Journal of Global Optimization (2024) 89:927–948
https://doi.org/10.1007/s10898-024-01382-4

An inertial ADMM for a class of nonconvex composite optimization with nonlinear coupling constraints

Le Thi Khanh Hien[1] · **Dimitri Papadimitriou**[2]

Received: 21 December 2022 / Accepted: 27 February 2024 / Published online: 19 March 2024

Abstract
In this paper, we propose an inertial alternating direction method of multipliers for solving a class of non-convex multi-block optimization problems with *nonlinear coupling constraints*. Distinctive features of our proposed method, when compared with other alternating direction methods of multipliers for solving non-convex problems with nonlinear coupling constraints, include: (i) we apply the inertial technique to the update of primal variables and (ii) we apply a non-standard update rule for the multiplier by scaling the multiplier by a factor before moving along the ascent direction where a relaxation parameter is allowed. Subsequential convergence and global convergence are presented for the proposed algorithm.

Keywords Alternating direction methods of multipliers · Nonlinear coupling constraints · Logistic matrix factorization · Multiblock nonconvex optimization · Majorization minimization

1 Introduction

We consider the following composite problem with nonlinear coupling constraints

$$\begin{aligned} \underset{x=(x_1,\ldots,x_s)\in\mathbb{R}^n, y\in\mathbb{R}^m}{\text{minimize}} \quad & \Theta(x,y) := F(x) + \sum_{i=1}^{s} f_i(x_i) + G(y) \\ \text{subject to} \quad & h(x) + \mathcal{B}y = 0, \end{aligned} \tag{1}$$

where $h(x) = (h_1(x), \ldots, h_q(x))$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^q$, $h_i(x)$ for $i = 1, \ldots, q$, are continuously differentiable functions, $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are proper lower semi-continuous functions, $F : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, G is L_G-smooth (that is, ∇G is L_G-Lipschitz continuous), and $\mathcal{B}$ is a linear mapping from $\mathbb{R}^m$ to $\mathbb{R}^q$. The

✉ Le Thi Khanh Hien
khanhhiennt@gmail.com

Dimitri Papadimitriou
dimitrios.papadimitriou.ext@huawei.com

[1] Huawei Belgium Research Center, 3001 Leuven, Belgium

[2] FTNO, Huawei Belgium Research Center, 3001 Leuven, Belgium

following Assumption 1 is necessary for our convergence analysis. In Remark 2, we will also consider a situation when Assumption 1(i) can be removed. Note that Assumption 1(ii) is typical in the literature of ADMM for nonconvex problem.

Assumption 1 (i) $x_i \mapsto h_2(x) := \frac{1}{2}\|h(x)\|^2$ is $l_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_s)$-smooth, for $i = 1, \ldots, s$; that is, $\nabla_{x_i} h_2(x)$ is l_i-Lipschiz continuous.
(ii) $\sigma_{\mathcal{B}} := \lambda_{\min}(\mathcal{B}\mathcal{B}^\top) > 0$ ($\sigma_{\mathcal{B}}$ denotes the smallest eigenvalue of $\mathcal{B}\mathcal{B}^\top$) and $\Theta(x, y)$ is bounded from below by ν.

Let us provide some examples that satisfy Assumption 1(i).

- If $h(x) = \sum_{i=1}^s \mathcal{A}_i x_i$ (h is linear), then $x_i \mapsto \frac{1}{2}\|h(x_i, y_{\neq i})\|^2$ is $\|\mathcal{A}_i^\top \mathcal{A}_i\|$-smooth.
- If $h(x_1, x_2) = x_1 x_2$ (bi-linear function), then $x_1 \mapsto \frac{1}{2}\|h(x_1, x_2)\|^2$ is $\|x_2 x_2^\top\|$-smooth and $x_2 \mapsto \frac{1}{2}\|h(x_1, x_2)\|^2$ is $\|x_1^\top x_1\|$-smooth.
- If h is a multilinear mapping then h satisfies Assumption 1(i).

Alternating direction methods of multipliers (ADMM) for solving Problem (1) with *linear* coupling constraints (that is when $h(x)$ is an affine mapping) have been deeply studied in the literature, see e.g., [8, 13, 18, 37] and the references therein. However, ADMM with convergence guarantee[1] for solving nonconvex composite problem with *nonlinear* coupling constraints have only appeared in [7], [9] and [15]. The authors in [7] consider Problem (1) with a more general nonlinear coupling constraint, which replaces $\mathcal{B}y$ by a mapping $g(y)$ from $\mathbb{R}^m$ to $\mathbb{R}^q$, and propose a universal framework to study global convergence analysis of *Lagrangian sequences* (see [7, Section 3.3]). To establish the global convergence of the Lagrangian sequences, the notion of *information zone* was introduced in [7] and the boundedness of the multiplier sequence is the key assumption to fulfill the role of the information zone. The authors in [15] develop the idea of information zone of [7] to propose mADMM - a multiblock alternating direction method of multipliers - for solving the general problem with any $s \geq 1$. Although the convergence analysis presented in [15] does not use the property of the Lagrangian sequence proposed in [7] to establish the global convergence of its generated sequence, it still relies on the boundedness assumption of the multiplier sequence (as a consequence of using the information zone). To avoid this unrealistic assumption, the authors in [9] consider Problem (1) (instead of the general nonlinear coupling constraints as in [7]) with $s = 1$. They design a proximal linearized alternating direction method of multipliers that requires a backtracking procedure to generate the proximal parameters. The convergence analysis of [9] does not require the boundedness of the multiplier sequence but the backtracking procedure used in [9] relies on the boundedness assumption of the generated sequence (see, [9, Lemma 5.2]) to guarantee the boundedness of the proximal parameters; however, [9] does not present a sufficient condition to guarantee the boundedness of the generated sequence.

Utilizing inertial techniques has become a prevalent approach to numerically accelerate an algorithm, frequently resulting in improved convergence rates. Let us provide a very brief review for inertial techniques. As far as we know, [31] is the first paper using an inertial technique that adds an inertial force, also known as a "momentum", $\zeta^k(x^k - x^{k-1})$ (where x^k is the current iterate, x^{k-1} is the previous iterate, and ζ^k is an extrapolation parameter) to the gradient direction to accelerate the gradient descent method. Later, Nesterov proposed his well-known accelerated fast gradient methods in the series of works [24–27]. Since the

[1] We thank the reviewer for bringing to our attention the reference [36], which explores the application of ADMM for addressing a nonconvex optimization problem involving nonlinear coupling constraints. Nevertheless, it is noteworthy that the paper does not provide a convergence analysis.

appearance of Nesterov's accelerated gradient methods, inertial techniques have been widely applied in convex as well as non-convex problems to accelerate the convergence of first-order methods, see e.g., [13, 14, 16, 28–30, 38, 39, 42] and the references therein.

In this paper, we employ the inertial technique proposed in [16] to propose iADMMn - an inertial alternating direction method of multipliers for solving Problem (1). When $\mathcal{B} = -\mathcal{I}$, where $\mathcal{I}$ is an identity mapping, Problem (1) is equivalent to

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ F(x) + \sum_{i=1}^{s} f_i(x_i) + G(h(x)). \tag{2}$$

An example of (2) is the following logistic matrix factorization problem [19]

$$\min_{U,V} \sum_{i=1}^{m}\sum_{j=1}^{n}(1 + cy_{ij} - y_{ij})\log(1 + \exp(u_i v_j^\top)) - cy_{ij}u_i v_j^\top + \frac{\lambda_d}{2}\|U\|_F^2 + \frac{\lambda_t}{2}\|V\|_F^2, \tag{3}$$

where $Y = \mathbb{R}^{m\times n}$ is a given data set with each element $y_{ij} \in \{0, 1\}$, λ_d and λ_t are regularization parameters, and c is some given constant. Problem (3) has the form of Problem (2) with $x = (U, V)$, $F(x) = \frac{\lambda_d}{2}\|U\|_F^2 + \frac{\lambda_t}{2}\|V\|_F^2$, $f_i = 0$, $h(U, V) = UV$, and $G(W) = \sum_{i,j}(1 + cy_{ij} - y_{ij})\log(1 + \exp(W_{ij})) - cy_{ij}W_{ij}$. A few other examples of Problem (2) are the PDE-constrained inverse problem [33], the risk parity portfolio selection problem [20], the robust phase retrieval problem [10], the nonlinear regression problem [11, 12] (h represents the model to train, G is a loss function, and $F(x) + \sum_{i=1}^{s} f_i(x_i)$ is a regularizer), and the generative adversarial networks [22].

It is worth noting that inertial techniques have also been applied in [13] to accelerate the convergence of ADMM for solving Problem (1) with h being a linear mapping. We allow h to be nonlinear in this paper. On the other hand, iADMMn uses a non-standard update rule for the multiplier. More specifically, the multiplier is scaled by a factor before moving along the ascent direction, see (9). This update rule for the multipliers is inspired by recent papers [34, 40] which give new perspectives on the multiplier update of primal-dual methods in the non-convex setting. Specifically, classical primal-dual methods are interpreted as methods that alternatively update the primal variable by minimizing the primal problem (primal descent) and update the dual (the multiplier) by maximizing the dual problem (dual ascent), see [5, Chapter 7]. However, when the primal problem is highly non-convex and may not be done in closed form, the classic dual ascent step may lose its valid interpretation. Hence, it makes sense to consider some modifications for the dual update. For example, [34] proposes a scaled dual *descent* update and [40] imposes a discounting factor to the multiplier before moving along the ascent direction. The multiplier update (9) of our iADMMn is similar to the multiplier update proposed in [40] in the sense that it scales the multiplier by a factor τ_1 before moving along the ascent direction, but iADMMn also allows a relaxation parameter τ_2 in the step-size of updating the multiplier.

The paper is organized as follows. In the next section, we provide some preliminaries on block surrogate functions, inertial techniques, and the augmented Lagrangian function. In Sect. 3, we describe iADMMn and analyze its convergence properties. We conclude the paper in Sect. 5.

Notations We denote $[s] = \{1, \ldots, s\}$. For $y \in \mathbb{R}^s$, we use $y_{\neq i}$ to denote $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_s)$. For a given mapping $h : \mathbb{R}^n \to \mathbb{R}^s$, we use $\nabla h(x) \in \mathbb{R}^{s\times n}$ to denote the Jacobian of h at x, that is $\nabla h(x) = [\nabla h_1(x) \ldots \nabla h_s(x)]^\top$, and we use

$\nabla_{x_i} h(x) \in \mathbb{R}^{s \times n_i}$ to denote the partial Jacobian of h with respect to x_i, that is $\nabla_{x_i} h(x) = [\nabla_{x_i} h_1(x) \dots \nabla_{x_i} h_s(x)]^\top$. For a given sequence $\{y^k\}$, we denote $\Delta y^k = y^k - y^{k-1}$.

2 Preliminaries

We refer the readers to [13, Appendix 1] for some preliminaries of non-convex optimization such as the definition of limiting subdifferential and the definition of the KŁ property. In the following, we give some preliminaries of block surrogate functions, inertial technique, augmented Lagrangian function, and ε-stationary point.

2.1 Block surrogate and nearly sufficiently decrease property

In our upcoming analysis, we use the following definition for a surrogate function.

Definition 1 (Block surrogate function) Let $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$, for $i = 1, \dots s$, and $\mathcal{X} \subseteq \mathbb{R}^n$. A continuous function $u_i : \mathcal{X}_i \times \mathcal{X} \to \mathbb{R}$ is called a block x_i surrogate function of $\varphi : \mathbb{R}^n \to \mathbb{R}$ on $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_s$ if the following conditions are satisfied:

(a) $u_i(z_i, z) = \varphi(z)$ for all $z \in \mathcal{X}$,
(b) $u_i(x_i, z) \geq \varphi(x_i, z_{\neq i})$ for all $x_i \in \mathcal{X}_i$ and $z \in \mathcal{X}$, where $(x_i, z_{\neq i})$ denotes $(z_1, \dots, z_{i-1}, x_i, z_{i+1}, \dots, z_s)$,
(c) the block approximation error is defined as $e_i(x_i, z) := u_i(x_i, z) - \varphi(x_i, z_{\neq i})$, $x_i \mapsto e_i(x_i, z)$ is continuously differentiable and it satisfies $\nabla_{x_i} e_i(x_i, x) = 0$ for all $x \in \mathcal{X}$.

For example, when $\nabla_{x_i} \varphi(\cdot, z_{\neq i})$ is $L_i^{(z)}$-Lipschitz continuous (note that $L_i^{(z)}$ may depend on z), the Lipschitz gradient surrogate function [14, 21, 39] has the form $u_i(x_i, z) = \varphi(z) + \langle \nabla_i \varphi(z), x_i - z_i \rangle + \frac{\kappa_i L_i^{(z)}}{2} \|x_i - z_i\|^2$, where $\kappa_i \geq 1$. Finding $\arg\min_{x_i} u_i(x_i, z) + f_i(x_i)$, where z plays the role of the current iterate, would lead to the block proximal gradient step of block coordinate methods [4, 6, 32, 35] for solving $\min_x \varphi(x) + \sum_{i=1}^s f_i(x_i)$. More examples such as a quadratic surrogate, a DC programming surrogate, a saddle point surrogate, and a Jensen surrogate can be found in [13, 16, 21]. Considering the optimization problem $\min_x \varphi(x) + \sum_{i=1}^s f_i(x_i)$, convergence analysis of various block coordinate methods corresponding to different choices of the surrogates can be unified by studying the convergence analysis of the block majorization minimization algorithm, which updates block x_i by finding $\arg\min_{x_i} u_i(x_i, z) + f_i(x_i)$, given z being the current iterate [32]. Furthermore, a suitable surrogate may lead to a closed-form solution for the update of block x_i while alternatingly minimizing the objective function does not have a closed-form solution and requires an outer optimizer to solve the subproblem, see e.g., [16, Section 6.2].

The authors in [16] propose TITAN, an inertial block majorization minimization framework for solving non-convex composite optimization problems without coupling constraints. TITAN updates one block of variables at a time by minimizing an inertial block surrogate function that is formed by adding an inertial force to a block surrogate function of the objective. The key property to establish the convergence of TITAN is the following nearly sufficiently decreasing property (NSDP), which holds when an inertial block surrogate is minimized (that is when (4) is performed). We will use the NSDP for the convergence analysis of our proposed method.

Proposition 1 *[16, Theorem 3] Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a lower semi-continuous function and $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper lower semi-continuous functions. Denote $x^{k,0} = x^k$,*

$x^{k,i} = (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_s^k)$ *for* $i \in [s]$, $x^{k+1} = x^{k,s}$, *and* $\Psi(x) := \Phi(x) + \sum_{i=1}^s f_i(x_i)$. *Suppose* $\mathcal{G}_i^k : \mathbb{R}^{n_i} \times \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ *be some extrapolation operator that satisfies* $\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \le a_i^k \|x_i^k - x_i^{k-1}\|$ *for some* a_i^k. *Let* $u_i(x_i, z)$ *be a block* x_i *surrogate function of* $\Phi(x)$ *and*

$$x_i^{k+1} = \operatorname*{argmin}_{x_i} u_i(x_i, x^{k,i-1}) + f_i(x_i) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle. \tag{4}$$

If one of the following conditions holds:

(i) $x_i \mapsto u_i(x_i, z) + f_i(x_i)$ *is* ρ_i*-strongly convex,*
(ii) the approximation error $e_i(x_i, z) := u_i(x_i, z) - \Phi(x_i, z_{\ne i})$ *satisfying*

$$e_i(x_i, z) \ge \frac{\rho_i(z)}{2}\|x_i - z_i\|^2 \text{for all } x_i,$$

(note that $\rho_i(z)$ *may depend on* z*), then we have the following NSDP*

$$\Psi(x^{k,i-1}) + \gamma_i^k \|x_i^k - x_i^{k-1}\|^2 \ge \Psi(x^{k,i}) + \eta_i^k \|x_i^{k+1} - x_i^k\|^2, \tag{5}$$

where

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

$\rho_i^k = \rho_i(x^{k,i-1})$, *and* $\nu \in (0,1)$ *is a constant. If we do not apply extrapolation, that is* $a_i^k = 0$, *then (5) is satisfied with* $\gamma_i^k = 0$ *and* $\eta_i^k = \rho_i^k/2$.

The following proposition, which is derived from [14, Remark 3] and [38, Lemma 2.1], provides an NSDP when Φ is multi-convex, Lipschitz gradient suggorates are used for Φ, and $f_i, i \in [s]$, are also convex. We then obtain better values of γ_i^k for the NSDP and larger extrapolation parameters could be used.

Proposition 2 *Let* Φ, f_i *and* Ψ *be defined as in Proposition 1. For* $i \in [s]$, *suppose* $x_i \mapsto \Phi(x)$ *is an* $L_i(x_{\ne i})$*-smooth convex function (the Lipschitz constant may depend on* $x_{\ne i}$*, the values of the other blocks) and* $f_i(x_i)$ *is convex. Define* $\hat{x}_i^k = x_i^k + \alpha_i^k(x_i^k - x_i^{k-1})$, $\bar{x}_i^k = x_i^k + \zeta_i^k(x_i^k - x_i^{k-1})$, *and* $\bar{x}^{k,i-1} = (x_1^{k+1}, \ldots, x_{i-1}^{k+1}, \bar{x}_i^k, x_{i+1}^k, \ldots, x_s^k)$, *where* α_i^k *and* ζ_i^k *are extrapolation parameters. Suppose*

$$x_i^{k+1} = \operatorname*{argmin}_{x_i} \langle \nabla_i \Phi(\bar{x}^{k,i-1}), x_i \rangle + f_i(x_i) + \frac{L_i^k}{2}\|x_i - \hat{x}_i^k\|^2, \tag{6}$$

where $L_i^k = L_i(x_{\ne i}^{k,i-1})$. *Note that (6) is equivalent to (4) with*

$$u_i(x_i, x^{k,i-1}) = \Phi(x^{k,i-1}) + \langle \nabla_i \Phi(x^{k,i-1}), x_i - x_i^k \rangle + \frac{L_i^k}{2}\|x_i - x_i^k\|^2,$$

and $\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \nabla_i \Phi(\bar{x}^{k,i-1}) - \nabla_i \Phi(x^{k,i-1}) + L_i^k(x_i^k - \hat{x}_i^k)$. *Then we have*

$$\Phi(x^{k,i-1}) + f_i(x_i^k) + \gamma_i^k \|x_i^k - x_i^{k-1}\|^2 \ge \Phi(x^{k,i}) + f_i(x_i^{k+1}) + \eta_i^k \|x_i^{k+1} - x_i^k\|^2,$$

where

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

It implies that the NSDP (5) is also satisfied.

If $\alpha_i^k = \zeta_i^k$ then we have Inequality (5) is satisfied with

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

2.2 Augmented Lagrangian and stationary point

Considering Problem (1), its augmented Lagrangian is

$$\mathcal{L}_\beta(x, y, \omega) = \Theta(x, y) + \langle h(x) + \mathcal{B}y, \omega\rangle + \frac{\beta}{2}\|h(x) + \mathcal{B}y\|^2. \tag{7}$$

Definition 2 We call (x^*, y^*) a stationary point of Problem (1) if there exists ω^* such that the following conditions are satisfied.

$$0 \in \partial_{x_i}(f_i(x_i^*) + F(x^*)) + \nabla_{x_i} h(x^*)^\top \omega^*, \ \nabla G(y^*) + \mathcal{B}^\top \omega^* = 0, \ h(x^*) + \mathcal{B}y^* = 0. \tag{8}$$

We know that finding a stationary point of (1) is equivalent to finding a critical point of the augmented Lagrangian $\mathcal{L}_\beta$, see [15, Section 2.2]. We are also interested in an ε-stationary point of (1), which is defined as follows.

Definition 3 We call (x^*, y^*) an ε-stationary point of Problem (1) if there exists ω^* and $\chi_i \in \partial_{x_i}(f_i(x_i^*) + F(x^*))$ such that

$$\|\chi_i + \nabla_{x_i} h(x^*)^\top \omega^*\| \le \varepsilon, \|\nabla G(y^*) + \mathcal{B}^\top \omega^*\| \le \varepsilon, \|h(x^*) + \mathcal{B}y^*\| \le \varepsilon.$$

3 Algorithm description and convergence analysis

Before describing iADMMn and presenting its convergence analysis, let us have a discussion on the proof methodology for a convergence guarantee of an alternating direction method of multipliers for solving a *non-convex* optimization problem. An informal general proof recipe that describes the main steps to prove the global convergence of a first-order method to a critical point of the objective function of a *non-convex* optimization problem without coupling constraints was introduced in [3, 6]. The three main steps of the recipe are (i) sufficient decrease property of the objective sequence, (ii) a subgradient lower bound for the iterates gap, and (iii) using the KŁ property, see [6, Section 3.2] for more details. This proof methodology has been used wisely to prove the global convergence of the proximal point algorithm, the forward-backward splitting algorithm, the regularized Gauss-Seidel method [2, 3], the proximal alternating linearized minimization algorithm [6], and the multi-block Bregman proximal alternating linearized minimization algorithm [1]. By replacing the sufficient decrease property of the objective sequence in Step (i) with the sufficient decrease property of the augmented Lagrangian sequence or the regularized augmented Lagrangian sequence, the proof methodology of [3, 6] is developed to prove the global convergence of ADMM for solving non-convex problems with linear/nonlinear coupling constraints, see [37, Section 3], [7, Section 3.3], [8, Theorem 1], [15, Theorem 2], [41, Section 3]. The augmented Lagrangian function or its regularized/auxiliary form plays the role of a Lyapunov function (sometimes it is called a potential function) of an original Lyapunov methodology. The sufficient decrease property of the Lyapunov function, the subgradient lower bound for the iterate gaps, together with the KŁ property of the Lyapunov function are sufficient to prove the

global convergence of the generated sequence to a critical point of the Lyapunov function. This general proof recipe is also applied when an inertial technique is used to accelerate the convergence, see [13, 14, 16, 17, 29, 38]. Choosing the appropriate Lyapunov function and establishing its sufficient decrease property is the cornerstone of proving global convergence. In this paper, we will also use this general proof recipe to prove the global convergence of iADMMn. To that end, in the following, we sequentially describe the update of x_i, y, and ω, and establish the necessary NSDPs. These NSDPs will be used to prove the sufficient decrease property of an appropriate Lyapunov function that is defined later in (40).

3.1 Algorithm description and the NSDPs

Let us remind the notations

$$x^{k,0} = x^k,\ \ x^{k,i} = (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_s^k) \ \text{ for } i \in [s], \ \text{ and } \ x^{k+1} = x^{k,s}.$$

Note that the augmented Lagrangian in (7) of Problem (1) can be rewritten as

$$\mathcal{L}_\beta(x, y, \omega) = \sum_{i=1}^{s} f_i(x_i) + \varphi(x, y, \omega),$$

where

$$\begin{aligned}\varphi(x, y, \omega) &= F(x) + G(y) + \langle h(x) + \mathcal{B}y, \omega\rangle + \frac{\beta}{2}\|h(x) + \mathcal{B}y\|^2 \\ &= F(x) + G(y) + \langle h(x) + \mathcal{B}y, w\rangle + \frac{\beta}{2}\|h(x)\|^2 + \frac{\beta}{2}\|\mathcal{B}y\|^2 + \beta\langle h(x), \mathcal{B}y\rangle.\end{aligned}$$

We summarize the description of iADMMn in Algorithm 1. The detailed update rules of x_i and y are elaborated in the following.

Algorithm 1: iADMMn for solving Problem (1)

Suppose $\hat{u}_i$ is a block surrogate function of F with respect to block x_i.
Parameters of the algorithm are chosen as in (28).
Choose initial points x^0, y^0, ω^0.
for $k = 0, \ldots$ **do**
  for $i = 1, \ldots, s$ **do**
    Compute the extrapolation point $\bar{x}_i^k = x_i^k + \alpha_i^k(x_i^k - x_i^{k-1})$.
    Under Assumption 1(i), we update x_i as in (14).
    If together with Assumption 1(i) we also have $x_i \mapsto F(x_i, y_{\neq i})$ is $L_i(y_{\neq i})$-smooth then we update x_i as in (20), and if Assumption 1(i) is not satisfied but $x_i \mapsto F(x_i, y_{\neq i})$ is $L_i(y_{\neq i})$-smooth then we update x_i as in (24).
  end for
  Update y as in (26).
  Update ω as

$$\omega^{k+1} = \tau_1\omega^k + \tau_2\beta(h(x^{k+1}) + \mathcal{B}y^{k+1}). \tag{9}$$

end for

Update rule for block x_i and NSDP when updating x_i Following the inertial technique proposed in [16], we conduct three steps to update block x_i: (1) form a block x_i surrogate function of $\varphi(x, y, \omega)$, (2) add an inertial term $\langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i\rangle$ to this surrogate function,

and (3) update x_i by minimizing the sum of $f_i(x_i)$ and the inertial surrogate function. To form a block x_i surrogate function of $\varphi(x, y, \omega)$, we note that if $u_i(x_i, z)$ is a block x_i surrogate function of $F(x) + \frac{\beta}{2}\|h(x)\|^2$ then

$$u_i(x_i, z) + G(y) + \langle h(x_i, z_{\neq i}) + \mathcal{B}y, w\rangle + \frac{\beta}{2}\|\mathcal{B}y\|^2 + \beta\langle h(x_i, z_{\neq i}), \mathcal{B}y\rangle \tag{10}$$

is a block x_i surrogate function of $\varphi(x, y, \omega)$. A general update rule for x_i can be described as follows.

$$\begin{aligned} x_i^{k+1} \in \operatorname*{argmin}_{x_i} \Big\{ & f_i(x_i) + \langle h(x_i, x_{\neq i}^{k,i-1}), \omega^k + \beta\mathcal{B}y^k\rangle + u_i(x_i, x^{k,i-1}) \\ & - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i\rangle \Big\}. \end{aligned} \tag{11}$$

Denote $h_2(x) = \frac{1}{2}\|h(x)\|^2$. As we assume $x_i \mapsto h_2(x_i, y_{\neq i})$ is $l_i(y_{\neq i})$-smooth (see Assumption 1(i)), then we take the Lipschitz gradient surrogate for $\frac{1}{2}\|h(x)\|^2$ and formulate u_i of (10) as follows

$$u_i(x_i, z) = \hat{u}_i(x_i, z) + \beta h_2(z) + \beta\langle \nabla_i h_2(z), x_i - z_i\rangle + \frac{\beta\kappa_i(z_{\neq i})}{2}\|x_i - z_i\|^2, \tag{12}$$

where $\kappa_i(z_{\neq i}) \geq l_i(z_{\neq i})$ and $\hat{u}_i$ is a block surrogate function of F with respect to x_i.

After forming the block surrogate for φ, the second step is to take the inertial term

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \beta\nabla_i h_2(x_i^k, x_{\neq i}^{k,i-1}) - \beta\nabla_i h_2(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \beta\kappa_i^k(\bar{x}_i^k - x_i^k), \tag{13}$$

where $\kappa_i^k \geq l_i(x_{\neq i}^{k,i-1}) = l_i^k$ and $\bar{x}_i^k = x_i^k + \alpha_i^k(x_i^k - x_i^{k-1})$, here α_i^k is an extrapolation parameter. Then the update (11) of x_i is rewritten as follows.

$$\begin{aligned} x_i^{k+1} \in \operatorname*{argmin}_{x_i} \Big\{ & f_i(x_i) + \hat{u}_i(x_i, x_{\neq i}^{k,i-1}) + \langle h(x_i, x_{\neq i}^{k,i-1}), \omega^k + \beta\mathcal{B}y^k\rangle \\ & + \big\langle \beta\nabla_i h(\bar{x}_i^k, x_{\neq i}^{k,i-1})^\top h(\bar{x}_i^k, x_{\neq i}^{k,i-1}), x_i\big\rangle + \frac{\beta\kappa_i^k}{2}\|x_i - \bar{x}_i^k\|^2 \Big\}. \end{aligned} \tag{14}$$

Let us establish the NSDP for $\mathcal{L}_\beta$ when the update in (14) is used. We have

$$\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \leq \beta(l_i^k + \kappa_i^k)\alpha_i^k\|\Delta x_i^k\|.$$

The approximation error between $\varphi(x, y, \omega)$ and its block x_i surrogate in (10) satisfies that

$$\begin{aligned} & \hat{u}_i(x_i, z) - F(x_i, z_{\neq i}) + \beta h_2(z) + \beta\langle \nabla_i h_2(z), x_i - z_i\rangle \\ & \qquad + \frac{\kappa_i(z_{\neq i})\beta}{2}\|x_i - z_i\|^2 - \beta h_2(x_i, z_{\neq i}) \\ & \geq \theta_i(x_i, z) := \beta h_2(z) - \beta h_2(x_i, z_{\neq i}) + \beta\langle \nabla_i h_2(z), x_i - z_i\rangle + \frac{\kappa_i(z_{\neq i})\beta}{2}\|x_i - z_i\|^2. \end{aligned}$$

Note that $\nabla_{x_i}^2\theta_i(x_i, z) = \kappa_i(z_{\neq i})\beta\mathcal{I} - \beta\nabla_{x_i}^2 h_2(x_i, z_{\neq i})$. This implies $x_i \mapsto \theta_i(x_i, z)$ is $\beta(\kappa_i(z_{\neq i}) - l_i(z_{\neq i}))$-strongly convex (since $\nabla_{x_i}^2 h_2(x_i, z_{\neq i}) \preceq l_i(z_{\neq i})\mathcal{I}$). On the other hand, $\nabla_{x_i}\theta_i(z_i, z) = 0$. It follows from [13, Lemma 1] that

$$\theta_i(x_i, z) \geq \frac{1}{2}\beta(\kappa_i(z_{\neq i}) - l_i(z_{\neq i}))\|x_i - z_i\|^2.$$

We then apply Proposition 1 to obtain the following NSDP.

$$\mathcal{L}_\beta(x^{k,i}, y^k, \omega^k) + \eta_i^k\|\Delta x_i^{k+1}\|^2 \leq \mathcal{L}_\beta(x^{k,i-1}, y^k, \omega^k) + \gamma_i^k\|\Delta x_i^k\|^2, \tag{15}$$

where $\kappa_i^k > l_i^k$ and

$$\eta_i^k = \frac{(1-\nu_i)(\kappa_i^k - l_i^k)\beta}{2}, \quad \gamma_i^k = \frac{(a_i^k)^2}{2\nu_i(\kappa_i^k - l_i^k)\beta}, \quad a_i^k = \beta(l_i^k + \kappa_i^k)\alpha_i^k. \tag{16}$$

If $x_i \mapsto f_i(x_i) + \hat{u}_i(x_i, x_{\neq i}^{k,i-1})$ is convex and $x_i \mapsto h(x)$ is linear then we take $\kappa_i^k = l_i^k$ and the NSDP in (15) can be tighter. Indeed, as $x_i \mapsto h_2(x)$ and $x_i \mapsto f_i(x_i) + \hat{u}_i(x_i, x_{\neq i}^{k,i-1}) + \langle h(x_i, x_{\neq i}^{k,i-1}), \omega^k + \beta \mathcal{B} y^k \rangle$ are convex, we apply Proposition 2 to obtain

$$\begin{aligned} &h_2(x^{k,i}) + f_i(x_i^k) + \hat{u}_i(x_i^k, x_{\neq i}^{k,i-1}) + \langle h(x_i^k, x_{\neq i}^{k,i-1}), \omega^k + \beta \mathcal{B} y^k \rangle + \eta_i^k \|\Delta x_i^{k+1}\|^2 \\ &\leq h_2(x^{k,i-1}) + f_i(x_i^{k-1}) + \hat{u}_i(x_i^{k-1}, x_{\neq i}^{k,i-1}) + \langle h(x_i^{k-1}, x_{\neq i}^{k,i-1}), \omega^k + \beta \mathcal{B} y^k \rangle \\ &\quad + \gamma_i^k \|\Delta x_i^k\|^2, \end{aligned} \tag{17}$$

where $\gamma_i^k = \frac{1}{2}\beta l_i^k (\alpha_i^k)^2$ and $\eta_i^k = \frac{1}{2}\beta l_i^k$. Furthermore, note that $\hat{u}_i(x_i^{k-1}, x_{\neq i}^{k,i-1}) = F(x^{k,i-1})$ and $\hat{u}_i(x_i^k, x_{\neq i}^{k,i-1}) \geq F(x_i^k, x_{\neq i}^{k,i-1}) = F(x^{k,i})$. Therefore, (17) implies that (15) is satisfied with $\gamma_i^k = \frac{1}{2}\beta l_i^k (\alpha_i^k)^2$ and $\eta_i^k = \frac{1}{2}\beta l_i^k$. In the following proposition, we summarize the NSDP when using (14) to update x_i.

Proposition 3 *(NSDP when updating x_i) Suppose Assumption 1(i) is satisfied and x_i is updated as in (14). We choose $\kappa_i^k > l_i^k$, then the NSDP (15) is satisfied with η_i^k and γ_i^k defined in (16). If $x_i \mapsto f_i(x_i) + \hat{u}_i(x_i, x_{\neq i}^{k,i-1})$ is convex and $x_i \mapsto h(x)$ is linear then we take $\kappa_i^k = l_i^k$ and then the NSDP (15) is satisfied with $\gamma_i^k = \frac{1}{2}\beta l_i^k (\alpha_i^k)^2$ and $\eta_i^k = \frac{1}{2}\beta l_i^k$.*

We discuss two other situations in the following remarks.

Remark 1 Consider the case that together with Assumption 1(i) we also have $x_i \mapsto F(x_i, y_{\neq i})$ is an $L_i(y_{\neq i})$-smooth function. Then $x_i \mapsto \hat{h}(x) := F(x) + \frac{\beta}{2}\|h(x)\|^2$ is $\mathbf{l}_i(y_{\neq i}) = L_i(y_{\neq i}) + \beta l_i(y_{\neq i})$-smooth. We take the Lipschitz gradient surrogate for $F(x) + \frac{\beta}{2}\|h(x)\|^2$, that is, u_i in (10) is

$$u_i(x_i, z) = \hat{h}(z) + \langle \nabla_i \hat{h}(z), x_i - z_i \rangle + \frac{\kappa_i(z_{\neq i})}{2}\|x_i - z_i\|^2, \tag{18}$$

where $\kappa_i(z_{\neq i}) \geq \mathbf{l}_i(z_{\neq i})$. And we take the inertial term

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \nabla_i \hat{h}(x_i^k, x_{\neq i}^{k,i-1}) - \nabla_i \hat{h}(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \kappa_i^k(\bar{x}_i^k - x_i^k), \tag{19}$$

where $\kappa_i^k \geq \mathbf{l}_i^k = \mathbf{l}_i(x_{\neq i}^{k,i-1})$. The update (11) of x_i is rewritten as follows

$$\begin{aligned} x_i^{k+1} \in \operatorname*{argmin}_{x_i} \Big\{ &f_i(x_i) + \langle h(x_i, x_{\neq i}^{k,i-1}), \omega^k + \beta \mathcal{B} y^k \rangle \\ &+ \big\langle \nabla_i F(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \beta \nabla_i h(\bar{x}_i^k, x_{\neq i}^{k,i-1})^\top h(\bar{x}_i^k, x_{\neq i}^{k,i-1}), x_i \big\rangle + \frac{\kappa_i^k}{2}\|x_i - \bar{x}_i^k\|^2 \Big\}. \end{aligned} \tag{20}$$

Similarly to the above reasoning of Proposition 3, we take $\kappa_i^k > \mathbf{l}_i^k$ and apply Proposition 1 to obtain the NSDP in (15) with

$$\eta_i^k = \frac{(1-\nu_i)(\kappa_i^k - \mathbf{l}_i^k)\beta}{2}, \quad \gamma_i^k = \frac{(a_i^k)^2}{2\nu_i(\kappa_i^k - \mathbf{l}_i^k)\beta}, \quad a_i^k = \beta(\mathbf{l}_i^k + \kappa_i^k)\alpha_i^k. \tag{21}$$

If together with the smoothness of $x_i \mapsto \hat{h}(x)$ we assume that $x_i \mapsto f_i(x_i)$ and $x_i \mapsto F(x)$ are convex and $x_i \mapsto h(x)$ is linear, then we take $\kappa_i^k = \mathbf{l}_i^k$ and (15) holds with $\gamma_i^k = \frac{1}{2}\beta \mathbf{l}_i^k (\alpha_i^k)^2$ and $\eta_i^k = \frac{1}{2}\beta \mathbf{l}_i^k$.

Remark 2 In this remark, we discuss a case when Assumption 1(i) can be removed: we assume $x_i \mapsto F(x_i, y_{\neq i})$ is an $L_i(y_{\neq i})$-smooth function (but $x_i \mapsto h_2(x_i, y_{\neq i})$ can be non-smooth), then we take

$$u_i(x_i, z) = \beta h_2(x_i, z_{\neq i}) + F(z) + \langle \nabla_i F(z), x_i - z_i \rangle + \frac{\kappa_i(z_{\neq i})}{2}\|x_i - z_i\|^2, \tag{22}$$

where $\kappa_i(z_{\neq i}) \geq L_i(z_{\neq i})$. And we take the inertial term

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \nabla_i F(x_i^k, x_{\neq i}^{k,i-1}) - \nabla_i F(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \kappa_i^k(\bar{x}_i^k - x_i^k), \tag{23}$$

where $\kappa_i^k \geq L_i(x_{\neq i}^{k,i-1}) = L_i^k$. The update (11) of x_i is rewritten as follows

$$\begin{aligned} x_i^{k+1} \in \operatorname*{argmin}_{x_i} \Big\{ & f_i(x_i) + \langle h(x_i, x_{\neq i}^{k,i-1}), \omega^k + \beta \mathcal{B} y^k \rangle + \frac{\beta}{2}\|h(x_i, x_{\neq i}^{k,i-1})\|^2 \\ & + \big\langle \nabla_i F(\bar{x}_i^k, x_{\neq i}^{k,i-1}), x_i \big\rangle + \frac{\kappa_i^k}{2}\|x_i - \bar{x}_i^k\|^2 \Big\}. \end{aligned} \tag{24}$$

Similarly to the reasoning of Proposition 3, we take $\kappa_i^k > L_i^k$ and apply Proposition 1 to obtain the NSDP in (15) with

$$\eta_i^k = \frac{(1-\nu_i)(\kappa_i^k - L_i^k)\beta}{2}, \quad \gamma_i^k = \frac{(a_i^k)^2}{2\nu_i(\kappa_i^k - L_i^k)\beta}, \quad a_i^k = \beta(L_i^k + \kappa_i^k)\alpha_i^k. \tag{25}$$

If together with the smoothness of $x_i \mapsto F(x)$ we assume that $x_i \mapsto f_i(x_i)$ and $x_i \mapsto F(x)$ are convex and $x_i \mapsto h(x)$ is linear, then we take $\kappa_i^k = L_i^k$ and (15) holds with $\gamma_i^k = \frac{1}{2}\beta L_i^k (\alpha_i^k)^2$ and $\eta_i^k = \frac{1}{2}\beta L_i^k$.

Update y As G is L_G-smooth, we form a block y surrogate of φ by summing the Lipschitz gradient surrogate of G and the remaining part of φ:

$$\begin{aligned} \hat{\varphi}(y, x, y', \omega) = G(y') + \langle \nabla G(y'), y - y' \rangle + \frac{L_G}{2}\|y - y'\|^2 \\ + F(x) + \langle h(x) + \mathcal{B}y, \omega \rangle + \frac{\beta}{2}\|h(x) + \mathcal{B}y\|^2. \end{aligned}$$

Block y is updated as follows

$$\begin{aligned} y^{k+1} & \in \operatorname*{argmin}_y \hat{\varphi}(y, x^{k+1}, y^k, \omega^k) \\ & = \operatorname*{argmin}_y \langle \mathcal{B}^\top \omega^k + \nabla G(y^k), y \rangle + \frac{\beta}{2}\|h(x^{k+1}) + \mathcal{B}y\|^2 + \frac{L_G}{2}\|y - y^k\|^2, \\ & = (\beta \mathcal{B}^\top \mathcal{B} + L_G \mathbf{I})^{-1}\big(L_G y^k - \nabla G(y^k) - \mathcal{B}^\top(\omega^k + \beta h(x^{k+1}))\big). \end{aligned} \tag{26}$$

Proposition 4 *(Sufficient decrease when updating y) The update in (26) guarantees a sufficient decrease*

$$\mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^k) + \frac{\delta}{2}\|y^{k+1} - y^k\|^2 \leq \mathcal{L}_\beta(x^{k+1}, y^k, \omega^k), \tag{27}$$

where $\delta = L_G + \beta\lambda_{\min}(\mathcal{B}^\top \mathcal{B})$. If G is convex, then (27) is satisfied with $\delta = L_G$.

Table 1 γ_i^k and η_i^k in the NSDP (15)

γ_i^k and η_i^k are defined in:
Proposition 3 if we assume Assumption 1(i) and use the update rule in (14),
Remark 1 if we assume Assumption 1(i) together with $x_i \mapsto F(x_i, y_{\neq i})$ being an $L_i(y_{\neq i})$-smooth function and use the update rule in (20),
Remark 2 if Assumption 1(i) is not satisfied but $F(x_i, y_{\neq i})$ is an $L_i(y_{\neq i})$-smooth function and we use the update rule in (24).

Proof We note that $y \mapsto \hat{\varphi}(y, x, y', \omega)$ is $L_G + \beta\lambda_{\min}(\mathcal{B}^\top\mathcal{B})$-strongly convex. Applying Proposition 1 (note that we do not use inertial term for updating y), we obtain (27). If G is convex, then we apply Proposition 2 (we simply take $\alpha_i^k = \beta_i^k = 0$ in Proposition 2) to obtain the result. □

3.2 Convergence analysis

Given τ_1, τ_2, we denote $C_1 = \frac{(\tau_1+1)|\tau_1-\tau_2|}{2\sigma_{\mathcal{B}}\tau_2\beta(1-|\tau_1-\tau_2|)}$, $C_2 = \frac{(\tau_1+1)\tau_2/\tau_1}{2\sigma_{\mathcal{B}}\beta(1-|\tau_1-\tau_2|)(1-|1-\tau_2/\tau_1|)}$, $C_3 = \frac{\delta}{2} - 2C_2L_G^2$, where δ is defined in Proposition 4. Note that γ_i^k and η_i^k are the coefficients in the NSDP(15), their formulas are summarized in Table 1.

We need to choose the parameters such that they satisfy the following conditions.

$$\begin{aligned} &\tau_1 \in (0, 1], \quad \tau_2/\tau_1 \in (0, 2), \quad |\tau_1 - \tau_2| < 1 \\ &\gamma_i^k \leq B_1\eta_i^{k-1}, \quad 8C_2L_G^2 \leq B_2C_3, \end{aligned} \tag{28}$$

where $B_1, B_2 \in (0, 1)$ are some constants. In this section, we will establish the subsequential convergence and the global convergence for iADMMn. The following proposition provides a recursive inequality for $\{\mathcal{L}_\beta(x^k, y^k; \omega^k)\}$.

Proposition 5 *Considering Algorithm 1, the following inequality holds*

$$\begin{aligned} &\mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^{k+1}) - \frac{1-\tau_1}{2\tau_2\beta}\|\omega^{k+1}\|^2 + \sum_{i=1}^s \eta_i^k\|\Delta x_i^{k+1}\|^2 + C_3\|\Delta y^{k+1}\|^2 \\ &\leq \mathcal{L}_\beta(x^k, y^k, \omega^k) - \frac{1-\tau_1}{2\tau_2\beta}\|\omega^k\|^2 + \sum_{i=1}^s \gamma_i^k\|\Delta x_i^k\|^2 \\ &\quad + C_1\big(\|\mathcal{B}^\top\Delta\omega^k\|^2 - \|\mathcal{B}^\top\Delta\omega^{k+1}\|^2\big) + 8C_2L_G^2\|\Delta y^k\|^2. \end{aligned}$$

Proof Optimality condition of (26) gives us

$$\mathcal{B}^\top(\omega^k + \beta(h(x^{k+1}) + \mathcal{B}y^{k+1})) + \nabla G(y^k) + L_G(y^{k+1} - y^k) = 0 \tag{29}$$

Denote $t^{k+1} = L_G\Delta y^{k+1} + \nabla G(y^k)$. From (29) and the update of ω in (9) we have

$$\mathcal{B}^\top((1 - \frac{\tau_1}{\tau_2})\omega^k + \frac{1}{\tau_2}\omega^{k+1}) = -t^{k+1},$$

which implies that

$$\frac{1}{\tau_2}\mathcal{B}^\top\Delta\omega^{k+1} + (1 - \frac{\tau_1}{\tau_2})\mathcal{B}^\top\Delta\omega^k = -\Delta t^{k+1}.$$

Hence, we have

$$\frac{1}{\tau_1}\mathcal{B}^\top \Delta\omega^{k+1} = (1-\frac{\tau_2}{\tau_1})\mathcal{B}^\top \Delta\omega^k - \frac{\tau_2}{\tau_1}\Delta t^{k+1}. \tag{30}$$

Note that if $\tau_2/\tau_1 \in [1,2)$ then (30) can be rewritten as

$$\frac{1}{\tau_1}\mathcal{B}^\top \Delta\omega^{k+1} = -(\tau_2/\tau_1 - 1)\mathcal{B}^\top \Delta\omega^k - (2-\tau_2/\tau_1)\frac{\tau_2/\tau_1}{2-\tau_2/\tau_1}\Delta t^{k+1}.$$

Hence, from $\tau_2/\tau_1 \in (0,2)$ and the convexity of the norm $\|\cdot\|$, we can derive from (30) the following inequality

$$\frac{1}{\tau_1}\|\mathcal{B}^\top \Delta\omega^{k+1}\|^2 \le |1-\tau_2/\tau_1|\|\mathcal{B}^\top \Delta\omega^k\|^2 + \frac{(\tau_2/\tau_1)^2}{1-|1-\tau_2/\tau_1|}\|\Delta t^{k+1}\|^2. \tag{31}$$

On the other hand, from the definition of t^{k+1} and the Lipschitz continuity of ∇G, we have $\|\Delta t^{k+1}\| \le L_G\|\Delta y^{k+1}\| + L_G\|\Delta y^k\| + L_G\|\Delta y^k\|$. It implies that

$$\|\Delta t^{k+1}\|^2 \le 2L_G^2\|\Delta y^{k+1}\|^2 + 8L_G^2\|\Delta y^k\|^2.$$

Hence, from (31) we obtain

$$\begin{aligned}\|\mathcal{B}^\top \Delta\omega^{k+1}\|^2 \le{}& \frac{|\tau_1-\tau_2|}{(1-|\tau_1-\tau_2|)}\big(\|\mathcal{B}^\top \Delta\omega^k\|^2 - \|\mathcal{B}^\top \Delta\omega^{k+1}\|^2\big)\\ &+ \frac{\tau_2^2/\tau_1}{(1-|\tau_1-\tau_2|)(1-|1-\tau_2/\tau_1|)}\big(2L_G^2\|\Delta y^{k+1}\|^2 + 8L_G^2\|\Delta y^k\|^2\big).\end{aligned} \tag{32}$$

Furthermore,

$$\begin{aligned}&\mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^{k+1}) - \mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^k)\\ &= \big\langle h(x^{k+1}) + \mathcal{B}y^{k+1}, \omega^{k+1} - \omega^k\big\rangle\\ &= \big\langle \frac{1}{\tau_2\beta}(\omega^{k+1} - \tau_1\omega^k), \omega^{k+1} - \omega^k\big\rangle\\ &= \frac{1}{\tau_2\beta}\big\langle \tau_1(\omega^{k+1} - \omega^k) + (1-\tau_1)\omega^{k+1}, \omega^{k+1} - \omega^k\big\rangle\\ &= \frac{\tau_1}{\tau_2\beta}\|\Delta\omega^{k+1}\|^2 + \frac{1-\tau_1}{2\tau_2\beta}\big(\|\Delta\omega^{k+1}\|^2 + \|\omega^{k+1}\|^2 - \|\omega^k\|^2\big),\end{aligned}$$

which leads to

$$\begin{aligned}&\mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^{k+1}) - \mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^k)\\ &= \frac{\tau_1+1}{2\tau_2\beta}\|\Delta\omega^{k+1}\|^2 + \frac{1-\tau_1}{2\tau_2\beta}\big(\|\omega^{k+1}\|^2 - \|\omega^k\|^2\big).\end{aligned} \tag{33}$$

Therefore, from (32), (33), and noting that $\sigma_{\mathcal{B}}\|\Delta\omega^{k+1}\|^2 \le \|\mathcal{B}^\top \Delta\omega^{k+1}\|^2$, we have

$$\begin{aligned}&\mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^{k+1}) - \mathcal{L}_\beta(x^{k+1}, y^{k+1}, \omega^k)\\ &\le C_1\big(\|\mathcal{B}^\top \Delta\omega^k\|^2 - \|\mathcal{B}^\top \Delta\omega^{k+1}\|^2\big) + C_2 2L_G^2\|\Delta y^{k+1}\|^2 + C_2 8L_G^2\|\Delta y^k\|^2\\ &\quad + \frac{1-\tau_1}{2\tau_2\beta}\big(\|\omega^{k+1}\|^2 - \|\omega^k\|^2\big).\end{aligned}$$

Together with (15) and (27) we obtain the result. □

Lemma 1 *Denote*

$$\begin{aligned}\hat{\mathcal{L}}^k = \mathcal{L}_\beta(x^k, y^k, \omega^k) - \frac{1-\tau_1}{2\tau_2\beta}\|\omega^k\|^2 + \sum_{i=1}^{s} B_1\eta_i^{k-1}\|\Delta x_i^k\|^2 \\ + C_1\|\mathcal{B}^\top\Delta\omega^k\|^2 + B_2C_3\|\Delta y^k\|^2.\end{aligned} \tag{34}$$

We have $\hat{\mathcal{L}}^k \geq \nu$ for all $k \geq 1$, where ν is the lower bound of $\Theta(x, y)$ (see Assumption 1).

Proof We use the technique of [23, Lemma 2.9]. From Proposition 5 and the conditions $\gamma_i^k \leq B_1\eta_i^{k-1}$ and $8C_2L_G^2 \leq B_2C_3$ for some constants $B_1, B_2 \in (0, 1)$, we have $\hat{\mathcal{L}}^{k+1} \leq \hat{\mathcal{L}}^k$, that is, $\{\hat{\mathcal{L}}^k\}$ is a non-increasing sequence. Suppose there exists k_0 such that $\hat{\mathcal{L}}^k < \nu$ for all $k \geq k_0$, then we have

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

This implies that $\sum_{k=1}^{\infty}(\hat{\mathcal{L}}^k - \vartheta) = -\infty$. However, since

$$\begin{aligned}\hat{\mathcal{L}}^k &\geq \mathcal{L}_\beta(x^k, y^k, \omega^k) - \frac{1-\tau_1}{2\tau_2\beta}\|\omega^k\|^2 \\ &\geq \nu + \langle h(x^k) + \mathcal{B}y^k, \omega^k\rangle - \frac{1-\tau_1}{2\tau_2\beta}\|\omega^k\|^2 \\ &= \nu + \frac{1}{\tau_2\beta}\langle\omega^k - \tau_1\omega^{k-1}, \omega^k\rangle - \frac{1-\tau_1}{2\tau_2\beta}\|\omega^k\|^2 \\ &= \nu + \frac{1+\tau_1}{2\tau_2\beta}\|\omega^k\|^2 - \frac{\tau_1}{\tau_2\beta}\langle\omega^{k-1}, \omega^k\rangle \\ &= \nu + \frac{1+\tau_1}{2\tau_2\beta}\|\omega^k\|^2 + \frac{\tau_1}{2\tau_2\beta}(\|\Delta\omega^k\|^2 - \|\omega^k\|^2 - \|\omega^{k-1}\|^2) \\ &\geq \nu + \frac{1}{2\tau_2\beta}\|\omega^k\|^2 - \frac{\tau_1}{2\tau_2\beta}\|\omega^{k-1}\|^2 \\ &\geq \nu + \frac{\tau_1}{2\tau_2\beta}(\|\omega^k\|^2 - |\omega^{k-1}\|^2),\end{aligned}$$

we have

$$\sum_{k=1}^{K}(\hat{\mathcal{L}}^k - \vartheta) \geq \frac{\tau_1}{2\tau_2\beta}\sum_{k=1}^{K}(\|\omega^k\|^2 - \|\omega^{k-1}\|^2) \geq \frac{\tau_1}{2\tau_2\beta}(-\|\omega^0\|^2),$$

which gives a contradiction to the fact $\sum_{k=1}^{\infty}(\hat{\mathcal{L}}^k - \vartheta) = -\infty$. □

Proposition 6 *Consider Algorithm 1. The values of γ_i^k and η_i^k are given in Table 1. Suppose there exist $\tilde{\eta}_i > 0$ such that $\eta_i^k \geq \tilde{\eta}_i$ for all $i \in [s]$ and $k \geq 1$. Then the sequence $\{\Delta y^k\}_{k\geq 0}$, $\{\Delta x_i^k\}_{k\geq 0}$, and $\{\Delta\omega^k\}_{k\geq 0}$ converge to 0.*

Proof From Proposition 5 we have

$$\hat{\mathcal{L}}^{k+1} + (1 - B_1)\sum_{i=1}^{s}\eta_i^k\|\Delta x_i^{k+1}\|^2 + (1 - B_2)C_3\|\Delta y^{k+1}\|^2 \leq \hat{\mathcal{L}}^k, \tag{35}$$

where $\hat{\mathcal{L}}^k$ is defined in (34). For all $K \geq 1$, by summing (35) from $k = 0$ to K we get

$$\hat{\mathcal{L}}^{K+1} + (1 - B_1) \sum_{k=0}^{K} \sum_{i=1}^{s} \eta_i^k \|\Delta x_i^{k+1}\|^2 + (1 - B_2) C_3 \sum_{k=0}^{K} \|\Delta y^{k+1}\|^2 \leq \hat{\mathcal{L}}^0. \tag{36}$$

From (36), Lemma 1 and the assumption $\eta_i^k \geq \tilde{\eta}_i > 0$ we derive that the sequence $\{\Delta y^k\}_{k\geq 0}$ and $\{\Delta x_i^k\}_{k\geq 0}$ converge to 0.

From (32), we have

$$\sum_{k=0}^{K} \|\mathcal{B}^\top \Delta\omega^{k+1}\|^2 \leq \frac{|\tau_1 - \tau_2|}{(1 - |\tau_1 - \tau_2|)} \|\mathcal{B}^\top \Delta\omega^0\|^2 + \frac{\tau_2^2/\tau_1}{(1 - |\tau_1 - \tau_2|)(1 - |1 - \tau_2/\tau_1|)} \big(2L_G^2 \sum_{k=0}^{K} \|\Delta y^{k+1}\|^2 + 8L_G^2 \sum_{k=0}^{K} \|\Delta y^k\|^2\big).$$

It implies that $\sum_{k=0}^{\infty} \|\Delta\omega^k\|^2 \leq \sum_{k=0}^{\infty} 1/\sigma_{\mathcal{B}} \|\mathcal{B}^\top \Delta\omega^k\|^2 < +\infty$. Hence $\{\Delta\omega^k\}_{k\geq 0}$ converge to 0. □

Theorem 1 *(Subsequential convergence) Consider Algorithm 1 and the parameters are chosen to satisfy the conditions in (28). The values of γ_i^k and η_i^k are summarized in Table 1. Suppose there exist $\tilde{\eta}_i > 0$ such that $\eta_i^k \geq \tilde{\eta}_i$ for all $i \in [s]$ and $k \geq 1$. And suppose a_i^k (which are defined in (16) if we use the update rule in (14)), defined in (21) if we use use the update rule in (20), and defined in (25) if we use use the update rule in (24)) are upper bounded.*[2] *It holds that if (x^*, y^*, ω^*) is a limit point of the generated sequence of iADMMn then (x^*, y^*, ω^*) is an $\frac{1-\tau_1}{\tau_2\beta}\|\omega^*\|$-approximate stationary point of Problem (1). When $\tau_1 = 1$, we have (x^*, y^*) is a stationary point of Problem (1) (or equivalently, (x^*, y^*, ω^*) is a critical point of $\mathcal{L}_\beta$).*

Proof We remind that the update rule of x_i^k in (14), (20) and (24) are just special cases of (11). In (11), $u_i(x_i, z)$ is a block x_i surrogate function of $F(x) + \frac{\beta}{2}\|h(x)\|^2$ and the formulas of $\mathcal{G}$ in (13), (19), and (23), which respectively correspond to the update in (14), (20), and (24), satisfy the condition $\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \leq a_i^k \|\Delta x_i^k\|$.

From (11), we know that the following inequality holds for all x_i

$$\begin{aligned} &f_i(x_i^{k+1}) + \langle h(x_i^{k+1}, x_{\neq i}^{k,i-1}), \omega^k + \beta\mathcal{B}y^k\rangle + u_i(x_i^{k+1}, x^{k,i-1}) \\ &\leq f_i(x_i) + \langle h(x_i, x_{\neq i}^{k,i-1}), \omega^k + \beta\mathcal{B}y^k\rangle + u_i(x_i, x^{k,i-1}) \\ &\quad - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i - x_i^{k+1}\rangle. \end{aligned} \tag{37}$$

Let $\{(x^{k_n}, y^{k_n}, \omega^{k_n})\}$ be a subsequence that converges to (x^*, y^*, ω^*). As $\Delta x_i^k \to 0$ (see Proposition 6), we have $x_i^{k_n+1}$, $x_i^{k_n-1}$ and $x_i^{k_n-2}$ also converge to x_i^* for all $i \in [s]$. Consequently, $\|\mathcal{G}_i^k(x_i^{k_n-1}, x_i^{k_n-2})\| \to 0$. Taking $x_i = x_i^*$ and $k = k_n - 1$ in (37), we obtain $\limsup_{n\to\infty} f_i(x_i^{k_n}) \leq f_i(x_i^*)$, which implies $f_i(x_i^{k_n}) \to f_i(x_i^*)$ as f_i is lower semi-continuous. Taking $k = k_n - 1$ in (37) and let $k_n \to \infty$, we obtain the following inequality

[2] In fact, if the Lipschitz constants of $x_i \mapsto h_2(x)$ in Assumption 1(i), or that of $x_i \mapsto \hat{h}(x)$ in Remark 1, or that of $x_i \mapsto F(x)$ in Remark 2 are upper bounded over the set that contains $\{x^k\}_{k\geq 0}$ then this assumption is satisfied. Note that $x_i^k \in \text{dom}\,(f_i)$.

for all x_i

$$\begin{aligned} &f_i(x_i^*) + \langle h(x^*), \omega^* + \beta\mathcal{B}y^*\rangle + u_i(x_i^*, x^*) \\ &\le f_i(x_i) + \langle h(x_i, x_{\neq i}^*), \omega^* + \beta\mathcal{B}y^*\rangle + u_i(x_i, x^*). \end{aligned} \tag{38}$$

Note that $u_i(x_i^*, x^*) = F(x^*) + \frac{\beta}{2}\|h(x^*)\|^2$. Hence, (38) implies that for all x_i we have

$$\begin{aligned} \varphi(x^*, y^*, \omega^*) + f_i(x_i^*) \le \varphi(x_i, x_{\neq i}^*, y^*, \omega^*) + f_i(x_i) \\ + u_i(x_i, x^*) - F(x_i, x_{\neq i}^*) - \frac{\beta}{2}\|h(x_i, x_{\neq i}^*)\|^2. \end{aligned} \tag{39}$$

Let $e_i(x_i, z) = u_i(x_i, z) - F(x_i, z_{\neq i}) - \frac{\beta}{2}\|h(x_i, z_{\neq i})\|^2$. We have $e_i(x_i^*, x^*) = 0$. It follows from (39) that x_i^* is a solution of

$$\min_{x_i} \varphi(x_i, x_{\neq i}^*, y^*, \omega^*) + f_i(x_i) + e_i(x_i, x^*).$$

Writing the optimality condition for this problem and noting that $\nabla_{x_i} e_i(x_i^*, x^*) = 0$ for all the three surrogates in (12), (18), and (22) that respectively corresponds to the update rule of x_i in (14), (20) and (24), we obtain

$$0 \in \partial_{x_i}\mathcal{L}_\beta(x^*, y^*, \omega^*) = \partial_{x_i}(\varphi(x^*, y^*, \omega^*) + f_i(x_i^*)).$$

Since the essence of the update rule of y is to minimize a block y surrogate function of φ (see (26)), we can use the same reasoning for x_i to prove that $0 \in \nabla_y\mathcal{L}_\beta(x^*, y^*, \omega^*)$.

From (9) we have $h(x^{k_n}) + \mathcal{B}y^{k_n} = \frac{1}{\tau_2\beta}(\Delta\omega^{k_n} + (1-\tau_1)\omega^{k_n-1})$. Furthermore, $\Delta\omega^{k_n} \to 0$ (see Proposition 6), which implies $\omega^{k_n-1} \to \omega^*$. Hence, we have

$$h(x^*) + \mathcal{B}y^* = \frac{1-\tau_1}{\tau_2\beta}\omega^*.$$

The result follows then. □

Theorem 2 *(Global convergence) Suppose the conditions for Theorem 1 are satisfied and we also assume that $\Theta(x, y)$ has the KŁ property, the generated sequence $\{(x^k, y^k, \omega^k)\}$ is bounded, and let $\tau_1 = \tau_2 = 1$. Furthermore, together with the existence of the constants $\tilde{\eta}_i$ in Theorem 1 we assume there exist $\bar{\eta}_i > 0$ such that $\eta_i^k \le \bar{\eta}_i$ for all $i \in [s]$ and $k \ge 0$, and the constant B_1 in (28) satisfies $B_1 < \min\{\tilde{\eta}_i/\bar{\eta}_i\}$.*

Then the whole sequence $\{(x^k, y^k, \omega^k)\}$ converges to a critical point of $\mathcal{L}_\beta$.

Before proving Theorem 2, let us have some discussion on the assumptions used for Theorem 2.

- If $h(x) = \mathcal{A}x = \sum_{i=1}^s \mathcal{A}_i x_i$, where $\mathcal{A}_i$ are linear mappings, and we use the surrogate in (12) then $l_i^k = \|\mathcal{A}_i^\top \mathcal{A}_i\|$. In this case, we simply choose κ_i^k in the update (14) to be any constant $\kappa_i \ge \|\mathcal{A}_i^\top \mathcal{A}_i\|$, then η_i^k in Proposition 3 are constants. Hence, $\tilde{\eta}_i/\bar{\eta}_i = 1$.
- It is important to note that we require $\tau_1 = \tau_2 = 1$ for the global convergence, but this condition is not required for Theorem 1.
- The boundedness assumption of $\{(x^k, y^k, \omega^k)\}$ is necessary for the upcoming proof. Totally similarly to [15, Proposition 7], we can prove that if $\tau_1 = 1, ran\, h(x) \subseteq Im(\mathcal{B})$, $\lambda_{\min}(\mathcal{B}^\top\mathcal{B}) > 0$, and $\Theta(x, y)$ is coercive over the feasible set $\{(x, y) : h(x) + \mathcal{B}y = 0\}$ then the generated sequence $\{(x^k, y^k, \omega^k)\}$ is bounded. We omit the proof of this property.

– Once the global convergence of iADMMn is guaranteed, by using the same technique as in [2, Theorem 2] (see [38, Theorem 2.9] and [14, Theorem 3] for some examples of using this technique to establish the convergence rate), we can establish a convergence rate for iADMMn. The type of convergence rate depends on the value of the KŁ exponent. More specifically, when the KŁ exponent is 0, the algorithm converges after a finite number of steps, when the KŁ exponent is in $(0, 1/2]$, the algorithm has linear convergence, and when the KŁ exponent is in $(1/2, 1)$, the algorithm has sublinear convergence. The estimation of the KŁ exponent falls beyond the scope of this paper.

We now prove Theorem 2. As discussed at the beginning of Sect. 3, we use the same methodology employed in [7, 13, 16, 17, 29, 37, 38, 41] to prove the global convergence of iADMMn, which comprises of three main steps (i) derive sufficient decrease property of a Lyapunov function, (ii) derive a subgradient lower bound for the iterates gap, and (iii) using the KŁ property. Although the general methodology is the same in these papers as its essence is originally from the general proof recipe of [3, 6], how to choose a Lyapunov function and how to establish the two properties (i) the sufficient decrease and (ii) the boundedness of subgradient are different in these papers since these steps highly rely on the structure of the problem. In the following, we will prove these two properties. The remaining step to obtain the global convergence are totally similar to the proof of [13, Theorem 2]; hence, we omit the details.

Proof Denote $z = (x, y, \omega)$. We use the following Lyapunov function

$$\begin{aligned}\bar{\mathcal{L}}(z, \tilde{z}) = \mathcal{L}_\beta(x, y, \omega) + \sum_{i=1}^{s} \frac{\tilde{\eta}_i + B_1\bar{\eta}_i}{2}\|x_i - \tilde{x}_i\|^2 \\ + \frac{(1 + B_2)C_3}{2}\|y - \tilde{y}\|^2 + C_1\|\mathcal{B}^\top(\omega - \tilde{\omega})\|^2.\end{aligned} \tag{40}$$

Sufficient decrease property From Proposition 5 and the condition $\tilde{\eta}_i \le \eta_i^k$, $\gamma_i^k \le B_1\eta_i^{k-1} \le B_1\bar{\eta}_i$, and $8C_2L_G^2 \le B_2C_3$, we obtain

$$\begin{aligned}\bar{\mathcal{L}}(z^{k+1}, z^k) + \sum_{i=1}^{s} \frac{\tilde{\eta}_i - B_1\bar{\eta}_i}{2}\big(\|\Delta x_i^{k+1}\|^2 + \|\Delta x_i^k\|^2\big) \\ + \frac{(1 - B_2)C_3}{2}\big(\|\Delta y^{k+1}\|^2 + \|\Delta y^k\|^2\big) \le \bar{\mathcal{L}}(z^k, z^{k-1}).\end{aligned}$$

Boundedness of subgradient In the following, we will work on the bounded set that contains the generated sequence (as we assume the generated sequence is bounded). The optimality condition of (11) gives us

$$\begin{aligned}\mathcal{G}_i^k(x_i^k, x_i^{k-1}) - \nabla_{x_i}h(x_i^{k+1}, x_{\neq i}^{k,i-1})^\top(\omega^k + \beta\mathcal{B}y^k) \\ \in \partial_{x_i}f_i(x_i^{k+1}) + \nabla_{x_i}u_i(x_i^{k+1}, x^{k,i-1}).\end{aligned} \tag{41}$$

Note that $\nabla_{x_i}u_i(x_i^{k+1}, x^{k+1}) = \nabla_{x_i}\big(F(x^{k+1}) + \beta h_2(x^{k+1})\big)$, $\nabla_{x_i}u_i(\cdot,\cdot)$ is continuously differentiable, and we are working on the bounded set containing the generated sequence. So there exists a constant $\hat{L}_i$ such that

$$\|\nabla_{x_i}u_i(x_i^{k+1}, x^{k,i-1}) - \nabla_{x_i}\big(F(x^{k+1}) + \beta h_2(x^{k+1})\big)\| \le \hat{L}_i\|x^{k+1} - x^{k,i-1}\|. \tag{42}$$

From (41) we know that there exists $d_i^{k+1} \in \partial_{x_i}f_i(x_i^{k+1})$ such that

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) - \nabla_{x_i} h(x_i^{k+1}, x_{\neq i}^{k,i-1})^\top(\omega^k + \beta\mathcal{B}y^k) = d_i^{k+1} + \nabla_{x_i} u_i(x_i^{k+1}, x^{k,i-1}). \tag{43}$$

Since h_2 is continuously differentiable and

$$\begin{aligned}\partial_{x_i}\mathcal{L}_\beta(z^{k+1}) = \partial_{x_i} f_i(x_i^{k+1}) + \nabla_{x_i}\big(F(x^{k+1}) + \beta h_2(x^{k+1})\big)\\ + \nabla_{x_i} h(x^{k+1})^\top(\omega^{k+1} + \beta\mathcal{B}y^{k+1}),\end{aligned}$$

we have (denote $\bar{\xi}_i^{k+1} = \nabla_{x_i}\big(F(x^{k+1}) + \beta h_2(x^{k+1})\big)$

$$D_i^{k+1} := d_i^{k+1} + \bar{\xi}_i^{k+1} + \nabla_{x_i} h(x^{k+1})^\top(\omega^{k+1} + \beta\mathcal{B}y^{k+1}) \in \partial_{x_i}\mathcal{L}_\beta(z^{k+1}).$$

Together with (43) and (42) we get

$$\begin{aligned}\|D_i^{k+1}\| &= \|d_i^{k+1} + \xi_i^{k+1} + \bar{\xi}_i^{k+1} - \xi_i^{k+1} + \nabla_{x_i} h(x^{k+1})^\top(\omega^{k+1} + \beta\mathcal{B}y^{k+1})\|\\ &\le \big\|\mathcal{G}_i^k(x_i^k, x_i^{k-1}) - \nabla_{x_i} h(x_i^{k+1}, x_{\neq i}^{k,i-1})^\top(\omega^k + \beta\mathcal{B}y^k)\\ &\quad + \nabla_{x_i} h(x^{k+1})^\top(\omega^{k+1} + \beta\mathcal{B}y^{k+1})\big\| + \|\bar{\xi}_i^{k+1} - \nabla_{x_i} u_i(x_i^{k+1}, x^{k,i-1})\|\\ &\le \|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| + \|\nabla_{x_i} h(x^{k+1})^\top\|(\|\Delta\omega^{k+1}\| + \beta\|\mathcal{B}\Delta y^{k+1}\|)\\ &\quad + \hat{L}_i\|x^{k+1} - x^{k,i-1}\|.\end{aligned} \tag{44}$$

On the other hand, as $h_i, i = 1, \ldots, q$, are continuously differentiable, hence $\|\nabla_{x_i} h(x^{k+1})^\top\|$ is bounded on the bounded set that contains the generated sequence. Since $\tau_1 = \tau_2 = 1$, we get from (32) that

$$\|\Delta\omega^{k+1}\|^2 \le \frac{1}{\sigma_{\mathcal{B}}}\|\mathcal{B}^\top\Delta\omega^{k+1}\|^2 \le \frac{1}{\sigma_{\mathcal{B}}}(2L_G^2\|\Delta y^{k+1}\|^2 + 8L_G^2\|\Delta y^k\|^2). \tag{45}$$

Therefore, from (44) we derive that

$$\|D_i^{k+1}\| \le a_1(\|\Delta x^{k+1}\| + \|\Delta x^k\| + \|\Delta y^{k+1}\| + \|\Delta y^k\|) \tag{46}$$

for some positive constant a_1. Note that

$$D_y^{k+1} := \nabla G(y^{k+1}) + \mathcal{B}^\top(\omega^{k+1} + \beta(h(x^{k+1}) + \mathcal{B}y^{k+1})) \in \nabla_y\mathcal{L}_\beta(z^{k+1}),$$

which together with(29) and (45) leads to

$$\begin{aligned}\|D_y^{k+1}\| &= \|\nabla G(y^{k+1}) - \nabla G(y^k) + \mathcal{B}^\top\Delta\omega^{k+1} - L_G\Delta y^{k+1}\|\\ &\le 2L_G\|\Delta y^{k+1}\| + \|\mathcal{B}^\top\Delta\omega^{k+1}\|\\ &\le a_2(\|\Delta y^{k+1}\| + \|\Delta y^k\|)\end{aligned} \tag{47}$$

for some positive constant a_2. On the other hand, we have

$$\|\nabla_\omega\mathcal{L}_\beta(z^{k+1})\| = \|h(x^{k+1}) + \mathcal{B}y^{k+1}\| = 1/\beta\|\Delta\omega^{k+1}\|, \tag{48}$$

and

$$\begin{aligned}\partial\bar{\mathcal{L}}(z,\tilde{z}) = \partial\mathcal{L}_\beta(z) + \partial\Big(\sum_{i=1}^s\frac{\tilde{\eta}_i + B_1\bar{\eta}_i}{2}\|x_i - \tilde{x}_i\|^2 + B_2C_3\|y - \tilde{y}\|^2\\ + C_1\|\mathcal{B}^\top(\omega - \tilde{\omega})\|^2\Big).\end{aligned} \tag{49}$$

Therefore, from (49), (48), (45), (47), and (44), it is not difficult to derive that

$$\|D^{k+1}\| \le a_3(\|\Delta x^{k+1}\| + \|\Delta x^k\| + \|\Delta y^{k+1}\| + \|\Delta y^k\|),$$

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

4 Numerical results

In this section, we test iADMMn on Problem (3). All tests are performed using Matlab R2021b on a laptop 2.5 GHz Intel Core i5 of 16GB RAM. The code is available from https://github.com/LeThiKhanhHien/iADMMn.

Problem (3) can be rewritten in the form of (1) as follows.

$$\begin{aligned} \min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{r\times n}} \quad & F(U,V) + G(W) \\ s.t \quad & UV - W = 0, \end{aligned} \tag{50}$$

where $F(U,V) = \frac{\lambda_d}{2}\|U\|_F^2 + \frac{\lambda_t}{2}\|V\|_F^2$ and $G(W) = \sum_{i,j}(1+cy_{ij}-y_{ij})\log(1+\exp(W_{ij})) - cy_{ij}W_{ij}$. The augmented Lagrangian of Problem (50) is

$$\mathcal{L}_\beta(U,V,W,\omega) = F(U,V) + G(W) + \langle UV - W, \omega\rangle + \frac{\beta}{2}\|UV - W\|^2.$$

The update in (14) for U (we keep F as a surrogate of itself) is

$$\begin{aligned} U^{k+1} &\in \arg\min_U \left\{ \frac{\lambda_d}{2}\|U\|_F^2 + \langle U, (w-\beta W)V^\top\rangle + \beta\langle U_{ex}VV^\top, U\rangle \right. \\ &\qquad \left. + \frac{\beta\|VV^\top\|}{2}\|U - U_{ex}\|^2 \right\} \\ &= \frac{\beta\|VV^\top\|}{\beta\|VV^\top\| + \lambda_d}U_{ex} - \frac{1}{\beta\|VV^\top\| + \lambda_d}wV^\top - \frac{\beta(U_{ex}V - W)V^\top}{\beta\|VV^\top\| + \lambda_d}, \end{aligned}$$

where $W = W^k$, $V = V^k$, $\omega = \omega^k$ and $U_{ex} = U^k + \alpha_U^k(U^k - U^{k-1})$. Similarly, the update of V is

$$V^{k+1} = \frac{\beta\|U^\top U\|}{\beta\|U^\top U\| + \lambda_t}V_{ex} - \frac{1}{\beta\|U^\top U\| + \lambda_t}U^\top w - \frac{\beta U^\top(UV_{ex} - W)}{\beta\|U^\top U\| + \lambda_t},$$

where $U = U^{k+1}$, $W = W^k$, $\omega = \omega^k$, and $V_{ex} = V^k + \alpha_V^k(V^k - V^{k-1})$. The update of W in (26) is

$$W^{k+1} = \frac{1}{\beta + L_G}\left(L_G W^k - \nabla G(W^k) + \omega^k + \beta U^{k+1}V^{k+1}\right),$$

and the update of ω is

$$\omega^{k+1} = \tau_1\omega^k + \tau_2\beta(U^{k+1}V^{k+1} - W^{k+1}).$$

We choose the following extrapolation parameters satisfying $\gamma_i^k \le B_1\eta_i^{k-1}$:

$$\begin{aligned} B_1 &= 0.9999,\ t_0 = 1,\ t_k = \tfrac{1}{2}(1 + \sqrt{1 + 4t_{k-1}^2}), \\ \alpha_U^k &= \min\left\{\tfrac{t_{k-1}-1}{t_k}, B_1\sqrt{\tfrac{\|V^{k-1}(V^{k-1})^\top\|}{\|V^k(V^k)^\top\|}}\right\}, \\ \alpha_V^k &= \min\left\{\tfrac{t_{k-1}-1}{t_k}, B_1\sqrt{\tfrac{\|(U^{k-1})^\top U^{k-1}\|}{\|(U^k)^\top U^k\|}}\right\}. \end{aligned}$$

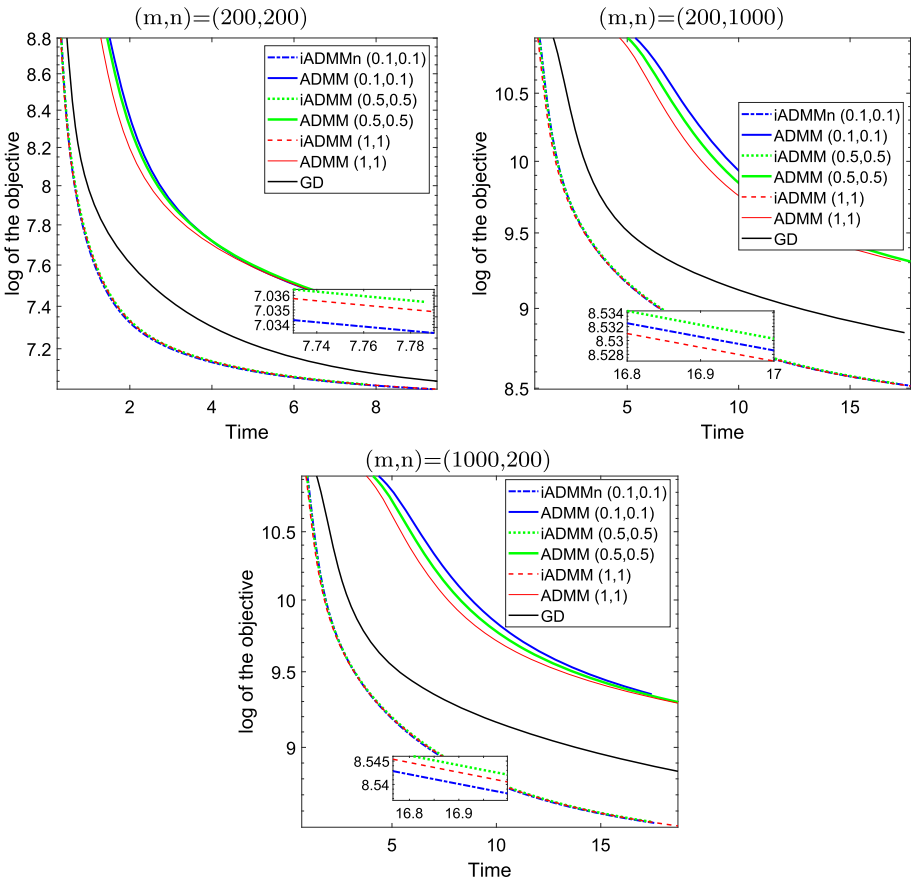

Fig. 1 Evolution of the log of mean of the objective values with respect to time

To generate a sparse data $Y \in \mathbb{R}^{m\times n}$ with each element $y_{ij} \in \{0, 1\}$, we use Matlab command `sprand(m,n,s);` and assign `Y(Y>0)=1;` In the upcoming experiments, we set $s = 0.1$ (that is, 90% of the elements of Y are 0), $r = 100, c = 1, \lambda_d = \lambda_t = \frac{1}{4}$, and $\beta = 1$. For each size $(m, n) \in \{(200, 200), (200, 1000), (1000, 200)\}$, we generate 5 random sparse matrices Y. And for each Y, we generate 5 random initial points. We run each algorithm with the same initial point and the same running time: 10 s for the size $(m, n) = (200, 200)$ and 25 s for the size $(m, n) \in \{(200, 1000), (1000, 200)\}$.

We compare iADMMn with GD - the alternating gradient descent method, which alternatively updates U and V by gradient descent step, and implement three iADMMn versions corresponding to $(\tau_1, \tau_2) \in \{(0.1, 0.1), (0.5, 0.5), (1, 1)\}$, together with their non-inertial versions. We compute mean of the objective values of Problem (3) over 25 trials (5 random datasets and 5 random initial points) and report the evolution of their log with respect to time in Fig. 1. We also report the mean $\pm$ std of the final objective values in Table 2.

We observe from Fig. 1 and Table 2 that iADMMn consistently outperforms its non-inertial version and the alternating gradient descent method. iADMMn with $\tau_1 = \tau_2 = 0.1$ more often produces better final objective values than the other two iADMMn variants with $\tau_1 = \tau_2 = 0.5$ and $\tau_1 = \tau_2 = 1$.

Table 2 Mean (and std) of the final objective values of Problem (3) over 25 trials

Algorithm / (m, n)	(200x200)	(200,1000)	(1000,200)
iADMMn (0.1,0.1)	$\mathbf{1.106 \times 10^3}$ (2.977)	4.807×10^3 (50.639)	$\mathbf{4.846 \times 10^3}$ (68.128)
ADMMn (0.1,0.1)	1.469×10^3 (15.751)	1.035×10^4 (427.431)	1.052×10^4 (271.288)
iADMMn (0.5,0.5)	1.108×10^3 (5.994)	4.813×10^3 (56.740)	4.864×10^3 (59.246)
ADMMn (0.5,0.5)	1.477×10^3 (28.267)	1.020×10^4 (178.253)	1.047×10^4 (124.764)
iADMMn (1,1)	1.107×10^3 (2.430)	$\mathbf{4.800 \times 10^3}$ (24.680)	4.855×10^3 (30.969)
ADMMn (1,1)	1.476×10^3 (11.517)	1.013×10^4 (204.938)	1.042×10^4 (122.554)
GD	1.144×10^3 (3.476)	6.491×10^3 (102.308)	6.725×10^3 (44.983)

The best means are highlighted in bold

5 Conclusion

We have analyzed iADMMn, an inertial alternating direction method of multipliers for solving Problem (1). In essence, iADMMn is an extension of iADMM proposed in [13]: iADMMn allows the nonlinearity of the coupling constraint whereas iADMM only considers linear coupling constraint. iADMMn is also an extension of [15, Algorithm 2]: iADMMn allows inertial term in the updating of x_i whereas [15, Algorithm 2] does not. On the other hand, iADMMn considers a non-standard update rule for the multiplier ω when it allows a scaling factor $\tau_1 \in (0, 1]$ for ω^k whereas both iADMM and [15, Algorithm 2] use standard rule $\tau_1 = 1$. Theorem 1 shows that every limit point of the generated sequence is an ε-stationary point of Problem (1). In Theorem 2, $\tau_1 = 1$ is required to guarantee a global convergence for the generated sequence. However, when $\tau_1 \neq 1$, if there exist many limit points of the generated sequence or not is still an open question for us. We take it as a future research direction.

Acknowledgements We express our sincere appreciation to the reviewers for their comments, which greatly helped improve the paper.

Data availability The manuscript has synthetic data.

References

1. Ahookhosh, M., Hien, L.T.K., Gillis, N., Patrinos, P.: Multi-block Bregman proximal alternating linearized minimization and its application to sparse orthogonal nonnegative matrix factorization. Comput. Optim. Appl. **79**, 681–715 (2021)
2. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1), 5–16 (2009)
3. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss-Seidel methods. Math. Program. **137**(1), 91–129 (2013)
4. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM J. Optim. **23**, 2037–2060 (2013)
5. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Nashua (2016)
6. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1), 459–494 (2014)
7. Bolte, J., Sabach, S., Teboulle, M.: Nonconvex Lagrangian-based optimization: monitoring schemes and global convergence. Math. Oper. Res. **43**(4), 1210–1232 (2018)

8. Bot, R.I., Nguyen, D.K.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. Math. Oper. Res. **45**(2), 682–712 (2020)
9. Cohen, E., Hallak, N., Teboulle, M.: A dynamic alternating direction of multipliers for nonconvex minimization with nonlinear functional equality constraints. J. Optim. Theory Appl. **193**, 324–353 (2022)
10. Duchi, J.C., Ruan, F.: Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. Inf. Inference J. IMA **8**(3), 471–529 (2018)
11. Dutter, R., Huber, P.J.: Numerical methods for the nonlinear robust regression problem. J. Stat. Comput. Simul. **13**(2), 79–113 (1981)
12. Goodfellow, I.J., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016). http://www.deeplearningbook.org
13. Hien, L., Phan, D., Gillis, N.: Inertial alternating direction method of multipliers for non-convex non-smooth optimization. Comput. Optim. Appl. **83**, 247–285 (2022)
14. Hien, L.T.K., Gillis, N., Patrinos, P.: Inertial block proximal method for non-convex non-smooth optimization. In: Thirty-seventh International Conference on Machine Learning ICML 2020 (2020)
15. Hien, L.T.K., Papadimitriou, D.: Multiblock ADMM for nonsmooth nonconvex optimization with nonlinear coupling constraints (2022). ArXiv:2201.07657
16. Hien, L.T.K., Phan, D.N., Gillis, N.: An inertial block majorization minimization framework for nonsmooth nonconvex optimization. J. Mach. Learn. Res. **24**(18), 1–41 (2023)
17. Hien, L.T.K., Phan, D.N., Gillis, N., Ahookhosh, M., Patrinos, P.: Block Bregman majorization minimization with extrapolation. SIAM J. Math. Data Sci. **4**(1), 1–25 (2022)
18. Hong, M., Chang, T.H., Wang, X., Razaviyayn, M., Ma, S., Luo, Z.Q.: A block successive upper-bound minimization method of multipliers for linearly constrained convex optimization. Math. Oper. Res. **45**(3), 833–861 (2020)
19. Liu, Y., Wu, M., Miao, C., Zhao, P., Li, X.L.: Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput. Biol. **12**(2), 1–26 (2016)
20. Maillard, S., Roncalli, T., Teiletche, J.: The properties of equally weighted risk contribution portfolios. J. Portf. Manage. **36**(4), 60–70 (2010)
21. Mairal, J.: Optimization with first-order surrogate functions. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Vol. 28, ICML'13, pp. 783–791. JMLR.org (2013)
22. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
23. Melo, J.G., Monteiro, R.D.C.: Iteration-complexity of a Jacobi-type non-Euclidean ADMM for multi-block linearly constrained nonconvex programs (2017). ArXiv:1705.07229
24. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Dokl. **269**(3), 543 (1983)
25. Nesterov, Y.: On an approach to the construction of optimal methods of minimization of smooth convex functions. Ekonom. i. Mat. Metody **24**, 509–517 (1998)
26. Nesterov, Y.: Introductory lectures on convex optimization: a basic course. Kluwer Academic Publ. (2004)
27. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Prog. **103**(1), 127–152 (2005)
28. Ochs, P.: Unifying abstract inexact convergence theorems and block coordinate variable metric ipiano. SIAM J. Optim. **29**(1), 541–570 (2019)
29. Ochs, P., Chen, Y., Brox, T., Pock, T.: iPiano: inertial proximal algorithm for nonconvex optimization. SIAM J. Imag. Sci. **7**(2), 1388–1419 (2014)
30. Pock, T., Sabach, S.: Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM J. Imag. Sci. **9**(4), 1756–1787 (2016)
31. Polyak, B.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
32. Razaviyayn, M., Hong, M., Luo, Z.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Optim. **23**(2), 1126–1153 (2013)
33. Roosta-Khorasani, F., van den Doel, K., Ascher, U.: Stochastic algorithms for inverse problems involving pdes and many measurements. SIAM J. Sci. Comput. **36**(5), S3–S22 (2014)
34. Sun, K., Sun, A.: Dual descent ALM and ADMM (2021). ArXiv:2109.13214
35. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**(1), 387–423 (2009)
36. Wang, J., Zhao, L.: Nonconvex generalization of alternating direction method of multipliers for nonlinear equality constrained problems. Results Control Optim. **2**, 100009 (2021)
37. Wang, Y., Yin, W., Zeng, J.: Global convergence of ADMM in nonconvex nonsmooth optimization. J. Sci. Comput. **78**, 29–63 (2019)

38. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imag. Sci. **6**(3), 1758–1789 (2013)
39. Xu, Y., Yin, W.: A globally convergent algorithm for nonconvex optimization based on block coordinate update. J. Sci. Comput. **72**(2), 700–734 (2017)
40. Yang, Y., Jia, Q.S., Xu, Z., Guan, X., Spanos, C.J.: Proximal ADMM for nonconvex and nonsmooth optimization. Automatica **146**, 110551 (2022)
41. Yashtini, M.: Convergence and rate analysis of a proximal linearized ADMM for nonconvex nonsmooth optimization. J. Glob. Optim. **84**, 913–939 (2022)
42. Zavriev, S., Kostyuk, F.: Heavy-ball method in nonconvex optimization problems. Comput. Math. Model. **4**, 336–341 (1993)