k), \tag{4b}$$

$$\omega^{k+1} = \omega^k + \beta(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b). \tag{4c}$$

For a multi-block problem, with $s > 1$, the scheme is similar, see for example [42]. The update of x in (4a) (a similar discussion is applicable to (4b)) can be rewritten as

$$x^{k+1} \in \operatorname*{argmin}_x F(x) + \varphi^k(x),$$

where

$$\varphi^k(x) = \frac{\beta}{2}\|\mathcal{A}x + \mathcal{B}y^k - b\|^2 + \langle \omega^k, \mathcal{A}x + \mathcal{B}y^k - b\rangle. \tag{5}$$

Solving the subproblem (4a) is usually very expensive especially when F is not smooth. A remedy is minimizing a suitable surrogate function of $\mathcal{L}(\cdot, y^k, \omega^k)$ that allows a more efficient update for x. For example, since $\varphi^k(x)$ is upper bounded by

$$\hat{\varphi}(x) = \varphi^k(x^k) + \langle \nabla\varphi^k(x^k), x - x^k\rangle + \frac{\kappa\beta}{2}\|x - x^k\|^2 \tag{6}$$

where $\kappa \geq \|\mathcal{A}^*\mathcal{A}\|$ (because $\nabla\varphi^k(x)$ is $\beta\|\mathcal{A}^*\mathcal{A}\|$-Lipschitz continuous), x can be updated by

$$x^{k+1} \in \operatorname*{argmin}_x F(x) + \hat{\varphi}(x), \tag{7}$$

which leads to the linearized ADMM method, see [25, 45]. The update in (7) has a closed form for some nonsmooth F; see [34]. When $F = f + g$ and f is L_f-smooth then we can also use the upper bound $\hat{F}(x) = f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L_f}{2}\|x - x^k\|^2 + g(x)$ of F to derive the following update for x:

$$x^{k+1} \in \operatorname*{argmin}_x \hat{F}(x) + \hat{\varphi}(x). \tag{8}$$

This leads to the proximal linearized ADMM method, see [7, 28]. We note that $\mathcal{L}(\cdot, y^k, \omega^k)$ is always upper bounded by $\mathcal{L}(\cdot, y^k, \omega^k) + \mathbf{D}_\phi(x, x^k)$, where $\mathbf{D}_\phi$ is the Bregman distance associated with a continuously differentiable convex function ϕ on $\mathbb{R}^n$:

$$\mathbf{D}_\phi(a, b) := \phi(a) - \phi(b) - \langle \nabla\phi(b), a - b\rangle, \forall a, b \in \mathbb{R}^n. \tag{9}$$

For example, if $\phi(x) = \|x\|_{\mathcal{Q}}^2 = \langle x, \mathcal{Q}x\rangle$ then $\mathbf{D}_\phi(a, b) = \|a - b\|_{\mathcal{Q}}^2$. This upper bound leads to proximal ADMM, see [12, 23].

The above mentioned upper bound functions are specific examples of surrogate functions for $\mathcal{L}(\cdot, y^k, \omega^k)$ (see Definition 2.1) while each method of updating x corresponds to a majorization-minimization (MM) step. In the convex setting (that is, $f(x, y)$ is convex), [11] and [20] use the MM principle to unify and generalize the convergence analysis of many ADMM for multi-blocks problems (that is, $s > 1$). However, ADMM with the MM principle has not been studied for the nonconvex problem (1), to the best of our knowledge.

When the linear coupling constraint is absent, the block coordinate descent (BCD) method is a standard approach to solve (1). [37] proposed the block successive upper-bound minimization (BSUM) framework that employs the MM principle in each block update. By employing suitable surrogate functions in each block update, BSUM recovers the typical BCD methods, for example of [16, 19, 36, 40, 4, 6, 41]. In the non-convex setting, BCD methods with inertial terms have also been studied and

they have showed significant improvement in their practical performance, see for example [32, 46, 47, 35, 17]. Recently, the authors in [18] propose an inertial block MM framework for solving (1) without the linear coupling constraint. To the best of our knowledge, ADMM with inertial terms for the primal variables have not been studied for the nonconvex setting although they have been analysed for the convex setting; see [24, 33].

1.3. Contribution. In this paper, we propose iADMM, a framework of inertial alternating direction methods of multipliers, for solving the nonconvex nonsmooth problem (1). When no extrapolation is used, iADMM becomes a general ADMM framework that employs the minimization-majorization principle in each block update. For the first time in the *nonconvex* nonsmooth setting of Problem (1), we study ADMM and its inertial version combined with the MM principle when updating each block of variables. Moreover, our framework allows to use an over-relaxation parameter $\alpha \in (0, 2)$ to set $\alpha\beta$ as the constant stepsize for updating the dual variable ω. Note that $\alpha = 1$ is the standard choice in the nonconvex setting, see, e.g., [20, 23, 42]. In the convex setting, [14] showed that α can be chosen in $(0, 2)$. However, in the nonconvex setting, most of the works assume that $\alpha \in \big(0, \frac{1+\sqrt{5}}{2}\big)$, see, e.g., [13, 49]. Recently, [7] proposed proximal ADMM that use $\alpha \in (0, 2)$ for solving a special case of the nonconvex Problem (1) with $s = 1$ and $\mathcal{A} = -\mathcal{I}$.

Under mild assumptions, we analyse the subsequential convergence guarantee for the generated sequence of iADMM and ADMM in parallel. When $F(x) + h(y)$ satisfies the KŁ property and $\alpha = 1$, we prove the global convergence for the generated sequence. Finally, we apply the proposed framework to solve a class of Problem (2) and report its numerical results to illustrate the efficacy of iADMM.

2. An inertial ADMM framework. In this section, we describe the iADMM framework and prove its subsequential and global convergence. Throughout the paper, we make the following assumptions that are standard for studying Problem (1) and the convergence of ADMMs in the nonconvex setting, see for example [42, 7, 23].

ASSUMPTION 1. *(i)* $\sigma_{\mathcal{B}} := \lambda_{\min}(\mathcal{B}\mathcal{B}^*) > 0$.
(ii) $F(x) + h(y)$ *is lower bounded.*
(iii) h *is a* L_h*-smooth function (that is,* ∇h *is* L_h*-Lipschitz continuous with constant* L_h*).*

2.1. iADMM description. Let us first formally define a surrogate function. Some examples were given in the introduction.

DEFINITION 2.1 (Surrogate function). *Let* $\mathcal{X} \subseteq \mathbb{R}^{\mathbf{n}}$. *A function* $u : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is called a surrogate function of a function* f *on* $\mathcal{X}$ *if the following conditions are satisfied:*
(a) $u(z, z) = f(z)$ *for all* $z \in \mathcal{X}$,
(b) $u(x, z) \geq f(x)$ *for all* $x, z \in \mathcal{X}$.

As we are considering multi-block problems, we need the following definition of a block surrogate function, which is a generalization of Definition 2.1.

DEFINITION 2.2 (Block surrogate function). *Let Let* $\mathcal{X}_i \subseteq \mathbb{R}^{\mathbf{n}_i}$, $\mathcal{X} \subseteq \mathbb{R}^{\mathbf{n}}$. *A function* $u_i : \mathcal{X}_i \times \mathcal{X} \to \mathbb{R}$ *is called a block* i *surrogate function of* f *on* $\mathcal{X}$ *if the following conditions are satisfied:*
(a) $u_i(z_i, z) = f(z)$ *for all* $z \in \mathcal{X}$,

Algorithm 1 iADMM for solving Problem (1)

Choose $x^0 = x^{-1}$, $y^0 = y^{-1}$, ω^0. Let u_i, $i \in [s]$, be block i surrogate functions of $f(x)$ on $\mathbb{R}^{\mathbf{n}}$.

for $k = 0, \ldots$ **do**

Set $x^{k,0} = x^k$

for $i = 1, ..., s$ **do**

Compute $\bar{x}_i^k = x_i^k + \zeta_i^k(x_i^k - x_i^{k-1})$.

Update block x_i by

$$x_i^{k,i} \in \operatorname*{argmin}_{x_i} \Big\{ u_i(x_i, x^{k,i-1}) + g_i(x_i) + \langle \mathcal{A}_i^*\big(\omega^k + \beta(\mathcal{A}\bar{x}^{k,i-1} + \mathcal{B}y^k - b)\big), x_i\rangle + \frac{\kappa_i\beta}{2}\|x_i - \bar{x}_i^k\|^2 \Big\}, \tag{10}$$

where $\kappa_i \geq \|\mathcal{A}_i^*\mathcal{A}_i\|$, and
$\bar{x}^{k,i-1} = (x_1^{k+1}, \ldots, x_{i-1}^{k+1}, \bar{x}_i^k, x_{i+1}^k, \ldots, x_s^k)$.

Set $x_j^{k,i} = x_j^{k,i-1}$ for all $j \neq i$.

end for

Set $x^{k+1} = x^{k,s}$.

Compute $\hat{y}^k = y^k + \delta_k(y^k - y^{k-1})$.

Update y by

$$y^{k+1} \in \operatorname*{argmin}_{y} \Big\{ \langle \mathcal{B}^*\omega^k + \nabla h(\hat{y}^k), y\rangle + \frac{\beta}{2}\|\mathcal{A}x^{k+1} + \mathcal{B}y - b\|^2 + \frac{L_h}{2}\|y - \hat{y}^k\|^2 \Big\}. \tag{11}$$

Update ω by

$$\omega^{k+1} = \omega^k + \alpha\beta(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b). \tag{12}$$

end for

(b) $u_i(x_i, z) \geq f(x_i, z_{\neq i})$ for all $x_i \in \mathcal{X}_i$ and $z \in \mathcal{X}$, where

$$(x_i, z_{\neq i}) := (z_1, \ldots, z_{i-1}, x_i, z_{i+1}, \ldots, z_s).$$

The block approximation error is defined as

$$e_i(x_i, z) := u_i(x_i, z) - f(x_i, z_{\neq i}).$$

The inertial alternating direction method of multipliers (iADMM) framework is described in Algorithm 1. iADMM cyclically update the blocks $x_1, \ldots, x_s$ and y. Let us use $x^{k,i} = (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_s^k)$ and $x^{k+1} = x^{k,s}$, where k is the outer iteration index, and i the cyclic inner iteration index ($i \in [s]$). The update of block x_i in (10) (note that $x_i^{k+1} = x_i^{k,i}$) means that iADMM chooses a surrogate function for $x_i \mapsto \mathcal{L}(x_i, x_{\neq i}^{k,i}, y^k, \omega^k)$, which is formed by summing a surrogate function of $x_i \mapsto f(x_i, x_{\neq i}^{k,i}) + g_i(x_i)$ and a surrogate function of $x_i \mapsto \varphi^k(x_i, x_{\neq i}^{k,i})$ where $\varphi^k(x)$ is defined in (5), then apply extrapolation to the latter surrogate function[2]. To update block y, as $h(y)$ is L_h-smooth, we apply Nesterov type acceleration on h as in (11).

[2]It is important noting that it is possible to embed the general inertial term $\mathcal{G}_i^k$ to the surrogate of

Together with Assumption 1, we make the following standard assumption for u_i throughout the paper.

ASSUMPTION 2. *(i) The block surrogate function $u_i(x_i, z)$ is continuous.*
(ii) Given $z \in \mathbb{R}^{\mathbf{n}}$, for $i \in [s]$, there exists a function $x_i \mapsto \bar{e}_i(x_i, z)$ such that $\bar{e}_i(\cdot, z)$ is continuously differentiable at z_i, $\bar{e}_i(z_i, z) = 0$, $\nabla_{x_i}\bar{e}_i(z_i, z) = 0$, and the block approximation error $x_i \mapsto e_i(x_i, z)$ satisfies

$$e_i(x_i, z) \le \bar{e}_i(x_i, z) \quad \text{for all} \quad x_i. \tag{13}$$

The condition in Assumption 2 (ii) is satisfied when we simply choose $u_i(x_i, z) = f(x_i, z_{\neq i})$ (that is, $f(x_i, z_{\neq i})$ is a surrogate function of itself), or when $e_i(\cdot, z)$ is continuously differentiable at z_i and $\nabla_{x_i} e_i(z_i, z) = 0$, or when $e_i(x_i, z) \le c\|x_i - z_i\|^{1+\epsilon}$ for some $\epsilon > 0$ and $c > 0$; see [18, Lemma 3].

REMARK 1. *Before proceeding to the convergence analysis of iADMM, we make the following remark. As we target Nesterov-type acceleration in the update of y (note that h is assumed to be L_h-smooth), we analyse the update rule as in (11) for y. In case y is updated by*

$$y^{k+1} \in \operatorname*{argmin}_y \mathcal{L}(x^{k+1}, y, \omega^k),$$

iADMM still works and the convergence analysis would be simplified by using the same rationale to obtain subsequential as well as global convergence. We hence omit this case in our analysis.

2.2. Convergence analysis. Let us start by defining some additional notations and their convention that will be used later. Let $x^{k,i}$, y^k and ω^k be the iterates generated by iADMM. We denote $\Delta x_i^k = x_i^k - x_i^{k-1}$, $\Delta y^k = y^k - y^{k-1}$, $\Delta\omega^k = \omega^k - \omega^{k-1}$, $\alpha_1 = \frac{|1-\alpha|}{\alpha\sigma_{\mathcal{B}}(1-|1-\alpha|)}$, $\alpha_2 = \frac{3\alpha}{\sigma_{\mathcal{B}}(1-|1-\alpha|)^2}$ and $\mathcal{L}^k = \mathcal{L}(x^k, y^k, \omega^k)$. We let ν_i, $i \in [s]$, and ν_y be arbitrary constants in $(0, 1)$. We take the following convention in the notation that allows us to analyse iADMM and its non-inertial version in parallel:

- If $\zeta_i^k = 0$, that is, when we do not apply extrapolation in the update of x_i^k, we take $\zeta_i^k/\nu_i = 0$ and $\nu_i = 0$.
- If $\delta_k = 0$, that is, when we do not apply extrapolation in the update of y, we take $\delta_k/\nu_y = 0$ and $\nu_y = 0$.

Now we present our main convergence results. Their proofs can be found in the supplementary material.

As iADMM allows to use extrapolation in the update of x_i^k and y^k, the Lagrangian is not guaranteed to satisfy the sufficient descent property; in fact, it is not guaranteed to decrease at each iteration. Instead, it has the following nearly sufficiently decreasing property as stated in the following Propositions 2.3 and 2.4.

PROPOSITION 2.3. *(i) Considering the update in (10), in general when $x_i \mapsto u_i(x_i, z) + g_i(x_i)$ is nonconvex, we choose $\kappa_i > \|A_i^* A_i\|$. Denote $a_i^k = \beta\zeta_i^k(\kappa_i + \|\mathcal{A}_i^*\mathcal{A}_i\|)$. Then*

$$\mathcal{L}(x^{k,i}, y^k, \omega^k) + \eta_i\|\Delta x_i^{k+1}\|^2 \le \mathcal{L}(x^{k,i-1}, y^k, \omega^k) + \gamma_i^k\|\Delta x_i^k\|^2, \tag{14}$$

$x_i \mapsto \mathcal{L}(x_i, x_{\neq i}^{k,i}, y^k, \omega^k)$ as in [18]. This inertial term may also lead to the extrapolation for the block surrogate function of $f(x)$ or for both the two block surrogates. However, to simplify our analysis, we only consider here the effect of the inertial term for the block surrogate of $\varphi^k(x)$.

where

$$\begin{aligned}\eta_i &= \tfrac{(1-\nu_i)(\kappa_i-\|\mathcal{A}_i^*\mathcal{A}_i\|)\beta}{2},\\ \gamma_i^k &= \tfrac{(a_i^k)^2}{2\nu_i(\kappa_i-\|\mathcal{A}_i^*\mathcal{A}_i\|)\beta}.\end{aligned} \tag{15}$$

(ii) When $x_i \mapsto u_i(x_i,z)+g_i(x_i)$ is convex, we choose $\kappa_i = \|\mathcal{A}_i^*\mathcal{A}_i\|$ and Inequality (14) is satisfied with

$$\gamma_i^k = \frac{\beta\|\mathcal{A}_i^*\mathcal{A}_i\|(\zeta_i^k)^2}{2}, \quad \eta_i = \frac{\beta\|\mathcal{A}_i^*\mathcal{A}_i\|}{2}. \tag{16}$$

PROPOSITION 2.4. *Considering the update in (11), we have*

$$\mathcal{L}(x^{k+1},y^{k+1},\omega^k)+\eta_y\|\Delta y^{k+1}\|^2 \le \mathcal{L}(x^{k+1},y^k,\omega^k)+\gamma_y^k\|\Delta y^k\|^2,$$

where $\eta_y = \frac{(1-\nu_y)(\beta\|\mathcal{B}^*\mathcal{B}\|+L_h)}{2}$ and $\gamma_y^k = \frac{2L_h^2\delta_k^2}{\nu_y(\beta\|\mathcal{B}^*\mathcal{B}\|+L_h)}$ when $h(y)$ is nonconvex, and $\eta_y = \frac{L_h}{2}$ and $\gamma_y^k = \frac{L_h\delta_k^2}{2}$ when $h(y)$ is convex.

From Proposition 2.3 and Proposition 2.4, we obtain the following recursive inequality for $\{\mathcal{L}^k\}$ in Proposition 2.5 that serves as cornerstone to derive the bound for the extrapolation parameters ζ_i^k and δ_k in Proposition 2.6.

PROPOSITION 2.5. *We have*

$$\begin{aligned}\mathcal{L}^{k+1}+\eta_y\|\Delta y^{k+1}\|^2+\sum_{i=1}^s\eta_i\|\Delta x_i^{k+1}\|^2 \le \mathcal{L}^k+\sum_{i=1}^s\gamma_i^k\|\Delta x_i^k\|^2+\gamma_y^k\|\Delta y^k\|^2\\ +\frac{\alpha_1}{\beta}(\|B^*\Delta\omega^k\|^2-\|B^*\Delta\omega^{k+1}\|^2)+\frac{\alpha_2}{\beta}L_h^2\|\Delta y^{k+1}\|^2\\ +\frac{\alpha_2}{\beta}\big(\bar\delta_k L_h^2\|\Delta y^k\|^2+4L_h^2\delta_{k-1}^2\|\Delta y^{k-1}\|^2\big),\end{aligned} \tag{17}$$

where $\bar\delta_k = 2$ if $\delta_k = 0$ for all k and $4(1+\delta_k)^2$ otherwise.

Now we characterize the chosen parameters for Algorithm 1 in the following proposition.

PROPOSITION 2.6. *Let η_y, γ_y^k, η_i, γ_i^k, $i\in[s]$, and $\bar\delta_k$ be defined in Proposition 2.3 and Proposition 2.4. Denote*

$$\mu = \eta_y - \tfrac{\alpha_2 L_h^2}{\beta}. \tag{18}$$

For $k\ge 1$, suppose the parameters are chosen such that $\mu>0$, $\eta_i>0$, and the following conditions are satisfied for some constants $0<C_x,C_y<1$:

$$\begin{aligned}&\gamma_i^k \le C_x\eta_i, \quad \frac{4\alpha_2L_h^2\delta_{k-1}^2}{\beta}\le C_2\mu,\\ &\frac{\alpha_2L_h^2\bar\delta_k}{\beta}+\gamma_y^k \le C_1\mu,\end{aligned} \tag{19}$$

where $\begin{cases} C_1 = C_y \text{ and } C_2 = 0 & \text{if } \delta_k = 0\,\forall\, k,\\ 0<C_1<C_y \text{ and } C_2 = C_y - C_1 & \text{otherwise.}\end{cases}$

(i) For $K > 1$ we have

$$\begin{aligned}
&\mathcal{L}^{K+1} + \mu\|\Delta y^{K+1}\|^2 + \sum_{i=1}^{s} \eta_i \|\Delta x_i^{K+1}\|^2 + \frac{\alpha_1}{\beta}\|\mathcal{B}^*\Delta w^{K+1}\|^2 \\
&+(1-C_1)\mu\|\Delta y^K\|^2 + \sum_{k=1}^{K-1}\left[(1-C_y)\mu\|\Delta y^k\|^2 + (1-C_x)\sum_{i=1}^{s}\eta_i\|\Delta x_i^{k+1}\|^2\right] \\
&\leq \mathcal{L}^1 + \frac{\alpha_1}{\beta}\|\mathcal{B}^*\Delta\omega^1\|^2 + C_x\sum_{i=1}^{s}\eta_i\|\Delta x_i^1\|^2 + \mu\|\Delta y^1\|^2 + C_2\mu\|\Delta y^0\|^2.
\end{aligned} \tag{20}$$

(ii) If we use one of the following methods:

- *we choose $\delta_k = 0$ for all k, that is, there is no extrapolation in the update of y,*
- *we use extrapolation in the update of y and choose the parameters such that*

$$\begin{aligned}
&\beta \geq \frac{4L_h\alpha}{\sigma_{\mathcal{B}}(1-|1-\alpha|)}, \\
&\beta \geq \frac{6\alpha L_h^2}{\mu\sigma_{\mathcal{B}}(1-|1-\alpha|)}\max\left\{1, \frac{12\delta_k^2}{1-C_1}\right\}
\end{aligned} \tag{21}$$

then $\{\Delta y^k\}$, $\{\Delta x_i^k\}$ and $\{\Delta\omega^k\}$ converge to 0.

We will assume that Algorithm 1 generates a bounded sequence in our subsequential and global convergence results. Let us provide a sufficient condition that guarantees this boundedness assumption in the following proposition.

PROPOSITION 2.7. *If $b + Im(\mathcal{A}) \subseteq Im(\mathcal{B})$, $\lambda_{\min}(\mathcal{B}^*\mathcal{B}) > 0$ and $F(x) + h(y)$ is coercive over the feasible set $\{(x,y) : \mathcal{A}x + \mathcal{B}y = b\}$ then the sequences $\{x^k\}$, $\{y^k\}$ and $\{\omega^k\}$ generated by Algorithm 1 are bounded.*

It is important noting that the coercive condition of $F(x) + h(y)$ over the feasible set is weaker than the coercive condition of $F(x) + h(y)$ over $x \in \mathbb{R}^{\mathbf{n}}, y \in \mathbb{R}^{\mathbf{q}}$. Let us now present the subsequential convergence of the generated sequence.

THEOREM 2.8 (Subsequential convergence). *Suppose the parameters of Algorithm 1 are chosen such that the conditions in (19) of Proposition 2.6 are satisfied. If the generated sequence of Algorithm 1 is bounded, then every limit point of the generated sequence is a critical point of $\mathcal{L}$.*

To obtain a global convergence, we need the following Kurdyka-Łojasiewicz (KŁ) property for $F(x) + h(y)$.

DEFINITION 2.9. *A function $\phi(\cdot)$ is said to have the KŁ property at $\bar{\mathbf{x}} \in \operatorname{dom}\partial\,\phi$ if there exists $\varsigma \in (0, +\infty]$, a neighborhood U of $\bar{\mathbf{x}}$ and a concave function $\Upsilon : [0, \varsigma) \to \mathbb{R}_+$ that is continuously differentiable on $(0, \varsigma)$, continuous at 0, $\Upsilon(0) = 0$, and $\Upsilon'(t) > 0$ for all $t \in (0, \eta)$, such that for all $\mathbf{x} \in U \cap [\phi(\bar{\mathbf{x}}) < \phi(\mathbf{x}) < \phi(\bar{\mathbf{x}}) + \varsigma]$, we have*

$$\Upsilon'\left(\phi(\mathbf{x}) - \phi(\bar{\mathbf{x}})\right)\operatorname{dist}\left(0, \partial\phi(\mathbf{x})\right) \geq 1, \tag{22}$$

where $\operatorname{dist}(0, \partial\phi(\mathbf{x})) = \min\{\|\mathbf{z}\| : \mathbf{z} \in \partial\phi(\mathbf{x})\}$. If $\phi(\mathbf{x})$ has the KŁ property at each point of $\operatorname{dom}\partial\phi$ then ϕ is a KŁ function.

Many non-convex non-smooth functions in practical applications belong to the class of KŁ functions, for examples, real analytic functions, semi-algebraic functions, and locally strongly convex functions, see for example [5, 6].

THEOREM 2.10 (Global convergence). *Suppose we do not use extrapolation to update y (that is, $\delta_k = 0$ for all k) and we take $\alpha = 1$. Then the conditions in (19) become*

$$\gamma_i^k \le C_x \eta_i, \quad \frac{2\alpha_2 L_h^2}{\beta} \le C_y \mu, \tag{23}$$

for some constants $0 < C_x, C_y < 1$. Furthermore, we assume that (i) for any $x, z \in \mathbb{R}^{\mathbf{n}}$, $x_i \in \mathrm{dom}(g_i)$ we have

$$\begin{aligned} \partial_{x_i}\big(f(x) + g_i(x_i)\big) &= \partial_{x_i} f(x) + \partial_{x_i} g_i(x_i), \\ \partial_{x_i}\big(u_i(x_i, z) + g_i(x_i)\big) &= \partial_{x_i} u_i(x_i, z) + \partial_{x_i} g_i(x_i), \end{aligned} \tag{24}$$

and (ii) for any x, z in a bounded subset of $\mathbb{R}^{\mathbf{n}}$, if $\mathbf{s}_i \in \partial u_i(x_i, z)$, there exists $\xi_i \in \partial_{x_i} f(x)$ such that

$$\|\xi_i - \mathbf{s}_i\| \le L_i \|x - z\| \text{ for some constant } L_i. \tag{25}$$

If the generated sequence of Algorithm 1 is bounded and $F(x) + h(y)$ has the KŁ property, then the whole generated sequence of Algorithm 1 converges to a critical point of $\mathcal{L}$.

We refer the readers to [38, Corollary 10.9] for a sufficient condition for (24) (see supplementary material for more details). Some specific examples that satisfy (24) include: (i) $g_i = 0$, (ii) the functions $x_i \mapsto f(x)$ and $x_i \mapsto u_i(x_i, z)$ are strictly differentiable (see [38, Exercise 10.10]), (iii) the functions $x_i \mapsto f(x)$ and $x_i \mapsto u_i(x_i, z)$ are convex and the relative interior qualification conditions are satisfied: $\mathrm{ri}(\mathrm{dom}(f(\cdot, x_{\neq i})) \cap \mathrm{ri}(\mathrm{dom} g_i) \neq \emptyset$ and $\mathrm{ri}(\mathrm{dom}(g(\cdot, z)) \cap \mathrm{ri}(\mathrm{dom} g_i) \neq \emptyset$. We note that although the condition in (25) is necessary for our convergence proof, the Lipschitz constant L_i does not influence how to choose the parameters in our framework. We end this section by noting that a convergence rate for the generated sequence of iADMM can be derived using the same technique as in the proof of [1, Theorem 2]. Some examples of using the technique of [1, Theorem 2] to derive the convergence rate include [46, Theorem 2.9] and [17, Theorem 3]. Other than the convergence rate which appears to be the same in different papers using the technique in [1], determining the KŁ exponent, that is, the coefficient $\mathbf{a}$ when $\Upsilon(t) = ct^{1-\mathbf{a}}$, where c is a constant, is an active and challenging topic. The type of the convergence rate depends on the value of $\mathbf{a}$. Specifically, when $\mathbf{a} = 0$, the algorithm converges after a finite number of steps; when $\mathbf{a} \in (0, 1/2]$ it has linear convergence, and when $\mathbf{a} \in (1/2, 1]$ it has sublinear convergence. Determining the value of $\mathbf{a}$ is out of the scope of this paper.

3. Numerical results. In this section, we apply iADMM to solve a latent low-rank representation problem of the form of Problem (2); see Section 1.1. Specifically, we choose $r_1(t) = \lambda_1 t$, $r_3(Z) = \frac{1}{2}\|Z\|^2$ (hence Z represents a Gaussian noise), and consider a *nonconvex* regularization function for Y, $r_2(Y) = \lambda \sum_{i=1}^q \phi(\|Y_i\|_2)$, where Y_i is the i-th column of Y and $\phi(t) = 1 - \exp(-\theta t)$ [9]. In the upcoming experiments, we choose $A_1 = DP_1$ and $A_2 = P_2^* D$ as proposed in [27], where P_1 and P_2 are computed by orthogonalizing the columns of D^* and D, respectively.

Problem (2) in this case takes the form of (1) with B being the identity operator, b being the data set D, x_1 and x_2 being the matrices X and Y, y being the matrix Z, $f(X, Y) = \lambda_1 \|X\|_* + r_2(Y)$, $g_i = 0$ and $h(Z) = \frac{1}{2}\|Z\|^2$.

We choose the following block surrogate functions for f: $u_1(X, X^k, Y^k) = \lambda_1 \|X\|_* + r_2(Y^k)$, $u_2(Y, X^{k+1}, Y^k) = r_2(Y^k) + \sum_{i=1}^q \varsigma_i^k \|Y_i\|_2 + \lambda_1 \|X^{k+1}\|_*$, where

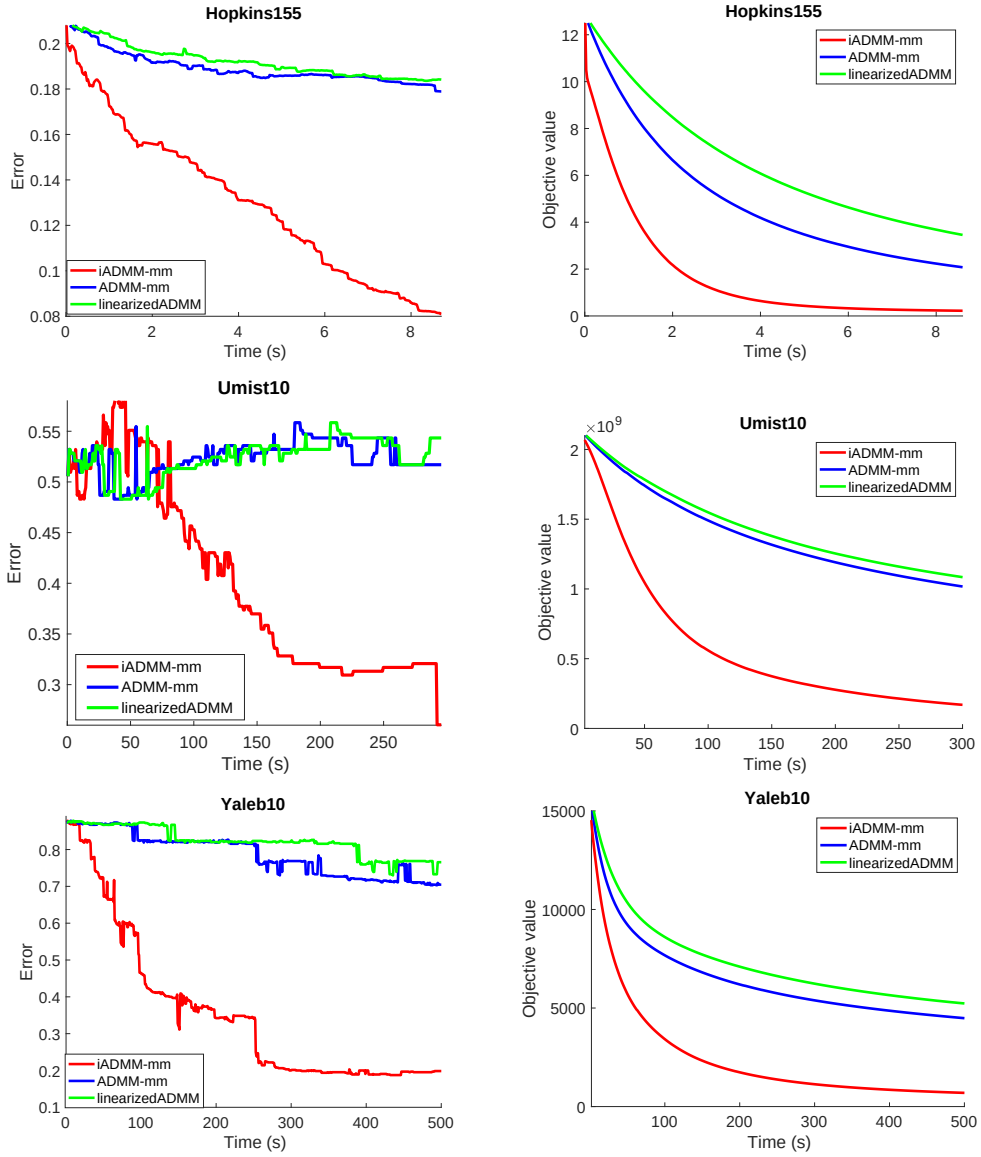

FIG. 1. *Evolution of the value of the segmentation error rate and the objective function value with respect to time. For Hopkins155, the results are the average values over 156 sequences.*

$\varsigma_i^k \in \lambda \nabla \phi(\|Y_i^k\|_2)$. Obviously u_1 satisfies Assumption 2 and u_2 satisfies Assumption 2 (i). Since ϕ is continuously differentiable with Lipschitz gradient on $[0, +\infty)$ and the Euclidean norm is Lipschitz continuous, it follows from Section 4.5 of [18] that u_2 satisfies Assumption 2 (ii).

For updating X, according to the update (10), X^{k+1} is computed by solving the

following nuclear norm problem

(26)
$$\min_X \lambda_1\|X\|_* + \left\langle A_1^*\Big(\beta(A_1\bar{X}^k + Y^kA_2 + Z^k - D) + W^k\Big), X\right\rangle + \frac{\kappa_1\beta}{2}\|X - \bar{X}^k\|^2,$$

where $\kappa_1 \geq \|A_1^*A_1\|$ and $\bar{X}^k = X^k + \zeta_1^k(X^k - X^{k-1})$. The sub-problem (26) has a closed-form solution given by

$$X^{k+1} = US_{\lambda_1/\kappa_1\beta}V^T,$$

where USV^T is the SVD of $\bar{X}^k - A_1^*(A_1\bar{X}^k + Y^kA_2 + Z^k - D + W^k)/\kappa_1\beta$ and $S_{\lambda_1/\kappa_1\beta} = \mathrm{diag}([S_{ii} - \lambda_1/\kappa_1\beta]_+)$, where $\mathrm{diag}(u)$ is a diagonal matrix whose diagonal elements are the entries of u, and $[.]_+$ is the projection onto the nonnegative orthant.

The update (11) for Y is

$$Y^{k+1} \in \arg\min_Y \sum_{i=1}^q \varsigma_i^k\|Y_i\|_2 + \langle (W^k + \beta(A_1X^{k+1} + \bar{Y}^kA_2 + Z^k - D))A_2^*, Y\rangle + \frac{\kappa_2\beta}{2}\|Y - \bar{Y}^k\|^2,$$

where $\kappa_2 \geq \|A_2A_2^*\|$ and $\bar{Y}^k = Y^k + \zeta_2^k(Y^k - Y^{k-1})$. The sub-problem above has a closed-form solution

$$Y_i^{k+1} = \big[\|P_i^k\| - \varsigma_i^k/(\kappa_2\beta)\big]_+ \frac{P_i^k}{\|P_i^k\|},$$

where P_i^k is the i-th column of $\bar{Y}^k - (A_1X^{k+1} + \bar{Y}^kA_2 + Z^k - D)/\kappa_2 - W^k/(\kappa_2\beta)$.

The updates (11) and (12) for Z and W are respectively given by

$$Z^{k+1} = -(W^k + \beta(A_1X^{k+1} + y^{k+1}A_2 - D))/(1+\beta),$$
$$W^{k+1} = W^k + \alpha\beta(A_1X^{k+1} + y^{k+1}A_2 + Z^{k+1} - D).$$

Let us determine the parameters. Note that $L_h = 1$, $\sigma_{\mathcal{B}} = 1$, and $\delta_k = 0$. Since $h(Z)$ is convex and we do not apply extrapolation for Z, hence by Proposition 2.4 we have $\eta_y = \frac{1}{2}$ and $\gamma_y^k = 0$. Since $\|X\|_*$ and $\sum_{i=1}^q \varsigma_i^k\|Y_i\|_2$ are convex, we choose $\kappa_1 = \|A_1^*A_1\|$, $\kappa_2 = \|A_2A_2^*\|$ and the conditions in (23) become $\zeta_i^k \leq \sqrt{C_x}$ (for $i = 1, 2$) and $\frac{(2+C_y)\alpha_2}{\beta} \leq \frac{C_y}{2}$. In our experiments, we choose $C_x = 1 - 10^{-15}$, $\alpha = 1$, $C_y = 1 - 10^{-6}$, $\beta = 2(2 + C_y)\alpha_2/C_y$, $a_0 = 1$, $a_k = \frac{1}{2}(1 + \sqrt{1 + 4a_{k-1}^2})$, and $\zeta_i^k = \min\left\{\frac{a_{k-1}-1}{a_k}, \sqrt{C_x}\right\}$.

We compare iADMM without extrapolation denoted by ADMM-mm, and iADMM with the extrapolation denoted by iADMM-mm, with a linearized ADMM that is only different from ADMM-mm for updating Y. In particular, the linearizedADMM method updates Y by solving the following nonconvex sub-problems

$$\min -\lambda\exp(\|Y_i\|_2) + \frac{\kappa_2\beta}{2}\|Y_i - V_i^k\|^2,$$

where V_i^k is the i-the column of $X^k - (W^k + \beta(A_1X^{k+1} + \bar{Y}^kA_2 + Z^k - D))A_2^*/(\kappa_2\beta)$. Since the sub-problems above do not have closed-form solutions, we employ an MM scheme to solve them.

TABLE 1
Comparison of segmentation error rate and final objective function values obtained within the allotted time. Bold values indicate the best results

	Method	Error mean ± std	Obj. value mean ± std
Hopkins	linearizedADMM	0.1579 ± 0.1550	3.0254 ± 2.4189
	ADMM-mm	0.1472 ± 0.1513	1.8081 ± 1.6674
	iADMM-mm	**0.0562** ± 0.1006	**0.2023** ± 0.1062
Umist	linearizedADMM	0.5170	1.0838×10^9
	ADMM-mm	0.5170	1.0167×10^9
	iADMM-mm	**0.2604**	$\mathbf{0.1694}\times10^9$
Yaleb	linearizedADMM	0.7656	5.2317×10^3
	ADMM-mm	0.7047	4.4829×10^3
	iADMM-mm	**0.1984**	$\mathbf{0.6951}\times10^3$

To examine the performance of the comparative algorithms, we consider subspace segmentation tasks. In particular, after obtaining X^*, we follow the setting in [26] to construct the affinity matrix Q by $Q_{ij} = (\tilde{U}\tilde{U}^T)_{ij}$, where $\tilde{U}$ is formed by $U^*(\Sigma^*)^{1/2}$ with normalized rows and $U^*\Sigma^*(V^*)^T$ being the SVD of X^*. Finally, we apply the Normalized Cuts [21] on W to cluster the data into groups.

The experiments are run on three data sets: Hopkins 155, extended Yale B and Umist. Hopskins 155 consists of 156 sequences, each of which has from 39 to 550 vectors drawn from two or three motions (one motion corresponds to one subspace). Each sequence is a sole segmentation task and thus there are 156 clustering tasks in total. Yale B contains 2414 frontal face images of 38 classes while Umist contains 564 images of 20 classes. To avoid computational issue when computing the segmentation error rate, we construct clustering tasks by using only the first 10 classes of these two data sets as proposed in [29].

All tests are preformed using Matlab R2019a on a PC 2.3 GHz Intel Core i5 of 8GB RAM. The code is available from https://github.com/nhatpd/iADMM

In our experiments, we choose $\theta = 5$, $\lambda_1 = \lambda = 0.01$ for Hopkins 155, and $\lambda_1 = \lambda = 1$ for the two other data sets. We note that we do not optimize numerical results by tweaking the parameters as this is beyond the scope of this work. It is important noting that we evaluate the algorithms on the same models. We set the initial points to zero. We run each algorithm 10, 300, and 500 seconds for each sequence of Hopkins 155, Umist10, and Yaleb10, respectively. We plot the curves of the value of the segmentation error rate and the objective function value versus the training time in Figure 1, and report the final values in Table 1. Since there are 156 sequences (data sets) in Hopkins 155, we plot the average values, and report the final average results and standard deviation over these sequences.

We observe that iADMM-mm converges the fastest on all the data sets, providing a significant acceleration of ADMM-mm. iADMM-mm achieves not only the best final objective function values but also the best segmentation error rates. This illustrates the usefulness of the acceleration technique. In addition, ADMM-mm outperforms linearizedADMM which illustrates the usefulness of properly choosing a proper surrogate function.

4. Conclusion. We have analysed iADMM, a framework of inertial alternating direction methods of multipliers, for solving a class of nonconvex nonsmooth optimization problem with linear constraints. The preliminary computational results in solving a class of nonconvex low-rank representation problems not only show the efficacy of using inertial terms for ADMM but also show the advantage of using suitable block surrogate functions that may lead to closed-form solutions in the block update of ADMM. We conclude the paper by mentioning two important questions that we consider as a future research directions:

- Can we extend the cyclic update rule of iADMM to randomized/non-cyclic setting?
- To guarantee the global convergence, iADMM does not allow extrapolation in the update of y; see Theorem 2.10. Can we extend the analysis to allow the extrapolation in the update of y?

REFERENCES

[1] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, Jan 2009.

[2] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4, 08 2011.

[4] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23:2037–2060, 2013.

[5] J. Bochnak, M. Coste, and M.-F. Roy. *Real Algebraic Geometry*. Springer, 1998.

[6] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, Aug 2014.

[7] R. I. Bot and D.-K. Nguyen. The proximal alternating direction method of multipliers in the nonconvex setting: Convergence analysis and rates. *Mathematics of Operations Research*, 45(2):682–712, 2020.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011.

[9] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceeding of international conference on machine learning ICML'98*, 1998.

[10] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3), 2011.

[11] L. Canyi, J. Feng, S. Yan, and Z. Lin. A unified alternating direction method of multipliers by majorization minimization. *IEEE transactions on pattern analysis and machine intelligence*, 40:527 – 541, 07 2018.

[12] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Rice CAAM tech report TR12-14*, 66, 01 2012.

[13] M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

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

[15] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

[16] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations Research Letters*, 26(3):127 – 136, 2000.

[17] L. T. K. Hien, N. Gillis, and P. Patrinos. Inertial block proximal method for non-convex non-smooth optimization. In *Thirty-seventh International Conference on Machine Learning*

ICML 2020, 2020.
[18] L. T. K. Hien, D. N. Phan, and N. Gillis. Inertial block majorization minimization framework for nonconvex nonsmooth optimization. arXiv:2010.12133, 2020.
[19] C. Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85, 1957.
[20] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo. A block successive upper-bound minimization method of multipliers for linearly constrained convex optimization. *Mathematics of Operations Research*, 45(3):833–861, 2020.
[21] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.
[22] R. Lai and S. Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58, 02 2014.
[23] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.
[24] H. Li and Z. Lin. Accelerated alternating direction method of multipliers: An optimal o(1 / k) nonergodic analysis. *Journal of Scientific Computing*, 79:671–699, 05 2019.
[25] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 612–620. Curran Associates, Inc., 2011.
[26] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):171–184, 2013.
[27] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. *2011 International Conference on Computer Vision*, pages 1615–1622, 2011.
[28] Q. Liu, X. Shen, and Y. Gu. Linearized admm for nonconvex nonsmooth optimization with convergence analysis. *IEEE Access*, 7:76131–76144, 2019.
[29] C. Lu, J. Tang, S. Yan, and Z. Lin. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25(2):829–839, 2016.
[30] J. G. Melo and R. D. C. Monteiro. Iteration-complexity of a jacobi-type non-euclidean admm for multi-block linearly constrained nonconvex programs, 2017.
[31] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publ., 2004.
[32] P. Ochs. Unifying abstract inexact convergence theorems and block coordinate variable metric ipiano. *SIAM Journal on Optimization*, 29(1):541–570, 2019.
[33] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015.
[34] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.
[35] T. Pock and S. Sabach. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787, 2016.
[36] M. J. D. Powell. On search directions for minimization algorithms. *Mathematical Programming*, 4(1):193–201, Dec 1973.
[37] M. Razaviyayn, M. Hong, and Z. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.
[38] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.
[39] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2101–2109. Curran Associates, Inc., 2010.
[40] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, Jun 2001.
[41] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, Mar 2009.
[42] Y. Wang, W. Yin, and J. Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78:29–63, 01 2019.
[43] Y. Wang, J. Zeng, Z. Peng, X. Chang, and Z. Xu. Linear convergence of adaptively iterative thresholding algorithms for compressed sensing. *IEEE Transactions on Signal Processing*, 63(11):2957–2971, 2015.
[44] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints.

Mathematical Programming, 142, 12 2010.
[45] M. Xu and T. Wu. A class of linearized proximal alternating direction methods. *J. Optimization Theory and Applications*, 151:321–337, 11 2011.
[46] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.
[47] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, Aug 2017.
[48] J. Yang, Y. Zhang, and W. Yin. An efficient tvl1 algorithm for deblurring multichannel images corrupted by impulsive noise. *SIAM Journal on Scientific Computing*, 31(4):2842–2865, 2009.
[49] L. Yang, T. K. Pong, and X. Chen. Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction. *SIAM Journal on Imaging Sciences*, 10(1):74–110, 2017.
[50] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for l(1)-minimization with applications to compressed sensing. *Siam Journal on Imaging Sciences*, 1:143–168, 01 2008.

Appendix A. Preliminaries of non-convex non-smooth optimization.

Let $g : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

DEFINITION A.1. *[38, Definition 8.3]*

(i) For any $x \in \operatorname{dom} g$, and $d \in \mathbb{E}$, we denote the directional derivative of g at x in the direction d by

$$g'(x;d) = \liminf_{\tau \downarrow 0} \frac{g(x+\tau d) - g(x)}{\tau}.$$

(ii) For each $x \in \operatorname{dom} g$, we denote $\hat{\partial} g(x)$ as the Frechet subdifferential of g at x which contains vectors $v \in \mathbb{E}$ satisfying

$$\liminf_{y \neq x, y \to x} \frac{1}{\|y-x\|} \left(g(y) - g(x) - \langle v, y-x \rangle\right) \geq 0.$$

If $x \notin \operatorname{dom} g$, then we set $\hat{\partial} g(x) = \emptyset$.

(iii) The limiting-subdifferential $\partial g(x)$ of g at $x \in \operatorname{dom} g$ is defined as follows:

$$\partial g(x) := \left\{ v \in \mathbb{E} : \exists x^{(k)} \to x,\, g\left(x^{(k)}\right) \to g(x),\, v^{(k)} \in \hat{\partial} g\left(x^{(k)}\right),\, v^{(k)} \to v \right\}.$$

(iv) The horizon subdifferential $\partial^{\infty} g(x)$ of g at x is defined as follows:

$$\begin{aligned} \partial^{\infty} g(x) := \Big\{ v \in \mathbb{E} : \exists \lambda^{(k)} \to 0, \lambda^{(k)} \geq 0, \lambda^{(k)} x^{(k)} \to x, \\ g\left(x^{(k)}\right) \to g(x),\, v^{(k)} \in \hat{\partial} g\left(x^{(k)}\right),\, v^{(k)} \to v \Big\}. \end{aligned}$$

DEFINITION A.2. *We call $x^* \in \operatorname{dom} \mathrm{F}$ a critical point of F if $0 \in \partial F(x^*)$.*

DEFINITION A.3. *[38, Definition 7.5] A function $f : \mathbb{R}^{\mathbf{n}} \to \mathbb{R} \cup \{+\infty\}$ is called subdifferentially regular at $\bar{x}$ if $f(\bar{x})$ is finite and the epigraph of f is Clarke regular at $(\bar{x}, f(\bar{x}))$ as a subset of $\mathbb{R}^{\mathbf{n}} \times \mathbb{R}$ (see [38, Definition 6.4] for the definition of Clarke regularity of a set at a point).*

PROPOSITION A.4. *[38, Corollary 10.9] Suppose $f = f_1 + \cdot + f_m$ for proper lower semi-continuous function $f_i : \mathbb{R}^{\mathbf{n}} \to \mathbb{R} \cup \{+\infty\}$ and let $\bar{x} \in \operatorname{dom} f$. Suppose each function f_i is subdifferential regular at $\bar{x}$, and the condition that the only combination of vector $\nu_i \in \partial^{\infty} f_i(\bar{x})$ with $\nu_1 + \ldots \nu_m = 0$ is $\nu_i = 0$ for $i \in [m]$. Then we have*

$$\partial f(\bar{x}) = \partial f_1(\bar{x}) + \ldots \partial f_m(\bar{x}).$$

Appendix B. Proofs. Before proving the propositions, let us give some preliminary results. We use x, z to denote the vectors in $\mathbb{R}^n$.

LEMMA B.1. *[18, Lemma 2.8] If the function $x_i \mapsto \Theta(x_i, z)$ is ρ-strongly convex, differentiable at z_i, and $\nabla_{x_i}\Theta(z_i, z) = 0$ then we have*

$$\Theta(x_i, z) \geq \frac{\rho}{2}\|x_i - z_i\|^2.$$

We recall the notation $(x_i, z_{\neq i}) = (z_1, \ldots, z_{i-1}, x_i, z_{i+1}, \ldots, z_s)$. Suppose we are trying to solve

$$\min_x \Psi(x) := \Phi(x) + \sum_{i=1}^{s} g_i(x_i).$$

PROPOSITION B.2. *[18, Theorem 2.7] Suppose $\mathcal{G}_i^k : \mathbb{R}^{\mathbf{n}_i} \times \mathbb{R}^{\mathbf{n}_i} \to \mathbb{R}^{\mathbf{n}_i}$ be some extrapolation operator that satisfies $\mathcal{G}_i^k(x_i^k, x_i^{k-1}) \leq a_i^k\|x_i^k - x_i^{k-1}\|$. Let $u_i(x_i, z)$ is a block surrogate function of $\Phi(x)$. We assume one of the following conditions holds:*

- *$x_i \mapsto u_i(x_i, z) + g_i(x_i)$ is ρ_i-strongly convex,*
- *the approximation error $\Theta(x_i, z) := u_i(x_i, z) - \Phi(x_i, z_{\neq i})$ satisfying $\Theta(x_i, z) \geq \frac{\rho_i}{2}\|x_i - z_i\|^2$ for all x_i.*

Note that ρ_i may depend on z. Let

$$x_i^{k+1} = \operatorname*{argmin}_{x_i} u_i(x_i, x^{k,i-1}) + g_i(x_i) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i\rangle.$$

Then we have

$$\Psi(x^{k,i-1}) + \gamma_i^k\|x_i^k - x_i^{k-1}\|^2 \geq \Psi(x^{k,i}) + \eta_i^k\|x_i^{k+1} - x_i^k\|^2, \tag{27}$$

where

$$\gamma_i^k = \tfrac{(a_i^k)^2}{2\nu\rho_i}, \qquad \eta_i^k = \tfrac{(1-\nu)\rho_i}{2},$$

and $0 < \nu < 1$ is a constant. If we do not apply extrapolation, that is $a_i^k = 0$, then (27) is satisfied with $\gamma_i^k = 0$ and $\eta_i^k = \rho_i/2$.

The following proposition is derived from [17, Remark 3] and [46, Lemma 2.1].

PROPOSITION B.3. *Suppose $x_i \mapsto \Phi(x)$ is a L_i-smooth convex function and $g_i(x_i)$ is convex. Define $\hat{x}_i^k = x_i^k + \alpha_i^k(x_i^k - x_i^{k-1})$, $\bar{x}_i^k = x_i^k + \beta_i^k(x_i^k - x_i^{k-1})$, and $\bar{x}^{k,i-1} = (x_1^{k+1}, \ldots, x_{i-1}^{k+1}, \bar{x}_i^k, x_{i+1}^k, \ldots, x_s^k)$. Let*

$$x_i^{k+1} = \operatorname*{argmin}_{x_i} \langle \nabla\Phi(\bar{x}^{k,i-1}), x_i\rangle + g_i(x_i) + \frac{L_i}{2}\|x_i - \hat{x}_i^k\|^2.$$

Then we have Inequality (27) is satisfied with

$$\gamma_i^k = \frac{L_i}{2}\big((\beta_i^k)^2 + \frac{(\gamma_i^k - \alpha_i^k)^2}{\nu}\big), \qquad \eta_i^k = \frac{(1-\nu)L_i}{2}.$$

If $\alpha_i^k = \beta_i^k$ then we have Inequality (27) is satisfied with

$$\gamma_i^k = \frac{L_i}{2}(\beta_i^k)^2, \qquad \eta_i^k = \frac{L_i}{2}.$$

B.1. Proof of Proposition 2.3. (i) Suppose we are updating x_i^k. Let us recall that $\mathcal{L}(x,y,\omega) := f(x) + \sum_{i=1}^s g_i(x_i) + h(y) + \varphi(x,y,\omega)$, where

$$\varphi(x,y,\omega) = \frac{\beta}{2}\|\mathcal{A}x + \mathcal{B}y - b\|^2 + \langle \omega, \mathcal{A}x + \mathcal{B}y - b\rangle. \tag{28}$$

Denote

$$\mathbf{u}_i(x_i,z,y,\omega) = u_i(x_i,z) + h(y) + \hat{\varphi}_i(x_i,z,y,\omega),$$

where

$$\hat{\varphi}_i(x_i,z,y,\omega) = \varphi(z,y,\omega) + \langle \mathcal{A}_i^*\big(\omega + \beta(\mathcal{A}z + \mathcal{B}y - b)\big), x_i - z_i\rangle + \frac{\kappa_i\beta}{2}\|x_i - z_i\|^2.$$

We see that $\hat{\varphi}_i(x_i,z,y,\omega)$ is a block surrogate function of $x \mapsto \varphi(x,y,\omega)$ with respect to block x_i, and $\mathbf{u}_i(x_i,z,y,\omega)$ is a block surrogate function of $x \mapsto f(x) + h(y) + \varphi(x,y,\omega)$ with respect to block x_i. The update in (10) can be rewritten as follows.

$$x_i^{k+1} = \operatorname*{argmin}_{x_i} \mathbf{u}_i(x_i, x^{k,i-1}, y^k, \omega^k) + g_i(x_i) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i\rangle, \tag{29}$$

where

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \beta\mathcal{A}_i^*\mathcal{A}\big(x^{k,i-1} - \bar{x}^{k,i-1}\big) + \kappa_i\beta\zeta_i^k(x_i^k - x_i^{k-1}). \tag{30}$$

The block approximation error function between $\mathbf{u}_i(x_i,z,y,\omega)$ and $x \mapsto f(x) + h(y) + \varphi(x,y,\omega)$ is defined as

$$\begin{aligned}
&\mathbf{e}_i(x_i,z,y,\omega) = \mathbf{u}_i(x_i,z,y,\omega) - \big(f(x_i,z_{\neq i}) + h(y) + \varphi((x_i,z_{\neq i}),y,\omega)\big)\\
&= u_i(x_i,z) - f(x_i,z_{\neq i}) + \hat{\varphi}_i(x_i,z,y,\omega) - \varphi((x_i,z_{\neq i}),y,\omega)\\
&\geq \theta_i(x_i,z,y,\omega) := \varphi(z,y,\omega) - \varphi((x_i,z_{\neq i}),y,\omega)\\
&+ \langle \mathcal{A}_i^*\big(\omega + \beta(\mathcal{A}z + \mathcal{B}y - b)\big), x_i - z_i\rangle + \frac{\kappa_i\beta}{2}\|x_i - z_i\|^2.
\end{aligned} \tag{31}$$

We have $\nabla_{x_i}\theta_i(x_i,z,y,\omega) = \kappa_i\beta(x_i - z_i) + \nabla_{x_i}\varphi(z,y,\omega) - \nabla_{x_i}\varphi((x_i,z_{\neq i}),y,\omega)$. Hence $\nabla_{x_i}\theta_i(z_i,z) = 0$. On the other hand, note that $x_i \mapsto \varphi((x_i,z_{\neq i}),y^k,\omega^k)$ is $\beta\|\mathcal{A}_i^*\mathcal{A}_i\|$ - smooth. So, $x_i \mapsto \theta_i(x_i,z,y,\omega)$ is a $\beta(\kappa_i - \|\mathcal{A}_i^*\mathcal{A}_i\|)$ - strongly convex function. From Lemma B.1 we have $\theta_i(x_i,z) \geq \frac{\beta(\kappa_i - \|\mathcal{A}_i^*\mathcal{A}_i\|)}{2}\|x_i - z_i\|^2$. The result follows from (29), (31) and Proposition (B.2).

(ii) When $x_i \mapsto u_i(x_i,z) + g_i(x_i)$ is convex and we apply the update as in (10), it follows from Proposition B.3 (see also [18, Remark 4.1]) that

$$\begin{aligned}
&u_i(x_i^k, x^{k,i-1}) + g_i(x_i^k) + \varphi(x^{k,i-1}, y^k, \omega^k) + \frac{\beta\|\mathcal{A}_i^*\mathcal{A}_i\|}{2}(\zeta_i^k)^2\|x_i^k - x_i^{k-1}\|^2\\
&\geq u_i(x_i^{k+1}, x^{k,i-1}) + g_i(x_i^{k+1}) + \varphi(x^{k,i}, y^k, \omega^k) + \frac{\beta\|\mathcal{A}_i^*\mathcal{A}_i\|}{2}\|x_i^{k+1} - x_i^k\|^2.
\end{aligned} \tag{32}$$

On the other hand, note that $u_i(x_i^k, x^{k,i-1}) = f(x^{k,i-1})$ and $u_i(x_i^{k+1}, x^{k,i-1}) \geq f(x^{k,i})$. The result follows then.

B.2. Proof of Proposition 2.4. Denote

$$\hat{h}(y,y') = h(y') + \langle \omega, \mathcal{A}x + \mathcal{B}y' - b\rangle + \langle \mathcal{B}^*\omega + \nabla h(y'), y - y'\rangle + \frac{L_h}{2}\|y - y'\|^2.$$

Then we have $\hat{h}(y,y') + \frac{\beta}{2}\|\mathcal{A}x+\mathcal{B}y-b\|^2$ is a surrogate function of $y \mapsto h(y)+\varphi(x,y,\omega)$. Note that the function $y \mapsto \hat{h}(y,y') + \frac{\beta}{2}\|\mathcal{A}x + \mathcal{B}y - b\|^2$ is $(L_h + \beta\|\mathcal{B}^*\mathcal{B}\|)$-strongly convex. The result follows from Proposition B.2 (see also [18, Section 4.2.1]).

Suppose $h(y)$ is convex. We note that $y \mapsto \frac{\beta}{2}\|\mathcal{A}x + \mathcal{B}y - b\|^2$ is also convex and plays the role of g_i in Proposition B.3. The result follows from Proposition B.3.

B.3. Proof of Proposition 2.5. Note that

$$\mathcal{L}(x^{k+1}, y^{k+1}, \omega^{k+1}) = \mathcal{L}(x^{k+1}, y^{k+1}, \omega^k) + \frac{1}{\alpha\beta}\langle \omega^{k+1} - \omega^k, \omega^{k+1} - \omega^k\rangle. \tag{33}$$

From the optimality condition of (11) we have

$$\nabla h(\hat{y}^k) + L_h(y^{k+1} - \hat{y}^k) + \mathcal{B}^*\omega^k + \beta\mathcal{B}^*(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b) = 0.$$

Together with (12) we obtain

$$\nabla h(\hat{y}^k) + L_h(\Delta y^{k+1} - \delta_k \Delta y^k) + \mathcal{B}^*\omega^k + \frac{1}{\alpha}\mathcal{B}^*(w^{k+1} - w^k) = 0. \tag{34}$$

Hence,

$$\mathcal{B}^* w^{k+1} = (1-\alpha)\mathcal{B}^*\omega^k - \alpha(\nabla h(\hat{y}^k) + L_h(\Delta y^{k+1} - \delta_k \Delta y^k)), \tag{35}$$

which implies that

$$\mathcal{B}^*\Delta w^{k+1} = (1-\alpha)\mathcal{B}^*\Delta w^k - \alpha\Delta z^{k+1}, \tag{36}$$

where $\Delta z^{k+1} = z^{k+1} - z^k$ and $z^{k+1} = \nabla h(\hat{y}^k) + L_h(\Delta y^{k+1} - \delta_k \Delta y^k)$. We now consider 2 cases.

Case 1: $0 < \alpha \leq 1$. From the convexity of $\|\cdot\|$ we have

$$\|\mathcal{B}^*\Delta w^{k+1}\|^2 \leq (1-\alpha)\|\mathcal{B}^*\Delta w^k\|^2 + \alpha\|\Delta z^{k+1}\|^2. \tag{37}$$

Case 2: $1 < \alpha < 2$. We rewrite (36) as

$$\mathcal{B}^*\Delta w^{k+1} = -(\alpha-1)\mathcal{B}^*\Delta w^k - \frac{\alpha}{2-\alpha}(2-\alpha)\Delta z^{k+1}.$$

Hence

$$\|\mathcal{B}^*\Delta w^{k+1}\|^2 \leq (\alpha-1)\|\mathcal{B}^*\Delta w^k\|^2 + \frac{\alpha^2}{(2-\alpha)}\|\Delta z^{k+1}\|^2. \tag{38}$$

Combine (37) and (38) we obtain

$$\|\mathcal{B}^*\Delta w^{k+1}\|^2 \leq |1-\alpha|\|\mathcal{B}^*\Delta w^k\|^2 + \frac{\alpha^2}{1-|1-\alpha|}\|\Delta z^{k+1}\|^2, \tag{39}$$

which implies

$$(1-|1-\alpha|)\|\mathcal{B}^*\Delta w^{k+1}\|^2 \leq |1-\alpha|(\|\mathcal{B}^*\Delta w^k\|^2 - \|\mathcal{B}^*\Delta w^{k+1}\|^2) + \frac{\alpha^2}{1-|1-\alpha|}\|\Delta z^{k+1}\|^2. \tag{40}$$

On the other hand, when we use extrapolation for the update of y we have

$$
\begin{aligned}
\|\Delta z^{k+1}\|^2 &= \|\nabla h(\hat{y}^k) - \nabla h(\hat{y}^{k-1}) + L_h(\Delta y^{k+1} - \delta_k \Delta y^k) - L_h(\Delta y^k - \delta_{k-1}\Delta y^{k-1})\|^2 \\
&\leq 3L_h^2\|\hat{y}^k - \hat{y}^{k-1}\|^2 + 3L_h^2\|\Delta y^{k+1}\|^2 + 3\|(1+\delta_k)L_h\Delta y^k - L_h\delta_{k-1}\Delta y^{k-1}\|^2 \\
&\leq 6L_h^2\left[(1+\delta_k)^2\|\Delta y^k\|^2 + \delta_{k-1}^2\|\Delta y^{k-1}\|^2\right] \\
&\quad + 3L_h^2\|\Delta y^{k+1}\|^2 + 6(1+\delta_k)^2L_h^2\|\Delta y^k\|^2 + 6L_h^2\delta_{k-1}^2\|\Delta y^{k-1}\|^2 \\
&= 3L_h^2\|\Delta y^{k+1}\|^2 + 12(1+\delta_k)^2L_h^2\|\Delta y^k\|^2 + 12L_h^2\delta_{k-1}^2\|\Delta y^{k-1}\|^2.
\end{aligned} \tag{41}
$$

If we do not use extrapolation for y then we have

$$
\begin{aligned}
\|\Delta z^{k+1}\|^2 &= \|\nabla h(y^k) - \nabla h(y^{k-1}) + L_h\Delta y^{k+1} - L_h\Delta y^k\|^2 \\
&\leq 3L_h^2\|\Delta y^k\|^2 + 3L_h^2\|\Delta y^{k+1}\|^2 + 3L_h^2\|\Delta y^k\|^2 \\
&= 6L_h^2\|\Delta y^k\|^2 + 3L_h^2\|\Delta y^{k+1}\|^2.
\end{aligned} \tag{42}
$$

Furthermore, note that $\sigma_{\mathcal{B}}\|\Delta w^{k+1}\|^2 \leq \|\mathcal{B}^*\Delta w^{k+1}\|^2$. Therefore, it follows from (40) that

$$
\begin{aligned}
\|\Delta w^{k+1}\|^2 &\leq \frac{|1-\alpha|}{\sigma_{\mathcal{B}}(1-|1-\alpha|)}(\|\mathcal{B}^*\Delta w^k\|^2 - \|\mathcal{B}^*\Delta w^{k+1}\|^2) \\
&+ \frac{\alpha^2 3L_h^2}{\sigma_{\mathcal{B}}(1-|1-\alpha|)^2}(\|\Delta y^{k+1}\|^2 + \bar{\delta}_k\|\Delta y^k\|^2 + 4\delta_{k-1}^2\|\Delta y^{k-1}\|^2).
\end{aligned} \tag{43}
$$

The result is obtained from (43), (33) and Proposition 2.3.

B.4. Proof of Proposition 2.6. (i) From Inequality (17) and the conditions in (19) we have

$$
\begin{aligned}
&\mathcal{L}^{k+1} + \mu\|\Delta y^{k+1}\|^2 + \sum_{i=1}^{s}\eta_i\|\Delta x_i^{k+1}\|^2 + \frac{\alpha_1}{\beta}\|\mathcal{B}^*\Delta w^{k+1}\|^2 \\
&\leq \mathcal{L}^k + C_1\mu\|\Delta y^k\|^2 + C_2\mu\|\Delta y^{k-1}\|^2 + C_x\sum_{i=1}^{s}\eta_i\|\Delta x_i^k\|^2 + \frac{\alpha_1}{\beta}\|\mathcal{B}^*\Delta w^k\|^2.
\end{aligned} \tag{44}
$$

By summing from $k=1$ to K Inequality (44) and noting that $C_1 + C_2 = C_y$ we obtain Inequality (20).

(ii) Let us prove $\{\Delta y^k\}$ and $\{\Delta x_i^k\}$ converge to 0.

Let us first prove the second situation, that is we use extrapolation for the update of y and Inequality (21) is satisfied. From (35) we have

$$
\alpha\mathcal{B}^*w^{k+1} = -(1-\alpha)\mathcal{B}^*\Delta\omega^{k+1} - \alpha z^{k+1},
$$

where $z^{k+1} = \nabla h(\hat{y}^k) + L_h(\Delta y^{k+1} - \delta_k\Delta y^k)$. Using the same technique that derives Inequality (39), we obtain the following

$$
\alpha\sigma_{\mathcal{B}}\|w^{k+1}\|^2 \leq \alpha\|\mathcal{B}^*w^{k+1}\|^2 \leq |1-\alpha|\|\mathcal{B}^*\Delta\omega^{k+1}\|^2 + \frac{\alpha^2}{1-|1-\alpha|}\|z^{k+1}\|^2. \tag{45}
$$

On the other hand, we have

$$
\mathcal{L}^k = F(x^k) + h(y^k) + \frac{\beta}{2}\|\mathcal{A}x^k + \mathcal{B}y^k - b + \frac{\omega^k}{\beta}\|^2 - \frac{1}{2\beta}\|\omega^k\|^2
$$

$$\geq F(x^k) + h(y^k) - \frac{1}{2\beta}\|\omega^k\|^2.$$

Together with (45) and

$$\begin{aligned}\|z^k\|^2 &= \|\nabla h(\hat{y}^{k-1}) - \nabla h(y^k) + \nabla h(y^k) + L_h(\Delta y^k - \delta_{k-1}\Delta y^{k-1})\|^2\\ &\leq 4\|\nabla h(\hat{y}^{k-1}) - \nabla h(y^k)\|^2 + 4\|\nabla h(y^k)\|^2 + 4L_h^2\|\Delta y^k\|^2 + 4L_h^2\delta_{k-1}^2\|\Delta y^{k-1}\|^2\\ &\leq 12L_h^2\|\Delta y^k\|^2 + 12\delta_{k-1}^2\|\Delta y^{k-1}\|^2 + 4\|\nabla h(y^k)\|^2.\end{aligned}$$

we obtain

$$\begin{aligned}\mathcal{L}^k &\geq F(x^k) + h(y^k) - \frac{1}{2\alpha\beta\sigma_{\mathcal{B}}}\big(|1-\alpha|\|B^*\Delta\omega^k\|^2 + \frac{\alpha^2}{1-|1-\alpha|}\|z^k\|^2\big)\\ &\geq F(x^k) + h(y^k) - \frac{|1-\alpha|}{2\alpha\beta\sigma_{\mathcal{B}}}\|B^*\Delta\omega^k\|^2\\ &\quad - \frac{\alpha}{2\beta\sigma_{\mathcal{B}}(1-|1-\alpha|)}\big(12L_h^2\|\Delta y^k\|^2 + 12\delta_{k-1}^2\|\Delta y^{k-1}\|^2 + 4\|\nabla h(y^k)\|^2\big)\end{aligned} \tag{46}$$

Since $h(y)$ is L_h-smooth, for all $y \in \mathbb{R}^q$ and $\alpha_L > 0$ we have (see [31])

$$h(y - \alpha_L\nabla f(y)) \leq h(y) - \alpha_L(1 - \frac{L_h\alpha_L}{2})\|\nabla h(y)\|^2.$$

Let us choose α_L such that $\alpha_L(1 - \frac{L_h\alpha_L}{2}) = \frac{4\alpha}{2\beta\sigma_{\mathcal{B}}(1-|1-\alpha|)}$. Note that this equation have, solution when $\beta \geq \frac{4L_h\alpha}{\sigma_{\mathcal{B}}(1-|1-\alpha|)}$. Then we have

$$h(y^k) - \frac{4\alpha}{2\beta\sigma_{\mathcal{B}}(1-|1-\alpha|)}\|\nabla h(y^k)\|^2 \geq h(y^k - \alpha_L\nabla f(y^k)).$$

Together with (46) we get

$$\begin{aligned}\mathcal{L}^k &\geq F(x^k) + h(y^k - \alpha_L\nabla f(y^k)) - \frac{|1-\alpha|}{2\alpha\beta\sigma_{\mathcal{B}}}\|B^*\Delta\omega^k\|^2\\ &\quad - \frac{\alpha}{2\beta\sigma_{\mathcal{B}}(1-|1-\alpha|)}(12L_h^2\|\Delta y^k\|^2 + 12\delta_{k-1}^2\|\Delta y^{k-1}\|^2).\end{aligned} \tag{47}$$

So from $\frac{\alpha_1}{\beta} \geq \frac{|1-\alpha|}{2\alpha\beta\sigma_{\mathcal{B}}}$, $\mu \geq \frac{\alpha 12L_h^2}{2\beta\sigma_{\mathcal{B}}(1-|1-\alpha|)}$, $(1-C_1)\mu \geq \frac{\alpha 12L_h^2 12\delta_k^2}{2\beta\sigma_{\mathcal{B}}(1-|1-\alpha|)}$ we have

$$\begin{aligned}&\mathcal{L}^{K+1} + \mu\|\Delta y^{K+1}\|^2 + \frac{\alpha_1}{\beta}\|B^*\Delta w^{K+1}\|^2 + (1-C_1)\mu\|\Delta y^K\|^2\\ &\geq F(x^{K+1}) + h(y^{K+1} - \alpha_L\nabla f(y^{K+1})).\end{aligned} \tag{48}$$

Hence $\mathcal{L}^{K+1} + \mu\|\Delta y^{K+1}\|^2 + \frac{\alpha_1}{\beta}\|B^*\Delta w^{K+1}\|^2 + (1-C_1)\mu\|\Delta y^K\|^2$ is lower bounded.

Furthermore, since η_i and μ are positive numbers we derive from Inequality (20) that $\sum_{k=1}^{\infty}\|\Delta y^k\|^2 < +\infty$ and $\sum_{k=1}^{\infty}\|\Delta x_i^k\|^2 < +\infty$. Therefore, $\{\Delta y^k\}$ and $\{\Delta x_i^k\}$ converge to 0.

Let us now consider the first situation when $\delta_k = 0$ for all k.

From Inequality (17) and the conditions in (19) we have

$$\begin{aligned}&\mathcal{L}^{k+1} + \mu\|\Delta y^{k+1}\|^2 + \sum_{i=1}^{s}\eta_i\|\Delta x_i^{k+1}\|^2 + \frac{\alpha_1}{\beta}\|B^*\Delta w^{k+1}\|^2\\ &\leq \mathcal{L}^k + C_y\mu\|\Delta y^k\|^2 + C_x\sum_{i=1}^{s}\eta_i\|\Delta x_i^k\|^2 + \frac{\alpha_1}{\beta}\|B^*\Delta w^k\|^2.\end{aligned} \tag{49}$$

By summing Inequality (49) from $k = 1$ to K we obtain

$$\begin{aligned}
&\mathcal{L}^{K+1} + C_y\mu\|\Delta y^{K+1}\|^2 + C_x\sum_{i=1}^{s}\eta_i\|\Delta x_i^{K+1}\|^2 + \frac{\alpha_1}{\beta}\|B^*\Delta w^{K+1}\|^2 \\
&\quad + \sum_{k=1}^{K}\left[(1-C_y)\mu\|\Delta y^{k+1}\|^2 + (1-C_x)\sum_{i=1}^{s}\eta_i\|\Delta x_i^{k+1}\|^2\right] \\
&\leq \mathcal{L}^1 + \frac{\alpha_1}{\beta}\|B^*\Delta\omega^1\|^2 + \sum_{i=1}^{s}\eta_i^0\|\Delta x_i^1\|^2 + C\mu\|\Delta y^1\|^2.
\end{aligned} \tag{50}$$

Denote the value of the right side of Inequality (49) by $\hat{\mathcal{L}}^k$. Note that $0 < C_x, C_y < 1$, then from (49) we have the sequence $\{\hat{\mathcal{L}}^k\}$ is non-increasing. It follows from [30, Lemma 2.9] that $\hat{\mathcal{L}}^k \geq \vartheta$ for all k, where ϑ is is the lower bound of $F(x^k) + h(y^k)$. For completeness, let us provide the proof in the following. We have

$$\begin{aligned}
\hat{\mathcal{L}}^k \geq \mathcal{L}^k &= F(x^k) + h(y^k) + \frac{\beta}{2}\|Ax^k + By^k - b\|^2 + \frac{1}{\alpha\beta}\langle\omega^k, \omega^k - \omega^{k-1}\rangle \\
&\geq \vartheta + \frac{1}{2\alpha\beta}(\|\omega^k\|^2 - \|\omega^{k-1}\|^2 + \|\Delta\omega^k\|^2) \\
&\geq \vartheta + \frac{1}{2\alpha\beta}(\|\omega^k\|^2 - \|\omega^{k-1}\|^2),
\end{aligned} \tag{51}$$

Assume that there exists k_0 such that $\hat{\mathcal{L}}^k < \vartheta$ for all $k \geq k_0$. As $\hat{\mathcal{L}}^k$ is non-increasing,

$$\sum_{k=1}^{K}(\hat{\mathcal{L}}^k - \vartheta) \leq \sum_{k=1}^{k_0}(\hat{\mathcal{L}}^k - \vartheta) + (K - k_0)(\hat{\mathcal{L}}^k - \vartheta).$$

Hence $\sum_{k=1}^{\infty}(\hat{\mathcal{L}}^k - \vartheta) = -\infty$. However, from (51) we have

$$\sum_{k=1}^{K}(\hat{\mathcal{L}}^k - \vartheta) \geq \sum_{k=1}^{K}\frac{1}{2\alpha\beta}\|\omega^k\|^2 - \frac{1}{2\alpha\beta}\|\omega^{k-1}\|^2 \geq \frac{1}{2\alpha\beta}(-\|\omega^0\|^2),$$

which gives a contradiction.

Since $\hat{\mathcal{L}}^K \geq \vartheta$ and η_i and μ are positive numbers we derive from Inequality (20) that $\sum_{k=1}^{\infty}\|\Delta y^k\|^2 < +\infty$ and $\sum_{k=1}^{\infty}\|\Delta x_i^k\|^2 < +\infty$. Therefore, $\{\Delta y^k\}$ and $\{\Delta x_i^k\}$ converge to 0.

Now we prove $\{\Delta\omega^k\}$ goes to 0. Since $\sum_{k=1}^{\infty}\|\Delta y^k\|^2 < +\infty$, we derive from (41) that $\sum_{k=1}^{\infty}\|\Delta z^k\|^2 < +\infty$. Summing up Equality (39) from $k = 1$ to K we have

$$\begin{aligned}
&(1 - |1-\alpha|)\sum_{k=1}^{K}\|\mathcal{B}^*\Delta\omega^k\|^2 + \|\mathcal{B}^*\Delta\omega^{K+1}\|^2 \\
&\leq \|\mathcal{B}^*\Delta\omega^1\|^2 + \frac{\alpha^2}{1-|1-\alpha|}\sum_{k=1}^{K}\|\Delta z^{k+1}\|^2,
\end{aligned}$$

which implies that $\sum_{k=1}^{\infty}\|\mathcal{B}^*\Delta\omega^k\|^2 < +\infty$. Hence, $\|\mathcal{B}^*\Delta\omega^k\|^2 \to 0$. Since $\sigma_{\mathcal{B}} > 0$ we have $\{\Delta\omega^k\}$ goes to 0.

B.5. Proof of Proposition 2.7. We remark that we use the idea in the proof of [42, Lemma 6] to prove the proposition. However, our proof is more complicated since in our framework $\alpha \in (0,2)$, the function h is linearized and we use extrapolation for y.

Note that as $\sigma_{\mathcal{B}} > 0$ we have $\mathcal{B}$ is a surjective. Together with the assumption $b + Im(\mathcal{A}) \subseteq Im(\mathcal{B})$ we have there exist $\bar{y}^k$ such that $\mathcal{A}x^k + \mathcal{B}\bar{y}^k - b = 0$.

Now we have

$$\begin{aligned}\mathcal{L}^k &= F(x^k) + h(y^k) + \frac{\beta}{2}\|\mathcal{A}x^k + \mathcal{B}y^k - b\|^2 + \langle \omega^k, \mathcal{A}x^k + \mathcal{B}y^k - b\rangle \\ &= F(x^k) + h(y^k) + \frac{\beta}{2}\|Ax^k + \mathcal{B}y^k - b\|^2 + \langle \mathcal{B}^*\omega^k, y^k - \bar{y}^k\rangle.\end{aligned} \tag{52}$$

From (34) we have

$$\begin{aligned}\langle \mathcal{B}^*\omega^k, y^k - \bar{y}^k\rangle &= \big\langle \nabla h(\hat{y}^k) + L_h(\Delta y^{k+1} - \delta_k \Delta y^k) + \frac{1}{\alpha}\mathcal{B}^*(w^{k+1} - w^k), \bar{y}^k - y^k\big\rangle \\ &\geq \langle \nabla h(y^k), \bar{y}^k - y^k\rangle - \big(\|\nabla h(y^k) - \nabla h(\hat{y}^k)\| + L_h\|\Delta y^{k+1}\| \\ &\quad + L_h\delta_k\|\Delta y^k\| + \frac{1}{\alpha}\|\mathcal{B}^*\Delta\omega^{k+1}\|\big)\|\bar{y}^k - y^k\|.\end{aligned}$$

Therefore, it follows from (52) and L_h-smooth property of h that

$$\begin{aligned}\mathcal{L}^k &\geq F(x^k) + h(\bar{y}^k) - \frac{L_h}{2}\|y^k - \bar{y}^k\|^2 \\ &\quad - \big(2L_h\delta_k\|\Delta y^k\| + L_h\|\Delta y^{k+1}\| + \frac{1}{\alpha}\|\mathcal{B}^*\Delta\omega^{k+1}\|\big)\|\bar{y}^k - y^k\|.\end{aligned} \tag{53}$$

On the other hand, we have

$$\begin{aligned}\|\bar{y}^k - y^k\|^2 &\leq \frac{1}{\lambda_{\min}(\mathcal{B}^*\mathcal{B})}\|\mathcal{B}(\bar{y}^k - y^k)\|^2 = \frac{1}{\lambda_{\min}(\mathcal{B}^*\mathcal{B})}\|\mathcal{A}x^k + \mathcal{B}y^k - b\|^2 \\ &= \frac{1}{\lambda_{\min}(\mathcal{B}^*\mathcal{B})}\big\|\frac{1}{\alpha\beta}\Delta\omega^k\big\|^2.\end{aligned} \tag{54}$$

We have proved in Proposition 2.6 that $\|\Delta\omega^k\|$, $\|\Delta x^k\|$ and $\|\Delta y^k\|$ converge to 0. Furthermore, from Proposition 2.6 we have $\mathcal{L}^k$ is upper bounded. Therefore, from (53), (54) and (20) we have $F(x^k) + h(\bar{y}^k)$ is upper bounded. So $\{x^k\}$ is bounded. Consequently, $\mathcal{A}x^k$ is bounded.

Furthermore, we have

$$\|y^k\|^2 \leq \frac{1}{\lambda_{\min}(\mathcal{B}^*\mathcal{B})}\|\mathcal{B}y^k\|^2 = \frac{1}{\lambda_{\min}(\mathcal{B}^*\mathcal{B})}\big\|\frac{1}{\alpha\beta}\Delta\omega^k - \mathcal{A}x^k - b\big\|^2.$$

Therefore, $\{y^k\}$ is bounded, which implies that $\|\nabla h(\hat{y}^k)\|$ is also bounded. Finally, from (34) and the assumption $\lambda_{\min}(\mathcal{B}\mathcal{B}^*) > 0$ we also have $\{\omega^k\}$ is bounded.

B.6. Proof of Theorem 2.8. Suppose $(x^{k_n}, y^{k_n}, \omega^{k_n})$ converges to (x^*, y^*, ω^*). Since Δx_i^k goes to 0, we have $x_i^{k_n+1}$ and $x_i^{k_n-1}$ also converge to x_i^* for all $i \in [s]$. From (29), for all x_i, we have

$$\begin{aligned}&\mathbf{u}_i(x_i^{k+1}, x^{k,i-1}, y^k, \omega^k) + g_i(x_i^{k+1}) \\ &\leq \mathbf{u}_i(x_i, x^{k,i-1}, y^k, \omega^k) + g_i(x_i) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i - x_i^{k+1}\rangle.\end{aligned} \tag{55}$$

Choosing $x_i = x_i^*$ and $k = k_n - 1$ in (55) and noting that $\mathbf{u}_i(x_i, z)$ is continuous by Assumption 2 (i), we have

$$\limsup_{n\to\infty} \mathbf{u}_i(x_i^*, x^*, y^*, \omega^*) + g_i(x_i^{k_n}) \le \mathbf{u}_i(x_i^*, x^*, y^*, \omega^*) + g_i(x_i^*).$$

On the other hand, as $g_i(x_i)$ is lower semi-continuous. Hence, $g_i(x_i^{k_n})$ converges to $g_i(x_i^*)$. Now we choose $k = k_n \to \infty$ in (55) for all x_i we obtain

$$\begin{aligned} L_0(x^*, y^*, \omega^*) + g_i(x_i^*) &\le \mathbf{u}_i(x_i, x^*, y^*, \omega^*) + g_i(x_i) \\ &= L_0(x_i, x_{\neq i}^*, y^*, \omega^*) + \mathbf{e}_i(x_i, x^*, y^*, \omega^*) + g_i(x_i), \end{aligned} \tag{56}$$

where $L_0(x, y, \omega) = f(x) + h(y) + \varphi(x, y, \omega)$ and $\mathbf{e}_i$ is the approximation error defined in (31). We have

$$\begin{aligned} \mathbf{e}_i(x_i, x^*, y^*, \omega^*) &= u_i(x_i, x^*) - f(x_i, x_{\neq i}^*) + \hat{\varphi}_i(x_i, x^*, y^*, \omega^*) - \varphi((x_i, x_{\neq i}^*), y^*, \omega^*) \\ &\le \bar{e}_i(x_i, x^*) + \hat{\varphi}_i(x_i, x^*, y^*, \omega^*) - \varphi((x_i, x_{\neq i}^*), y^*, \omega^*). \end{aligned}$$

Note that $\bar{e}_i(x_i^*, x^*) = 0$ by Assumption 2. From (56) we have x_i^* is a solution of

$$\min_{x_i} L(x_i, x_{\neq i}^*, y^*, \omega^*) + \bar{e}_i(x_i, x^*) + \hat{\varphi}_i(x_i, x^*, y^*, \omega^*) - \varphi((x_i, x_{\neq i}^*), y^*, \omega^*).$$

Writing the optimality condition for this problem we obtain $0 \in \partial_{x_i} \mathcal{L}(x^*, y^*, \omega^*)$. Totally similarly we can prove that $0 \in \partial_y \mathcal{L}(x^*, y^*, \omega^*)$. On the other hand, we have

$$\Delta\omega^k = \omega^k - \omega^{k-1} = \alpha\beta(\mathcal{A}x^k + \mathcal{B}y^k - b) \to 0.$$

Hence, $\partial_\omega \mathcal{L}(x^*, y^*, \omega^*) = \mathcal{A}x^* + \mathcal{B}y^* - b = 0$.

As we assume $\partial F(x) = \partial_{x_1} F(x) \times \ldots \times \partial_{x_s} F(x)$, we have

$$\begin{aligned} \partial\mathcal{L}(x, y, \omega) &= \partial F(x) + \nabla\Big(h(y) + \langle\omega, \mathcal{A}x + \mathcal{B}y - b\rangle + \frac{\beta}{2}\|\mathcal{A}x + \mathcal{B}y - b\|^2\Big) \\ &= \partial_{x_1}\mathcal{L}(x, y, \omega) \times \ldots \times \partial_{x_s}\mathcal{L}(x, y, \omega) \times \partial_y \mathcal{L}(x, y, \omega) \times \partial_\omega \mathcal{L}(x, y, \omega). \end{aligned}$$

So $0 \in \partial\mathcal{L}(x^*, y^*, \omega^*)$.

B.7. Proof of Theorem 2.10. Note that we assume the generated sequence of Algorithm 1 is bounded. The following analysis is considered in the bounded set that contains the generated sequence of Algorithm 1. We first prove some preliminary results.

(A) The optimality condition of (29) gives us

$$\begin{aligned} &\mathcal{G}_i^k(x_i^k - x_i^{k-1}) - \mathcal{A}_i^*\big(\omega^k + \beta(\mathcal{A}x^{k,i-1} + \mathcal{B}y^k - b)\big) - \kappa_i\beta(x_i^{k+1} - x_i^k) \\ &\in \partial_{x_i}\big(u_i(x_i^{k+1}, x^{k,i-1}) + g_i(x_i^{k+1})\big). \end{aligned} \tag{57}$$

As (24) holds, there exists $\mathbf{s}_i^{k+1} \in \partial u_i(x_i^{k+1}, x^{k,i-1})$ and $\mathbf{t}_i^{k+1} \in \partial g_i(x_i^{k+1})$ such that

$$\mathcal{G}_i^k(x_i^k - x_i^{k-1}) - \mathcal{A}_i^*\big(\omega^k + \beta(\mathcal{A}x^{k,i-1} + \mathcal{B}y^k - b)\big) - \kappa_i\beta(x_i^{k+1} - x_i^k) = \mathbf{s}_i^{k+1} + \mathbf{t}_i^{k+1}. \tag{58}$$

As (25) holds, there exists $\xi_i^{k+1} \in \partial_{x_i} f(x^{k+1})$ such that

$$\|\xi_i^{k+1} - \mathbf{s}_i^{k+1}\| \le L_i\|x^{k+1} - x^{k,i-1}\|. \tag{59}$$

Denote $\tau_i^{k+1} := \xi_i^{k+1} + \mathbf{t}_i^{k+1} \in \partial_{x_i} F(x^{k+1})$ (as (24) holds). Then, from (58) we have

$$\tau_i^{k+1} = \xi_i^{k+1} + \mathcal{G}_i^k(x_i^k - x_i^{k-1}) - \mathcal{A}_i^*\big(\omega^k + \beta(\mathcal{A}x^{k,i-1} + \mathcal{B}y^k - b)\big) - \kappa_i\beta(x_i^{k+1} - x_i^k) - \mathbf{s}_i^{k+1}. \tag{60}$$

On the other hand, we note that

$$\partial_{x_i}\mathcal{L}(x^{k+1}, y^{k+1}, \omega^{k+1}) = \partial_{x_i} F(x^{k+1}) + \mathcal{A}_i^*\big(\omega^{k+1} + \beta(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b)\big). \tag{61}$$

Let $d_i^{k+1} := \tau_i^{k+1} + \mathcal{A}_i^*\big(\omega^{k+1} + \beta(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b)\big) \in \partial_{x_i}\mathcal{L}(x^{k+1}, y^{k+1}, \omega^{k+1})$. From (60) we have

$$\begin{aligned}\|d_i^{k+1}\| = \Big\|&\xi_i^{k+1} + \mathcal{G}_i^k(x_i^k - x_i^{k-1}) - \mathcal{A}_i^*\big(\omega^k + \beta(\mathcal{A}x^{k,i-1} + \mathcal{B}y^k - b)\big)\\ &- \kappa_i\beta(x_i^{k+1} - x_i^k) - \mathbf{s}_i^{k+1} + \mathcal{A}_i^*\big(\omega^{k+1} + \beta(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b)\big)\Big\|\end{aligned} \tag{62}$$

Together with (59) we obtain

$$\begin{aligned}\|d_i^{k+1}\| \le a_i^k\|\Delta x_i^k\| + \beta\|\mathcal{A}_i^*A\|\|x^{k+1} - x^{k,i-1}\| + \beta\|\mathcal{A}_i^*\mathcal{B}\|\|\Delta y^{k+1}\| + \|\mathcal{A}_i^*\|\|\Delta\omega^{k+1}\|\\ + \kappa_i\beta\|\Delta x_i^{k+1}\| + L_i\|x^{k+1} - x^{k,i-1}\|.\end{aligned} \tag{63}$$

It follows from (11) that

$$\mathcal{B}^*\omega^k + \nabla h(\hat{y}^k) + \beta\mathcal{B}^*(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b) + L_h(y^{k+1} - \hat{y}^k) = 0.$$

Let $d_y^{k+1} := \nabla h(y^{k+1}) + \mathcal{B}^*\big(\omega^{k+1} + \beta(\mathcal{A}x^{k+1} + \mathcal{B}y^{k+1} - b)\big)$. We have

$$d_y^{k+1} \in \partial_y\mathcal{L}(x^{k+1}, y^{k+1}, \omega^{k+1})$$

and

$$\begin{aligned}\|d_y^{k+1}\| &= \|\nabla h(y^{k+1}) - \nabla h(\hat{y}^k) + \mathcal{B}^*(\omega^{k+1} - \omega^k) - L_h(y^{k+1} - \hat{y}^k)\|\\ &\le 2L_h\|y^{k+1} - \hat{y}^k\| + \|\mathcal{B}^*\|\|\Delta\omega^{k+1}\|\\ &\le 2L_h(\|\Delta y^{k+1}\| + \delta_k\|\Delta y^k\|) + \|\mathcal{B}^*\|\|\Delta\omega^{k+1}\|.\end{aligned}$$

Let $d_\omega^{k+1} := \mathcal{A}x^{k+1} + \mathcal{B}^{k+1} - b$. We have $d_\omega^{k+1} \in \partial_\omega\mathcal{L}(x^{k+1}, y^{k+1}, \omega^{k+1})$ and

$$d_\omega^{k+1} = (\omega^{k+1} - \omega^k)/(\alpha\beta) = \Delta\omega^{k+1}/(\alpha\beta).$$

(B) Let us now prove $F(x^{k_n})$ converges to $F(x^*)$. This implies $\mathcal{L}(x^{k_n}, y^{k_n}, \omega^{k_n})$ converges to $\mathcal{L}(x^*, y^*, \omega^*)$ since $\mathcal{L}$ is differentiable in y and ω. We have

$$F(x^{k_n}) = f(x^{k_n}) + \sum_{i=1}^s g_i(x_i^{k_n}) = u_s(x_s^{k_n}, x^{k_n}) + \sum_{i=1}^s g_i(x_i^{k_n}).$$

So $F(x^{k_n})$ converges to $u_s(x_i^*, x^*) + \sum_{i=1}^s g_i(x_i^*) = F(x^*)$.

We now proceed to prove the global convergence. Denote $\mathbf{z} = (x, y, \omega)$, $\tilde{\mathbf{z}} = (\tilde{x}, \tilde{y}, \tilde{\omega})$, and $\mathbf{z}^k = (x^k, y^k, \omega^k)$. We consider the following auxiliary function

$$\bar{\mathcal{L}}(\mathbf{z}, \tilde{\mathbf{z}}) = \mathcal{L}(x, y, \omega) + \sum_{i=1}^s \frac{\eta_i + C_x\eta_i}{2}\|x_i - \tilde{x}_i\|^2 + \frac{(1 + C_y)\mu}{2}\|y - \tilde{y}\|^2 + \frac{\alpha_1}{\beta}\|B^*(\omega - \tilde{\omega})\|^2.$$

The auxiliary sequence $\bar{\mathcal{L}}(\mathbf{z}^k, \mathbf{z}^{k-1})$ has the following properties.

1. **Sufficient decreasing property**. From (49) we have

$$\bar{\mathcal{L}}(\mathbf{z}^{k+1},\mathbf{z}^k)+\sum_{i=1}^{s}\frac{\eta_i-C_x\eta_i}{2}\big(\|x_i^{k+1}-x_i^k\|^2+\|x_i^k-x_i^{k-1}\|^2\big)$$
$$+\frac{(1-C_y)\mu}{2}\big(\|y^{k+1}-y^k\|^2+\|y^k-y^{k-1}\|^2\big)$$
$$\leq \bar{\mathcal{L}}(\mathbf{z}^k,\mathbf{z}^{k-1}).$$

2. **Boundedness of subgradient**. In the proof (A) above, we have proved that

$$\|d^{k+1}\|\leq a_1(\|x^{k+1}-x^k\|+\|x^k-x^{k-1}\|+\|y^{k+1}-y^k\|+\|\omega^{k+1}-\omega^k\|)$$

for some constant a_1 and $d^{k+1}\in\partial\mathcal{L}(\mathbf{z}^{k+1})$. On the other hand, as we use $\alpha=1$, from (36) we obtain

(64)
$$\begin{aligned}\sqrt{\sigma_{\mathcal{B}}}\|\omega^{k+1}-\omega^k\|&\leq\|B^*(\omega^{k+1}-\omega^k)\|\\&=\|\Delta z^{k+1}\|=\|\nabla h(y^k)-\nabla h(y^{k-1})+L_h(\Delta y^{k+1}-\Delta y^k)\|\\&\leq 2L_h\|y^k-y^{k-1}\|+L_h\|y^{k+1}-y^k\|.\end{aligned}$$

Hence,

$$\|d^{k+1}\|\leq a_2(\|x^{k+1}-x^k\|+\|x^k-x^{k-1}\|+\|y^{k+1}-y^k\|+\|y^k-y^{k-1}\|)$$

for some constant a_2. Note that

$$\partial\bar{\mathcal{L}}(\mathbf{z},\tilde{\mathbf{z}})=\partial\mathcal{L}(\mathbf{z},\tilde{\mathbf{z}})$$
$$+\partial\Big(\sum_{i=1}^{s}\frac{\eta_i+C_x\eta_i}{2}\|x_i-\tilde{x}_i\|^2+\frac{(1+C_y)\mu}{2}\|y-\tilde{y}\|^2+\frac{\alpha_1}{\beta}\|B^*(\omega-\tilde{\omega})\|^2\Big).$$

Hence, it is not difficult to show that

$$\|\mathbf{d}^{k+1}\|\leq a_3(\|x^{k+1}-x^k\|+\|x^k-x^{k-1}\|+\|y^{k+1}-y^k\|+\|y^k-y^{k-1}\|)$$

for some constant a_3 and $\mathbf{d}^{k+1}\in\partial\bar{\mathcal{L}}(\mathbf{z}^{k+1},\mathbf{z}^k)$.

3. **KL property**. Since $F(x)+h(y)$ has KL property, then $\bar{\mathcal{L}}(\mathbf{z},\tilde{\mathbf{z}})$ also has KŁ property.
4. **A continuity condition**. Suppose $\mathbf{z}^{k_n}$ converges to (x^*,y^*,ω^*). In the proof (B) above, we have proved that $\mathcal{L}(\mathbf{z}^{k_n})$ converges to $\mathcal{L}(x^*,y^*,\omega^*)$. Furthermore, from Proposition 2.6 we proved that $\|\mathbf{z}^{k+1}-\mathbf{z}^k\|$ goes to 0. Hence we have $\mathbf{z}^{k_n-1}$ converges to (x^*,y^*,ω^*). So, $\bar{\mathcal{L}}(\mathbf{z}^{k+1},\mathbf{z}^k)$ converges to $\bar{\mathcal{L}}(\mathbf{z}^*,\mathbf{z}^*)$.

Using the same technique as in [6, Theorem 1], see also [17, 32], we can prove that

$$\sum_{k=1}^{\infty}\big(\|x^{k+1}-x^k\|+\|x^k-x^{k-1}\|+\|y^{k+1}-y^k\|+\|y^k-y^{k-1}\|\big)<\infty.$$

which implies $\{(x^k,y^k)\}$ converges to (x^*,y^*). From (64) we obtain

$$\sum_{k=1}^{\infty}\|\omega^{k+1}-\omega^k\|\leq\sum_{k=1}^{\infty}\big(\|y^{k+1}-y^k\|+\|y^k-y^{k-1}\|\big)<\infty.$$

Hence, $\{\omega^k\}$ also converges to ω^*.