Math. Program., Ser. A (2019) 173:37–77
https://doi.org/10.1007/s10107-017-1205-9

FULL LENGTH PAPER

Extended ADMM and BCD for nonseparable convex minimization models with quadratic coupling terms: convergence analysis and insights

Caihua Chen[1] · Min Li[1] · Xin Liu[2,3] ·
Yinyu Ye[1,4]

Received: 27 December 2015 / Accepted: 20 October 2017 / Published online: 8 November 2017

Abstract In this paper, we establish the convergence of the proximal alternating direction method of multipliers (ADMM) and block coordinate descent (BCD) method for nonseparable minimization models with quadratic coupling terms. The novel convergence results presented in this paper answer several open questions that have been the subject of considerable discussion. We firstly extend the 2-block proximal ADMM to linearly constrained convex optimization with a coupled quadratic objective function, an area where theoretical understanding is currently lacking, and prove that the sequence generated by the proximal ADMM converges in point-wise manner to a primal-dual solution pair. Moreover, we apply randomly permuted ADMM (RPADMM) to nonseparable multi-block convex optimization, and prove its expected convergence for a class of nonseparable quadratic programming problems. When the

✉ Xin Liu
liuxin@lsec.cc.ac.cn

Caihua Chen
chchen@nju.edu.cn

Min Li
limin@nju.edu.cn

Yinyu Ye
yyye@stanford.edu

[1] International Center of Management Science and Engineering, School of Management and Engineering, Nanjing University, Nanjing, China

[2] State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

[3] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

[4] Department of Management Science and Engineering, School of Engineering, Stanford University, Stanford, CA, USA

linear constraint vanishes, the 2-block proximal ADMM and RPADMM reduce to the 2-block cyclic proximal BCD method and randomly permuted BCD (RPBCD). Our study provides the first iterate convergence result for 2-block cyclic proximal BCD without assuming the boundedness of the iterates. We also theoretically establish the expected iterate convergence result concerning multi-block RPBCD for convex quadratic optimization. In addition, we demonstrate that RPBCD may have a worse convergence rate than cyclic proximal BCD for 2-block convex quadratic minimization problems. Although the results on RPADMM and RPBCD are restricted to quadratic minimization models, they provide some interesting insights: (1) random permutation makes ADMM and BCD more robust for multi-block convex minimization problems; (2) cyclic BCD may outperform RPBCD for "nice" problems, and RPBCD should be applied with caution when solving general convex optimization problems especially with a few blocks.

Keywords Nonseparable convex minimization · Alternating direction method of multipliers · Block coordinate descent method · Iterate convergence · Random permutation

Mathematics Subject Classification 65K05 · 90C26

1 Introduction

In this paper we consider the linearly constrained convex minimization model with an objective function that is the sum of several separable functions and a coupled quadratic function:

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \ & \theta(x) := \sum_{i=1}^{n} \theta_i(x_i) + \frac{1}{2}x^\top Hx + g^\top x \\ \text{s.t.} \ & \sum_{i=1}^{n} A_i x_i = b, \end{aligned} \tag{1}$$

where $\theta_i : \mathbb{R}^{d_i} \mapsto (-\infty, +\infty]$ $(i = 1, 2, \ldots, n)$ are closed proper convex (not necessarily smooth) functions; $x_i \in \mathbb{R}^{d_i}$, $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^d$; $H \in \mathbb{R}^{d\times d}$ is a symmetric and positive semidefinite matrix; $g \in \mathbb{R}^d$; $A_i \in \mathbb{R}^{m\times d_i}$ and $b \in \mathbb{R}^m$. A point $(\bar{x}, \bar{\mu})$ is said to be a Karush-Kuhn-Tucker (KKT) point of (1) if it satisfies

$$\begin{cases} -(H\bar{x} + g)_i + A_i^\top \bar{\mu} \in \partial\theta_i(\bar{x}_i), \quad i = 1, \cdots, n, \\ \sum_{i=1}^{n} A_i\bar{x}_i = b. \end{cases} \tag{2}$$

The set consisting of the KKT points of (1) is assumed to be nonempty. Problem (1) has many applications in signal and imaging processing, machine learning, statistics, and engineering; e.g., see [1,14,19,29,41,42].

The augmented Lagrangian function of (1) is

$$\mathcal{L}_\beta(x_1,\ldots,x_n;\mu) := \sum_{i=1}^n \theta_i(x_i) + \frac{1}{2}x^\top Hx + g^\top x - \mu^\top \left(\sum_{i=1}^n A_i x_i - b\right) + \frac{\beta}{2}\Big\| \sum_{i=1}^n A_i x_i - b\Big\|^2, \tag{3}$$

where $\mu \in \mathbb{R}^m$ is the Lagrangian multiplier and $\beta > 0$ is the penalty parameter. In this paper, we extend the n-block proximal alternating direction method of multipliers (ADMM) to solve the nonseparable convex minimization problem (1), which consists of a cyclic update of the primal variables x_i $(i = 1, 2, \ldots, n)$ in the Gauss-Seidel fashion and a dual ascent type update of μ at each iteration, i.e.,

$$\begin{cases} x_1^{k+1} := \arg\min\limits_{x_1\in\mathbb{R}^{d_1}} \Big\{\mathcal{L}_\beta(x_1, x_2^k, \ldots, x_n^k; \mu^k) + \frac{1}{2}\|x_1 - x_1^k\|_{R_1}^2\Big\}, \\ x_2^{k+1} := \arg\min\limits_{x_2\in\mathbb{R}^{d_2}} \Big\{\mathcal{L}_\beta(x_1^{k+1}, x_2, x_3^k, \ldots, x_n^k; \mu^k) + \frac{1}{2}\|x_2 - x_2^k\|_{R_2}^2\Big\}, \\ \cdots\cdots \\ x_n^{k+1} := \arg\min\limits_{x_n\in\mathbb{R}^{d_n}} \Big\{\mathcal{L}_\beta(x_1^{k+1}, x_2^{k+1}, \ldots, x_{n-1}^{k+1}, x_n; \mu^k) + \frac{1}{2}\|x_n - x_n^k\|_{R_n}^2\Big\}, \\ \mu^{k+1} := \mu^k - \beta(\sum\limits_{i=1}^n A_i x_i^{k+1} - b), \end{cases} \tag{4}$$

where $R_i \in \mathbb{R}^{d_i\times d_i}$, $i = 1, \cdots, n$, are symmetric and positive semidefinite matrices.

Note that the algorithmic scheme (4) reduces to the classical ADMM when there are only two blocks $(n = 2)$, the coupled objective vanishes $(H = 0$ and $g = 0)$ and $R_i = 0$ $(i = 1, 2)$. ADMM was originally introduced in the early 1970s [20,23], and its convergence propertites have been studied extensively in the literature [6,15,17,18,22,28,40]. Because of its wide versatility and applicability in multiple fields, ADMM is a popular means of solving optimization problems, especially those related to big data; we refer to[8] for a survey on the modern applications of ADMM.

For the case of $n \geq 3$, numerous research efforts have been devoted to analyzing the convergence of multi-block ADMM and its variants for the linearly constrained separable convex optimization model, i.e., (1) without the coupled term. Recent work [10] has shown that the n-block ADMM (4) is not necessarily convergent, even for a nonsingular square system of linear equations. Various methods have been proposed to overcome the divergence issue of multi-block ADMM. One typical solution is to combine correction steps with the output of n-block ADMM (4) [25–27]. If at least $n - 2$ functions in the objective are strongly convex, it has been shown that (4) is globally convergent, provided that the penalty parameter β is restricted to a specific range [9,11,24,33,38,52]. Without strong convexity, it has been shown in [30] that the n-block ADMM with a small dual stepsize, where the multiplier update (4) is replaced by

$$\mu^{k+1} = \mu^k - \tau\beta\left(\sum_{i=1}^n A_i x_i^{k+1} - b\right),$$

is linearly convergent provided that the objective function satisfies certain error bound conditions. Some very recent studies [36,37] have demonstrated the convergence of multi-block ADMM under some other conditions, and some convergent proximal variants of the multi-block ADMM have been proposed for solving convex linear/quadratic conic programming problems [13,35,47]. A recent paper [48] proposed a randomly modified variant of the multi-block ADMM (4), called randomly permuted ADMM (RPADMM). At each step, RPADMM forms a random permutation of $\{1, 2, \ldots, n\}$ (known as block sampling without replacement), and updates the primal variables x_i ($i = 1, 2, \ldots, n$) in the order of the chosen permutation followed by the regular multiplier update. Surprisingly, RPADMM is convergent in expectation for any nonsingular square system of linear equations [48].

In contrast to the separable case, studies on the convergence properties of n-block ADMM for (1) with nonseparable objective, even for $n = 2$, are limited. In [29], the authors demonstrated that when problem (1) is convex but not necessarily separable[1], and certain error bound conditions are satisfied, the ADMM iteration converges to some primal-dual optimal solution, provided that the stepsize in the update of the multiplier is sufficiently small. Despite this conservative nature, the stepsize usually depends on some unknown parameters associated with the error bound, and may thus be difficult to compute, which often makes the algorithm less efficient. In view of this, it might be more beneficial to employ the classical ADMM (4) (with $\tau = 1$) or its variants with a large stepsize $\tau \geq 1$. However, as mentioned in [31], "when the objective function is not separable across the variables, the convergence of the ADMM (4) is still open, even in the case where $n = 2$ and $\theta(\cdot)$ is convex." Along slightly different lines, [14] investigated the convergence of a majorized ADMM for the convex optimization problem with a coupled smooth objective function, which includes the 2-block ADMM (4) for (1) as a special case. Convergence was established for the case when the subproblems of the ADMM admit unique solutions and H, A_1, A_2, R_1 and R_2 satisfy some additional restrictions; see Remark 4.2 in [14] for details. Very recently, [21] studied the convergence and ergodic complexity of a 2-block proximal ADMM and its variants for the nonseparable convex optimization by assuming some additional conditions on the problem data. As the positive definite proximal terms are indispensable in the analysis of these algorithms, the results derived in [21] are not applicable to the scheme (4) for problem (1) since R_1 and R_2 are only positive semidefinite.

In this paper, we analyze the iterate convergence of proximal ADMM (4) and the randomly permuted ADMM for solving the nonseparable convex optimization problem (1). The main contributions of our paper are threefold. Firstly, we prove that the 2-block proximal ADMM is convergent for (1) only under a condition that ensures the subproblems have unique solutions. Our condition is the weakest to ensure iterate convergence for the proximal ADMM since, as we will see in Sect. 2, it is not only sufficient but also necessary for the convergence of the proximal ADMM applied to some special problems. Our analysis partially answers the open question mentioned in

[1] The models considered in [29,31] are more general than problem (1), as the authors of [29,31] actually allow generally nonseparable smooth function in the objective, but in (1) the coupled objective is a quadratic function.

[31] on the convergence of ADMM for nonseparable convex optimization problems. Secondly, we extend the RPADMM proposed in [48] to solve the model (1), and prove its expected convergence in the case where $\theta_i \equiv 0\,(i = 1, 2, \ldots, n)$. This result is a non-trivial extension of the convergence result shown in [48], since the objective in (1) is more general and its solution set may not be a singleton. Thirdly, when restricted to the unconstrained case, that is, A_i $(i = 1, \cdots, n)$ and b are absent, the proximal ADMM and RPADMM reduce to the cyclic proximal block coordinate descent (BCD) method (also known as the alternating minimization method), i.e.,

$$\begin{cases} x_1^{k+1} := \underset{x_1 \in \mathbb{R}^{d_1}}{\arg\min} \Big\{ \theta(x_1, x_2^k, \ldots, x_n^k) + \frac{1}{2}\|x_1 - x_1^k\|_{R_1}^2 \Big\}, \\ x_2^{k+1} := \underset{x_2 \in \mathbb{R}^{d_2}}{\arg\min} \Big\{ \theta(x_1^{k+1}, x_2, x_3^k, \ldots, x_n^k) + \frac{1}{2}\|x_2 - x_2^k\|_{R_2}^2 \Big\}, \\ \cdots\cdots \\ x_n^{k+1} := \underset{x_n \in \mathbb{R}^{d_n}}{\arg\min} \Big\{ \theta(x_1^{k+1}, x_2^{k+1}, \ldots, x_{n-1}^{k+1}, x_n) + \frac{1}{2}\|x_n - x_n^k\|_{R_n}^2 \Big\}. \end{cases} \tag{5}$$

and randomly permuted BCD. An implication of our work is the iterate convergence of the 2-block cyclic proximal BCD method for the whole sequence and, in particular, the expected convergence of randomly permuted multi-block BCD. Although the literature on BCD-type methods is vast (e.g., [3–5,39,43,45,46,49,50]), there are very few results on the iterate convergence of BCD-type methods. As mentioned in [7], "in all these works [on BCD or its proximal variants] only convergence of the subsequences can be established." By assuming that the Kurdyka-Łojasiewicz property holds on the objective function and the iterates are bounded, [2] and [7] established the iterate convergence of the proximal BCD and proximal alternating linearized minimization, respectively. It is clear that these results are also applicable to the BCD type methods for convex minimization problems. While the boundedness assumption of the sequence are typical to establish the iterate convergence of algorithms for nonconvex optimization problems, it might be a bit restrictive to assume the boundedness for analyzing the iterate convergence for the convex cases. To the best of our knowledge, our convergence result for the 2-block proximal BCD method is the first for the proximal BCD that only requires the unique solutions-type condition of the subproblems, rather than any assumptions on the boundedness of the iterates.

It has been claimed that randomly permuted BCD (RPBCD, also known as the "sampling without replacement" variant of randomized BCD, and called "EPOCHS" in a recent survey [51]) tends to converge faster than the randomized BCD [51] , with the classical cyclic version performing even worse. Some numerical advantages of RPBCD compared with randomized BCD and cyclic BCD were discussed in [45]. In fact, it has been stated that "this kind of randomization [RPBCD] has been shown in several contexts to be superior to the sampling with replacement scheme analyzed above, but a theoretical understanding of this phenomenon remains elusive" [51]. Randomized BCD ("sampling with replacemen") has already been extensively studied [44], but its theoretical analysis does not apply to RPBCD. Although the function value convergence results [4,32,49] for cyclic or essential cyclic BCD can be simply extended to RPBCD, these analysis techniques are independent of permutation, so there remains a lack of direct theoretical analysis on the iterate convergence of RPBCD. Our expected iterate convergence of RPBCD for quadratic minimization problems can be regarded as the first direct analysis on the iterate convergence of the "sampling without replacement" variant of randomized BCD. We also prove that 2-block RPBCD may have a worse convergence rate than 2-block cyclic BCD for quadratic minimization problems. Thus, RPBCD should be used with caution for solving convex optimization problems with a few blocks.

The rest of this paper is organized as follows. In Sect. 2 , we prove the iterate convergence of the 2-block proximal ADMM and cyclic BCD for linearly constrained optimization problems with a coupled quadratic objective function (1) and its unconstrained variant, respectively. Section 3 illustrates the expected convergence of the RPADMM and the RPBCD for a class of linear constrained quadratic optimization problems and its unconstrained variant, respectively. Finally, we conclude our paper and present some insights into the use of ADMM and BCD in Sect. 4.

2 Convergence of 2-block proximal ADMM

In this section, we will specify $n = 2$ and analyze the iterate convergence of the 2-block proximal ADMM for the convex optimization model (1). For notational simplicity, we write

$$H := \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^\top & H_{22} \end{bmatrix}, \qquad R := \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \qquad \text{and} \qquad g := \begin{bmatrix} g_1 \\ g_2 \end{bmatrix},$$

and define the quadratic function $\phi(x_1, x_2)$ by

$$\phi(x_1, x_2) := \frac{1}{2} x_1^\top H_{11} x_1 + x_1^\top H_{12} x_2 + \frac{1}{2} x_2^\top H_{22} x_2 + g_1^\top x_1 + g_2^\top x_2. \tag{6}$$

Thus the problem under consideration can be written as

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^d} & \theta(x) := \theta_1(x_1) + \theta_2(x_2) + \phi(x_1, x_2) \\ \text{s.t.} & A_1 x_1 + A_2 x_2 = b. \end{array} \tag{7}$$

Since θ_1 and θ_2 are closed convex functions, there exist two symmetric positive semidefinite matrices Σ_1 and Σ_2 such that

$$(x_1 - \hat{x}_1)^\top (w_1 - \hat{w}_1) \geq \|x_1 - \hat{x}_1\|_{\Sigma_1}^2, \quad \forall\, x_1, \hat{x}_1 \in \text{dom}(\theta_1), w_1 \in \partial\theta_1(x_1), \hat{w}_1 \in \partial\theta_1(\hat{x}_1) \tag{8}$$

and

$$(x_2 - \hat{x}_2)^\top (w_2 - \hat{w}_2) \geq \|x_2 - \hat{x}_2\|_{\Sigma_2}^2, \quad \forall\, x_2, \hat{x}_2 \in \text{dom}(\theta_2),\ w_2 \in \partial\theta_2(x_2), \hat{w}_2 \in \partial\theta_2(\hat{x}_2), \tag{9}$$

where $\partial\theta_1$ and $\partial\theta_2$ are the subdifferential mappings of θ_1 and θ_2, respectively. By letting

$$x := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \hat{x} := \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}, \quad w := \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad \hat{w} := \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} \text{ and } \Sigma := \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{10}$$

we have

$$(x - \hat{x})^\top (w - \hat{w}) \geq \|x - \hat{x}\|_\Sigma^2. \tag{11}$$

The following lemma establishes the contraction property with respect to the solution set of (7) for the sequence generated by (4), which plays an important role in the subsequent analysis.

Lemma 1 *Assume the* 2*-block proximal ADMM* (4) *is well defined for problem* (7). *Let* $\{(x_1^k, x_2^k, \mu^k)\}$ *be the sequence generated by* (4). *Then, the following statements hold.*

(i) *If* $(\bar{x}_1, \bar{x}_2, \bar{\mu})$ *is a given KKT point of problem* (7)*, then we have*

$$\begin{aligned}
&\Big(\frac{7}{8}\|x^k - \bar{x}\|^2_{H+\Sigma+\frac{4}{7}R} + \frac{1}{2}\|x_2^k - \bar{x}_2\|^2_{H_{22}+\Sigma_2+\beta A_2^\top A_2} \\
&+\frac{1}{2\beta}\|\mu^k - \bar{\mu}\|^2 + \frac{1}{2}\|x_2^k - x_2^{k-1}\|^2_{R_2}\Big) \\
&-\Big(\frac{7}{8}\|x^{k+1} - \bar{x}\|^2_{H+\Sigma+\frac{4}{7}R} + \frac{1}{2}\|x_2^{k+1} - \bar{x}_2\|^2_{H_{22}+\Sigma_2+\beta A_2^\top A_2} \\
&+\frac{1}{2\beta}\|\mu^{k+1} - \bar{\mu}\|^2 + \frac{1}{2}\|x_2^{k+1} - x_2^k\|^2_{R_2}\Big) \\
&\geq \frac{1}{16}\|x^{k+1} - x^k\|^2_{H+\Sigma+8R} + \frac{1}{6}\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2+3\beta A_2^\top A_2} + \frac{1}{2\beta}\|\mu^{k+1} - \mu^k\|^2.
\end{aligned} \tag{12}$$

(ii) *It holds that*

$$\begin{cases}
\lim\limits_{k\to\infty} d(0,\ \partial\theta_1(x_1^{k+1}) + \nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) - A_1^\top \mu^{k+1}) = 0, \\
\lim\limits_{k\to\infty} d(0,\ \partial\theta_2(x_2^{k+1}) + \nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) - A_2^\top \mu^{k+1}) = 0, \\
\lim\limits_{k\to\infty} \|A_1 x_1^{k+1} + A_2 x_2^{k+1} - b\| = 0,
\end{cases} \tag{13}$$

where $d(\cdot,\cdot)$ *denotes the Euclidean distance of some point to a set.*

Proof (i) From the first order optimality condition of (4), we get

$$\begin{cases}
0 \in \partial\theta_1(x_1^{k+1}) + \nabla_{x_1}\phi(x_1^{k+1}, x_2^k) - A_1^\top \mu^k \\
\quad +\beta A_1^\top (A_1 x_1^{k+1} + A_2 x_2^k - b) + R_1(x_1^{k+1} - x_1^k), \\
0 \in \partial\theta_2(x_2^{k+1}) + \nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) \\
\quad -A_2^\top \mu^k + \beta A_2^\top (A_1 x_1^{k+1} + A_2 x_2^{k+1} - b) + R_2(x_2^{k+1} - x_2^k),
\end{cases}$$

where $\phi(\cdot,\cdot)$ is defined in (6). Using the definitions of ϕ and μ^{k+1}, the above formulas imply that

$$\begin{cases} -\nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) + A_1^\top \mu^{k+1} + (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) \\ \quad -R_1(x_1^{k+1} - x_1^k) \in \partial\theta_1(x_1^{k+1}), \\ -\nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) + A_2^\top \mu^{k+1} - R_2(x_2^{k+1} - x_2^k) \in \partial\theta_2(x_2^{k+1}). \end{cases} \tag{14}$$

Since $(\bar{x}_1, \bar{x}_2, \bar{\mu})$ is a KKT point of problem (7), we have that

$$\begin{cases} -\nabla_{x_1}\phi(\bar{x}_1, \bar{x}_2) + A_1^\top \bar{\mu} \in \partial\theta_1(\bar{x}_1), \\ -\nabla_{x_2}\phi(\bar{x}_1, \bar{x}_2) + A_2^\top \bar{\mu} \in \partial\theta_2(\bar{x}_2), \\ A_1\bar{x}_1 + A_2\bar{x}_2 = b. \end{cases} \tag{15}$$

From (11), (14) and (15), we obtain

$$\begin{aligned} &\|x^{k+1} - \bar{x}\|_\Sigma^2 \\ &\le (x_1^{k+1} - \bar{x}_1)^\top \Big\{\big[-\nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) + A_1^\top \mu^{k+1} + (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) \\ &\qquad -R_1(x_1^{k+1} - x_1^k)\big] - \big[-\nabla_{x_1}\phi(\bar{x}_1, \bar{x}_2) + A_1^\top \bar{\mu}\big]\Big\} \\ &\qquad +(x_2^{k+1} - \bar{x}_2)^\top \Big\{\big[-\nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) + A_2^\top \mu^{k+1} - R_2(x_2^{k+1} - x_2^k)\big] \\ &\qquad -\big[-\nabla_{x_2}\phi(\bar{x}_1, \bar{x}_2) + A_2^\top \bar{\mu}\big]\Big\} \\ &\quad = -(x_1^{k+1} - \bar{x}_1)^\top A_1^\top(\bar{\mu} - \mu^{k+1}) - (x_2^{k+1} - \bar{x}_2)^\top A_2^\top(\bar{\mu} - \mu^{k+1}) - (x^{k+1} - \bar{x})^T R(x^{k+1} - x^k) \\ &\qquad +(x_1^{k+1} - \bar{x}_1)^\top (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) - (x^{k+1} - \bar{x})^\top \big(\nabla\phi(x_1^{k+1}, x_2^{k+1}) \\ &\qquad -\nabla\phi(\bar{x}_1, \bar{x}_2)\big) \\ &= (x^{k+1} - x^k)^T R(\bar{x} - x^{k+1}) + \frac{1}{\beta}(\mu^{k+1} - \mu^k)^\top(\bar{\mu} - \mu^{k+1}) + (x_1^{k+1} - \bar{x}_1)^\top(H_{12} \\ &\qquad +\beta A_1^\top A_2)(x_2^{k+1} - x_2^k) - \|x^{k+1} - \bar{x}\|_H^2. \end{aligned} \tag{16}$$

By simple manipulations and using $A_1\bar{x}_1 + A_2\bar{x}_2 = b$, we can see that

$$\begin{aligned} &\beta(x_1^{k+1} - \bar{x}_1)^\top A_1^\top A_2(x_2^{k+1} - x_2^k) \\ &\quad = -\beta(A_2x_2^{k+1} - A_2\bar{x}_2)^\top(A_2x_2^{k+1} - A_2x_2^k) + \beta(A_1x_1^{k+1} + A_2x_2^{k+1} - b)^\top(A_2x_2^{k+1} - A_2x_2^k) \\ &\quad = \frac{\beta}{2}(\|A_2x_2^k - A_2\bar{x}_2\|^2 - \|A_2x_2^{k+1} - A_2\bar{x}_2\|^2) - \frac{\beta}{2}\|A_2x_2^{k+1} - A_2x_2^k\|^2 \\ &\qquad +\beta(A_1x_1^{k+1} + A_2x_2^{k+1} - b)^\top(A_2x_2^{k+1} - A_2x_2^k), \end{aligned} \tag{17}$$

$$(x^{k+1} - x^k)^T R(\bar{x} - x^{k+1}) = \frac{1}{2}(\|x^k - \bar{x}\|_R^2 - \|x^{k+1} - \bar{x}\|_R^2 - \|x^{k+1} - x^k\|_R^2) \tag{18}$$

and

$$\frac{1}{\beta}(\mu^{k+1} - \mu^k)^\top(\bar{\mu} - \mu^{k+1}) = \frac{1}{2\beta}(\|\mu^k - \bar{\mu}\|^2 - \|\mu^{k+1} - \bar{\mu}\|^2 - \|\mu^{k+1} - \mu^k\|^2). \tag{19}$$

On the other hand, it f ollows from (14) that

$$\begin{cases} -\nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) + A_2^\top \mu^{k+1} - R_2(x_2^{k+1} - x_2^k) \in \partial\theta_2(x_2^{k+1}), \\ -\nabla_{x_2}\phi(x_1^k, x_2^k) + A_2^\top \mu^k - R_2(x_2^k - x_2^{k-1}) \in \partial\theta_2(x_2^k), \end{cases}$$

which, together with (9), implies

$$\begin{aligned}(x_2^{k+1} - x_2^k)^\top\big[- \nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) \quad & +A_2^\top \mu^{k+1} - R_2(x_2^{k+1} - x_2^k) + \nabla_{x_2}\phi(x_1^k, x_2^k) \\ & -A_2^\top \mu^k + R_2(x_2^k - x_2^{k-1})\big] \geq \|x_2^{k+1} - x_2^k\|^2_{\Sigma_2}.\end{aligned} \tag{20}$$

Recall that

$$\mu^{k+1} - \mu^k = -\beta(A_1x_1^{k+1} + A_2x_2^{k+1} - b) \quad \text{and} \quad \nabla_{x_2}\phi(x_1, x_2) = H_{12}^\top x_1 + H_{22}x_2 + g_2.$$

Then, by using the Cauchy-Schwarz inequality, the inequality (20) gives

$$\begin{aligned}&\beta(A_1x_1^{k+1} + A_2x_2^{k+1} - b)^\top(A_2x_2^{k+1} - A_2x_2^k) \\ &\leq -\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2} + (x_2^{k+1} - x_2^k)^\top H_{12}^\top(x_1^k - x_1^{k+1}) - \|x_2^{k+1} - x_2^k\|^2_{R_2} \\ &\quad +(x_2^{k+1} - x_2^k)^T R_2(x_2^k - x_2^{k-1}) \\ &\leq -\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2} + (x_2^{k+1} - x_2^k)^\top H_{12}^\top(x_1^k - x_1^{k+1}) - \frac{1}{2}\|x_2^{k+1} - x_2^k\|^2_{R_2} \\ &\quad +\frac{1}{2}\|x_2^k - x_2^{k-1}\|^2_{R_2}.\end{aligned}$$

Substituting (17), (18), (19) and the above inequality into (16), we further get

$$\begin{aligned}&\frac{1}{2}\big(\|x^k - \bar{x}\|^2_R - \|x^{k+1} - \bar{x}\|^2_R\big) + \frac{1}{2\beta}\big(\|\mu^k - \bar{\mu}\|^2 - \|\mu^{k+1} - \bar{\mu}\|^2\big) + \frac{\beta}{2}\big(\|A_2x_2^k \\ &\quad -A_2\bar{x}_2\|^2 - \|A_2x_2^{k+1} - A_2\bar{x}_2\|^2\big) + \frac{1}{2}\big(\|x_2^k - x_2^{k-1}\|^2_{R_2} - \|x_2^{k+1} - x_2^k\|^2_{R_2}\big) \\ &\geq \|x^{k+1}-\bar{x}\|^2_{H+\Sigma}+\frac{1}{2}\|x^{k+1} - x^k\|^2_R + \frac{1}{2\beta}\|\mu^{k+1} - \mu^k\|^2 + \frac{1}{2}\|x_2^{k+1} - x_2^k\|^2_{\beta A_2^\top A_2} \\ &\quad -(x_2^{k+1} - x_2^k)^\top H_{12}^\top(x_1^k - \bar{x}_1) + \|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2}.\end{aligned} \tag{21}$$

Moreover, it follows from the Cauchy-Schwarz inequality and $H + \Sigma \succeq 0$ that

$$\begin{aligned}&(x_2^{k+1} - x_2^k)^\top H_{12}^\top(x_1^k - \bar{x}_1) - \|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2} \\ &\quad = (x_2^{k+1} - x_2^k)^\top H_{12}^\top(x_1^k - \bar{x}_1) + (x_2^{k+1} - x_2^k)^\top(H_{22} + \Sigma_2)(x_2^k - \bar{x}_2) \\ &\qquad -(x_2^{k+1} - x_2^k)^\top(H_{22} + \Sigma_2)(x_2^{k+1} - \bar{x}_2) \\ &\quad = \begin{bmatrix} 0 \\ x_2^{k+1} - x_2^k \end{bmatrix}^\top (H + \Sigma)(x^k - \bar{x}) - (x_2^{k+1} - x_2^k)^\top(H_{22} + \Sigma_2)(x_2^{k+1} - \bar{x}_2)\end{aligned}$$

$$
\begin{aligned}
&= \begin{bmatrix} 0 \\ x_2^{k+1} - x_2^k \end{bmatrix}^\top (H + \Sigma)(x^k - \bar{x}) - \frac{1}{2}\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2} \\
&\quad + \frac{1}{2}\big(\|x_2^k - \bar{x}_2\|^2_{H_{22}+\Sigma_2} - \|x_2^{k+1} - \bar{x}_2\|^2_{H_{22}+\Sigma_2}\big) \\
&\le \frac{3}{4}\|x^k - \bar{x}\|^2_{H+\Sigma} + \frac{1}{3}\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2} - \frac{1}{2}\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2} \\
&\quad + \frac{1}{2}\big(\|x_2^k - \bar{x}_2\|^2_{H_{22}+\Sigma_2} - \|x_2^{k+1} - \bar{x}_2\|^2_{H_{22}+\Sigma_2}\big) \\
&= \frac{3}{4}\|x^k - \bar{x}\|^2_{H+\Sigma} - \frac{1}{6}\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2} \\
&\quad + \frac{1}{2}\big(\|x_2^k - \bar{x}_2\|^2_{H_{22}+\Sigma_2} - \|x_2^{k+1} - \bar{x}_2\|^2_{H_{22}+\Sigma_2}\big),
\end{aligned} \tag{22}
$$

where the last inequality follows from the elementary inequality $a^T(H+\Sigma)b \le \frac{3}{4}\|a\|^2_{H+\Sigma} + \frac{1}{3}\|b\|^2_{H+\Sigma}$ for any $a, b \in \Re^d$ and the equality $\left\|\begin{matrix} 0 \\ x_2^{k+1} - x_2^k \end{matrix}\right\|^2_{H+\Sigma} = \|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2}$. Using the inequality $2(\|a\|^2_{H+\Sigma} + \|b\|^2_{H+\Sigma}) \ge \|a - b\|^2_{H+\Sigma}$, we obtain

$$
\begin{aligned}
&\|x^{k+1} - \bar{x}\|^2_{H+\Sigma} - \frac{3}{4}\|x^k - \bar{x}\|^2_{H+\Sigma} \\
&\quad = \frac{7}{8}\big(\|x^{k+1} - \bar{x}\|^2_{H+\Sigma} - \|x^k - \bar{x}\|^2_{H+\Sigma}\big) + \frac{1}{8}\big(\|x^{k+1} - \bar{x}\|^2_{H+\Sigma} + \|x^k - \bar{x}\|^2_{H+\Sigma}\big) \\
&\quad \ge \frac{7}{8}\big(\|x^{k+1} - \bar{x}\|^2_{H+\Sigma} - \|x^k - \bar{x}\|^2_{H+\Sigma}\big) + \frac{1}{16}\|x^{k+1} - x^k\|^2_{H+\Sigma}.
\end{aligned} \tag{23}
$$

Substituting (22) and (23) into (21), we get (12).
(ii) From (12), we can immediately see that

$$
\sum_{k=1}^{\infty}\Big(\frac{1}{16}\|x^{k+1} - x^k\|^2_{H+\Sigma+8R} + \frac{1}{6}\|x_2^{k+1} - x_2^k\|^2_{H_{22}+\Sigma_2+3\beta A_2^\top A_2} + \frac{1}{2\beta}\|\mu^{k+1} - \mu^k\|^2\Big) < \infty, \tag{24}
$$

and it therefore holds that

$$
\lim_{k\to\infty}\|x^{k+1} - x^k\|_{H+\Sigma+8R} = 0, \quad \lim_{k\to\infty}\|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2+3\beta A_2^\top A_2} = 0 \tag{25}
$$

and

$$
\lim_{k\to\infty}\|A_1 x_1^{k+1} + A_2 x_2^{k+1} - b\| = \lim_{k\to\infty}\frac{1}{\beta}\|\mu^{k+1} - \mu^k\| = 0. \tag{26}
$$

Since $H + \Sigma$, R and $H_{22} + \Sigma_2$ are positive semidefinite matrices, we deduce from (25) that

$$\begin{cases} \lim\limits_{k\to\infty} (H + \Sigma)\,(x^{k+1} - x^k) = 0, \\ \lim\limits_{k\to\infty} R\,(x^{k+1} - x^k) = 0, \\ \lim\limits_{k\to\infty} \|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2} = 0, \\ \lim\limits_{k\to\infty} \|A_2(x_2^{k+1} - x_2^k)\| = 0, \end{cases} \tag{27}$$

and hence

$$\lim_{k\to\infty} (H_{11} + \Sigma_1)(x_1^{k+1} - x_1^k) + H_{12}(x_2^{k+1} - x_2^k) = 0. \tag{28}$$

Using the triangle inequality, we have

$$\left\| \begin{bmatrix} x_1^{k+1} - x_1^k \\ 0 \end{bmatrix} \right\|_{H+\Sigma} \le \left\| \begin{bmatrix} x_1^{k+1} - x_1^k \\ x_2^{k+1} - x_2^k \end{bmatrix} \right\|_{H+\Sigma} + \left\| \begin{bmatrix} 0 \\ x_2^{k+1} - x_2^k \end{bmatrix} \right\|_{H+\Sigma},$$

and thus from (27), it follows

$$\lim_{k\to\infty} \|x_1^{k+1} - x_1^k\|_{H_{11}+\Sigma_1} \le \lim_{k\to\infty} \left(\|x^{k+1} - x^k\|_{H+\Sigma} + \|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2} \right) = 0.$$

From (27), (28) and the above formula, we obtain

$$\begin{cases} \lim\limits_{k\to\infty} R_1(x_1^{k+1} - x_1^k) = 0, \\ \lim\limits_{k\to\infty} R_2(x_2^{k+1} - x_2^k) = 0, \\ \lim\limits_{k\to\infty} H_{12}(x_2^{k+1} - x_2^k) = -\lim\limits_{k\to\infty} (H_{11} + \Sigma_1)(x_1^{k+1} - x_1^k) = 0, \\ \lim\limits_{k\to\infty} A_2(x_2^{k+1} - x_2^k) = 0. \end{cases} \tag{29}$$

This, together with (14) and (26), proves the assertion (13). □

To establish the convergence of ADMM, we make the following assumption:

Assumption 1 We assume

$$\begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} A_1^\top A_1 & 0 \\ 0 & A_2^\top A_2 \end{bmatrix} + \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \succ 0. \tag{30}$$

It is worth emphasizing that Assumption 1 means that the subproblems of 2-block proximal ADMM admit unique solutions, because Assumption 1 holds if and only if

$$\begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \beta \begin{bmatrix} A_1^\top A_1 & 0 \\ 0 & A_2^\top A_2 \end{bmatrix} + \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \succ 0$$

for any $\beta > 0$. However, the optimal solution to original problem (7) is not necessarily unique.

We are now ready to prove the iterate convergence of the 2-block proximal ADMM for the nonseparable convex optimization model (7).

Theorem 1 *Suppose Assumption 1 holds. Let $\{(x_1^k, x_2^k, \mu^k)\}$ be generated by the proximal ADMM (4) with $n = 2$ to solve problem (7). Then the sequence $\{(x_1^k, x_2^k, \mu^k)\}$ converges to a KKT point of problem (7).*

Proof It follows from (12) that the sequences $\{(H + \Sigma + R)x^{k+1}\}$, $\{(H_{22} + \Sigma_2 + \beta A_2^\top A_2 + R_2)x_2^{k+1}\}$ and $\{\mu^{k+1}\}$ are all bounded. Since $H_{22} + \Sigma_2 + \beta A_2^\top A_2 + R_2$ is positive definite, we know $\{x_2^{k+1}\}$ is bounded. Note that $A_1\bar{x}_1 + A_2\bar{x}_2 = b$. Using the triangle inequality

$$\begin{aligned}\|A_1(x_1^{k+1} - \bar{x}_1)\| &\le \|A_1x_1^{k+1} + A_2x_2^{k+1} - (A_1\bar{x}_1 + A_2\bar{x}_2)\| + \|A_2(x_2^{k+1} - \bar{x}_2)\|\\ &= \|A_1x_1^{k+1} + A_2x_2^{k+1} - b\| + \|A_2(x_2^{k+1} - \bar{x}_2)\|\\ &= \frac{1}{\beta}\|\mu^k - \mu^{k+1}\| + \|A_2(x_2^{k+1} - \bar{x}_2)\|\end{aligned}$$

and

$$\begin{aligned}\|x_1^{k+1} - \bar{x}_1\|_{H_{11}+\Sigma_1+R_1} &= \left\|\begin{bmatrix} x_1^{k+1} - \bar{x}_1 \\ 0 \end{bmatrix}\right\|_{H+\Sigma+R}\\ &\le \left\|\begin{bmatrix} x_1^{k+1} - \bar{x}_1 \\ x_2^{k+1} - \bar{x}_2 \end{bmatrix}\right\|_{H+\Sigma+R} + \left\|\begin{bmatrix} 0 \\ x_2^{k+1} - \bar{x}_2 \end{bmatrix}\right\|_{H+\Sigma+R}\\ &= \|x^{k+1} - \bar{x}\|_{H+\Sigma+R} + \|x_2^{k+1} - \bar{x}_2\|_{H_{22}+\Sigma_2+R_2},\end{aligned}$$

we further obtain the boundedness of the sequences $\{A_1x_1^{k+1}\}$ and $\{(H_{11} + \Sigma_1 + R_1)x_1^{k+1}\}$, and hence $\{(H_{11} + \Sigma_1 + \beta A_1^\top A_1 + R_1)x_1^{k+1}\}$ is bounded. Together with the positive definiteness of $H_{11} + \Sigma_1 + \beta A_1^\top A_1 + R_1$, this implies the boundedness of $\{x_1^{k+1}\}$. Thus, the sequence $\{(x_1^k, x_2^k, \mu^k)\}$ is bounded and there exists a triple $(x_1^\infty, x_2^\infty, \mu^\infty)$ and a subsequence $\{k_i\}$ such that

$$\lim_{i\to\infty} x_1^{k_i} = x_1^\infty, \qquad \lim_{i\to\infty} x_2^{k_i} = x_2^\infty \qquad \text{and} \qquad \lim_{i\to\infty} \mu^{k_i} = \mu^\infty.$$

Setting $k = k_i - 1$ and invoking the upper semicontinuity of $\partial\theta_1$ and $\partial\theta_2$ in (13), we then obtain

$$\begin{cases} -\nabla_{x_1}\phi(x_1^\infty, x_2^\infty) + A_1^\top\mu^\infty \in \partial\theta_1(x_1^\infty),\\ -\nabla_{x_2}\phi(x_1^\infty, x_2^\infty) + A_2^\top\mu^\infty \in \partial\theta_2(x_2^\infty),\\ A_1x_1^\infty + A_2x_2^\infty - b = 0,\end{cases}$$

which means $(x_1^\infty, x_2^\infty, \mu^\infty)$ is a KKT point of problem (7). Hence (12) is also valid if $(\bar{x}_1, \bar{x}_2, \bar{\mu})$ is replaced by $(x_1^\infty, x_2^\infty, \mu^\infty)$. Therefore, it holds for any $k \ge k_i$ that

$$\begin{aligned}&\frac{7}{8}\|x^{k+1} - x^\infty\|^2_{H+\Sigma+\frac{4}{7}R} + \frac{1}{2}\|x_2^{k+1} - x_2^\infty\|^2_{H_{22}+\Sigma_2+\beta A_2^\top A_2}\\ &+\frac{1}{2\beta}\|\mu^{k+1} - \mu^\infty\|^2 + \frac{1}{2}\|x_2^{k+1} - x_2^k\|^2_{R_2}\end{aligned}$$

$$\leq \frac{7}{8}\|x^{k_i}-x^{\infty}\|^2_{H+\Sigma+\frac{4}{7}R}+\frac{1}{2}\|x_2^{k_i}-x_2^{\infty}\|^2_{H_{22}+\Sigma_2+\beta A_2^\top A_2}$$
$$+\frac{1}{2\beta}\|\mu^{k_i}-\mu^{\infty}\|^2+\frac{1}{2}\|x_2^{k_i}-x_2^{k_i-1}\|^2_{R_2}. \quad (31)$$

It follows from (29) that

$$\lim_{k\to\infty}\|x_2^{k+1}-x_2^k\|_{R_2}=0.$$

Note that

$$\lim_{i\to\infty}\Big(\frac{7}{8}\|x^{k_i}-x^{\infty}\|^2_{H+\Sigma+\frac{4}{7}R}+\frac{1}{2}\|x_2^{k_i}-x_2^{\infty}\|^2_{H_{22}+\Sigma_2+\beta A_2^\top A_2}$$
$$+\frac{1}{2\beta}\|\mu^{k_i}-\mu^{\infty}\|^2+\frac{1}{2}\|x_2^{k_i}-x_2^{k_i-1}\|^2_{R_2}\Big)=0,$$

and so we can deduce from (31) that

$$\lim_{k\to\infty}\Big(\frac{7}{8}\|x^{k+1}-x^{\infty}\|^2_{H+\Sigma+\frac{4}{7}R}+\frac{1}{2}\|x_2^{k+1}-x_2^{\infty}\|^2_{H_{22}+\Sigma_2+\beta A_2^\top A_2}$$
$$+\frac{1}{2\beta}\|\mu^{k+1}-\mu^{\infty}\|^2+\frac{1}{2}\|x_2^{k+1}-x_2^{k}\|^2_{R_2}\Big)=0,$$

which implies

$$\lim_{k\to\infty}\|x_2^{k+1}-x_2^{\infty}\|^2_{H_{22}+\Sigma_2+\beta A_2^\top A_2+R_2}=0,\qquad \lim_{k\to\infty}\mu^{k+1}=\mu^{\infty}$$

and

$$\lim_{k\to\infty}\|x^{k+1}-x^{\infty}\|^2_{H+\Sigma+R}=0. \quad (32)$$

Since $H_{22}+\Sigma_2+\beta A_2^\top A_2+R_2$ is positive definite, we obtain

$$\lim_{k\to\infty}x_2^{k+1}=x_2^{\infty}. \quad (33)$$

On the other hand, by (13) and (33), it can easily be seen that

$$\|A_1(x_1^{k+1}-x_1^{\infty})\|\leq\|A_1x_1^{k+1}+A_2x_2^{k+1}-(A_1x_1^{\infty}+A_2x_2^{\infty})\|+\|A_2(x_2^{k+1}-x_2^{\infty})\|$$
$$=\|A_1x_1^{k+1}+A_2x_2^{k+1}-b\|+\|A_2(x_2^{k+1}-x_2^{\infty})\|\to 0, \quad (34)$$

as $k\to\infty$. Then, we obtain

$$\|x_1^{k+1}-x_1^{\infty}\|^2_{H_{11}+\Sigma_1+\beta A_1^\top A_1+R_1}=\|x_1^{k+1}-x_1^{\infty}\|^2_{H_{11}+\Sigma_1+R_1}+\beta\|A_1(x_1^{k+1}-x_1^{\infty})\|^2$$
$$=\left\|\begin{bmatrix}x_1^{k+1}-x_1^{\infty}\\0\end{bmatrix}\right\|^2_{H+\Sigma+R}+\beta\|A_1(x_1^{k+1}-x_1^{\infty})\|^2$$
$$\leq\left(\left\|\begin{bmatrix}x_1^{k+1}-x_1^{\infty}\\x_2^{k+1}-x_2^{\infty}\end{bmatrix}\right\|_{H+\Sigma+R}+\left\|\begin{bmatrix}0\\x_2^{k+1}-x_2^{\infty}\end{bmatrix}\right\|_{H+\Sigma+R}\right)^2+\beta\|A_1(x_1^{k+1}-x_1^{\infty})\|^2$$

$$= (\|x^{k+1} - x^{\infty}\|_{H+\Sigma+R} + \|x_2^{k+1} - x_2^{\infty}\|_{H_{22}+\Sigma_2+R_2})^2 + \beta\|A_1(x_1^{k+1} - x_1^{\infty})\|^2,$$

where "$\leq$" follows the triangle inequality of norms. Together with (32), (33), (34), and the positive definiteness of $H_{11} + \Sigma_1 + \beta A_1^\top A_1 + R_1$, this shows that

$$\lim_{k\to\infty} x_1^{k+1} = x_1^{\infty}.$$

Therefore, we have shown that the whole sequence $\{(x_1^k, x_2^k, \mu^k)\}$ converges to $(x_1^{\infty}, x_2^{\infty}, \mu^{\infty})$, which is a KKT point of problem (7). This comletes the proof. □

Remark 1 In fact, the iterate convergence of 2-block proximal ADMM can also be guaranteed if there is a fixed stepsize $\gamma \in (0, (1+\sqrt{5})/2)$ in the dual update. Namely, the proximal ADMM can be extended as follows:

$$\begin{cases} x_1^{k+1} := \arg\min\limits_{x_1\in\mathbb{R}^{d_1}} \left\{\mathcal{L}_\beta(x_1, x_2^k; \mu^k) + \frac{1}{2}\|x_1 - x_1^k\|_{R_1}^2\right\}, \\ x_2^{k+1} := \arg\min\limits_{x_2\in\mathbb{R}^{d_2}} \left\{\mathcal{L}_\beta(x_1^{k+1}, x_2; \mu^k)) + \frac{1}{2}\|x_2 - x_2^k\|_{R_2}^2\right\}, \\ \mu^{k+1} := \mu^k - \gamma\beta(A_1x_1^{k+1} + A_2x_2^{k+1} - b), \end{cases} \tag{35}$$

where $\beta > 0$ and $\gamma \in (0, (1+\sqrt{5})/2)$. Under the conditions of Theorem *1*, we can similarly prove the global iterate convergence of (35). For brevity, we omit the details here.

Remark 2 The proximal ADMM includes the ADMM and its linearized version as special cases. When $R_1 = 0$ and $R_2 = 0$, the proximal ADMM reduces to the ADMM and, according to Theorem 1, its convergence can be established under the condition that

$$\begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} A_1^\top A_1 & 0 \\ 0 & A_2^\top A_2 \end{bmatrix} \succ 0$$

The ADMM can be easily applied to the convex minimization problems where θ_i $(i = 1, 2)$ have closed form proximal operators and all the matrices H_{11}, H_{22}, A_1, A_2 are diagonal. Otherwise, we consider the linearized ADMM:

$$\begin{cases} x_1^{k+1/2} := \left[(r_1I - H_{11} - \beta A_1^TA_1)x_1^k - \beta A_1^T(A_2x_2^k - b) - H_{12}x_2^k - g_1\right]/r_1, \\ x_1^{k+1} := \arg\min\limits_{x_1\in\mathbb{R}^{d_1}} \theta_1(x_1) + \frac{r_1}{2}\|x_1 - x_1^{k+1/2}\|^2, \\ x_2^{k+1/2} := \left[(r_2I - H_{22} - \beta A_2^TA_2)x_2^k - \beta A_2^T(A_1x_1^{k+1} - b) - H_{12}^Tx_1^{k+1} - g_2\right]/r_2, \\ x_2^{k+1} := \arg\min\limits_{x_2\in\mathbb{R}^{d_2}} \theta_2(x_2) + \frac{r_2}{2}\|x_2 - x_2^{k+1/2}\|^2, \\ \mu^{k+1} := \mu^k - \beta(A_1x_1^{k+1} + A_2x_2^{k+1} - b), \end{cases}$$

which is equivalent to the proximal ADMM with $R_1 = r_1I - H_{11} - \beta A_1^TA_1$ and $R_2 = r_2I - H_{22} - \beta A_2^TA_2$. Thus the iterate convergence of linearized ADMM can be guaranteed under the condition that

$$r_i \geq \max_{1\leq j\leq d_i} \lambda_j(H_{ii} + \beta A_i^T A_i), \quad i = 1, 2,$$

where $\lambda_j(\cdot)$ represents the jth eigenvalue of a matrix.

By using the following proposition (see [16, Lemma 1.1] and [34, Lemma 3]), we can deliver sublinear convergence rates of the proximal ADMM, measured by the square of KKT violation and the function value.

Proposition 1 *For any sequence* $\{a_i\} \subseteq \Re$ *satisfying* $a_i \geq 0$ *and* $\sum_{i=1}^{\infty} a_i < +\infty$*, it holds that* $\min_{1\leq i\leq k}\{a_i\} = o(1/k)$.

Theorem 2 *Suppose Assumption 1 holds. Let* $\{(x_1^k, x_2^k, \mu^k)\}$ *be generated by the proximal ADMM (4) with* $n = 2$ *to solve problem (7). If* $(\bar{x}_1, \bar{x}_2, \bar{\mu})$ *is a given KKT point of problem (7), we have*

$$\min_{1\leq i\leq k} \Big\{d^2\big(0,\ \partial\theta_1(x_1^{i+1}) + \nabla_{x_1}\phi(x_1^{i+1}, x_2^{i+1}) - A_1^\top \mu^{i+1}\big) + d^2\big(0,\ \partial\theta_2(x_2^{i+1}) + \nabla_{x_2}\phi(x_1^{i+1}, x_2^{i+1}) - A_2^\top \mu^{i+1}\big) + \|A_1 x_1^{i+1} + A_2 x_2^{i+1} - b\|^2\Big\} = o(1/k) \tag{36}$$

and

$$\min_{1\leq i\leq k} |(\theta_1(x_1^i) + \theta_2(x_2^i) + \phi(x_1^i, x_2^i)) - (\theta_1(\bar{x}_1) + \theta_2(\bar{x}_2) + \phi(\bar{x}_1, \bar{x}_2))| = o(1/\sqrt{k}). \tag{37}$$

Proof From (4) and (14), we obtain

$$\begin{cases} -R_1(x_1^{k+1} - x_1^k) + (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) \in \partial\theta_1(x_1^{k+1}) \\ \quad +\nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) - A_1^\top \mu^{k+1}, \\ -R_2(x_2^{k+1} - x_2^k) \in \partial\theta_2(x_2^{k+1}) + \nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) - A_2^\top \mu^{k+1}, \\ \|A_1 x_1^{k+1} + A_2 x_2^{k+1} - b\|^2 = \dfrac{1}{\beta^2}\|\mu^{k+1} - \mu^k\|^2. \end{cases}$$

By using the Cauchy-Schwarz inequality and the above formulas, we obtain

$$\begin{aligned} & d^2\big(0,\ \partial\theta_1(x_1^{k+1}) + \nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) - A_1^\top \mu^{k+1}\big) + d^2\big(0,\ \partial\theta_2(x_2^{k+1}) \\ & \quad +\nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) - A_2^\top \mu^{k+1}\big) + \|A_1 x_1^{k+1} + A_2 x_2^{k+1} - b\|^2 \\ & \leq 2\|R_1(x_1^{k+1} - x_1^k)\|^2 + 2\|(H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k)\|^2 \\ & \quad +\|R_2(x_2^{k+1} - x_2^k)\|^2 + \frac{1}{\beta^2}\|\mu^{k+1} - \mu^k\|^2 \\ & \leq 2\|R_1^{\frac{1}{2}}\|^2 \|x_1^{k+1} - x_1^k\|_{R_1}^2 + (2\|H_{12} + \beta A_1^\top A_2\|^2 + \|R_2\|^2)\|x_2^{k+1} - x_2^k\|^2 \\ & \quad +\frac{1}{\beta^2}\|\mu^{k+1} - \mu^k\|^2. \end{aligned} \tag{38}$$

It follows from (24) that

$$\begin{cases} \sum_{k=1}^{\infty} \|x_1^{k+1} - x_1^k\|_{R_1}^2 < \infty, \\ \sum_{k=1}^{\infty} \|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2+A_2^\top A_2+R_2}^2 < \infty, \\ \sum_{k=1}^{\infty} \|\mu^{k+1} - \mu^k\|^2 < \infty. \end{cases} \tag{39}$$

Since $H_{22} + \Sigma_2 + A_2^\top A_2 + R_2 \succ 0$, from (39) we have

$$\sum_{k=1}^{\infty} \|x_2^{k+1} - x_2^k\|^2 < \infty. \tag{40}$$

Combining Proposition 1 with the relationships (39) and (40), we have

$$\min_{1\le i\le k} \Big\{ 2\|R_1^{\frac{1}{2}}\|^2 \|x_1^{i+1} - x_1^i\|_{R_1}^2 + (2\|H_{12} + \beta A_1^\top A_2\|^2 + \|R_2\|^2)\|x_2^{i+1} - x_2^i\|^2 \\ + \frac{1}{\beta^2}\|\mu^{i+1} - \mu^i\|^2 \Big\} = o(1/k),$$

which, together with (38), implies (36).

Since $(\bar{x}_1, \bar{x}_2, \bar{\mu})$ is a KKT point of problem (7), for any $(x_1, x_2) \in \text{dom}(\theta_1) \times \text{dom}(\theta_2)$, we get

$$\begin{cases} \theta_1(x_1) - \theta_1(\bar{x}_1) + (x_1 - \bar{x}_1)^\top (\nabla_{x_1}\phi(\bar{x}_1, \bar{x}_2) - A_1^\top \bar{\mu}) \ge 0, \\ \theta_2(x_2) - \theta_2(\bar{x}_2) + (x_2 - \bar{x}_2)^\top (\nabla_{x_2}\phi(\bar{x}_1, \bar{x}_2) - A_2^\top \bar{\mu}) \ge 0. \end{cases}$$

Since ϕ is convex, we obtain

$$\phi(x_1, x_2) - \phi(\bar{x}_1, \bar{x}_2) \ge (x_1 - \bar{x}_1)^\top \nabla_{x_1}\phi(\bar{x}_1, \bar{x}_2) + (x_2 - \bar{x}_2)^\top \nabla_{x_2}\phi(\bar{x}_1, \bar{x}_2).$$

Adding the above three inequalities, and using $A_1\bar{x}_1 + A_2\bar{x}_2 = b$, we get

$$\Big(\theta_1(x_1) + \theta_2(x_2) + \phi(x_1, x_2)\Big) - \Big(\theta_1(\bar{x}_1) + \theta_2(\bar{x}_2) + \phi(\bar{x}_1, \bar{x}_2)\Big) - \bar{\mu}^\top (A_1x_1 + A_2x_2 - b) \ge 0.$$

Setting $x_1 = x_1^i$ and $x_2 = x_2^i$ in the above inequality, we have

$$\Big(\theta_1(x_1^i) + \theta_2(x_2^i) + \phi(x_1^i, x_2^i)\Big) - \Big(\theta_1(\bar{x}_1) + \theta_2(\bar{x}_2) + \phi(\bar{x}_1, \bar{x}_2)\Big) \ge \bar{\mu}^\top (A_1x_1^i + A_2x_2^i - b). \tag{41}$$

Note that θ_1, θ_2 and ϕ are convex functions and the sequence $\{(x_1^i, x_2^i, \mu^i)\}$ generated by the proximal ADMM (4) is bounded. For any $u \in \partial\theta_1(x_1^i)$, $v \in \partial\theta_2(x_2^i)$, using $A_1\bar{x}_1 + A_2\bar{x}_2 = b$, we obtain

$$
\begin{aligned}
&\Big(\theta_1(\bar{x}_1)+\theta_2(\bar{x}_2)+\phi(\bar{x}_1,\bar{x}_2)\Big)-\Big(\theta_1(x_1^i)+\theta_2(x_2^i)+\phi(x_1^i,x_2^i)\Big)\\
&\geq(\bar{x}_1-x_1^i)^\top(u+\nabla_{x_1}\phi(x_1^i,x_2^i))+(\bar{x}_2-x_2^i)^\top(v+\nabla_{x_2}\phi(x_1^i,x_2^i))\\
&=(\bar{x}_1-x_1^i)^\top(u+\nabla_{x_1}\phi(x_1^i,x_2^i)-A_1^\top\mu^i)+(\bar{x}_2-x_2^i)^\top\\
&(v+\nabla_{x_2}\phi(x_1^i,x_2^i)-A_2^\top\mu^i)-(A_1x_1^i+A_2x_2^i-b)^\top\mu^i,
\end{aligned}
$$

which, together with (41) and (36), implies (37). We complete the proof. □

We remark that, in some sense, Assumption 1 actually acts as the weakest condition to guarantee the iterate convergence of the proximal ADMM for solving problem (7). Firstly, if Assumption 1 is violated, the solution sets of subproblems in (4) might be empty, in which case the 2-block proximal ADMM scheme is not well defined (see [12] for an illustration). Secondly, the following corollary shows that Assumption 1 is not only sufficient, but also necessary for the iterate convergence of the 2-block proximal ADMM for solving the coupled quadratic minimization problem. Thus, the conditions we proposed are already tight.

Corollary 1 *Assume problem (7) is a convex quadratic programming problem, that is $\theta_1(x_1)\equiv 0$ and $\theta_2(x_2)\equiv 0$. Then, any sequence generated by the 2-block proximal ADMM is convergent if and only if Assumption 1 holds.*

Proof The “if” part follows immediately from Theorem 1. For the “only if” part, we prove that if Assumption 1 fails to hold, there must exist some sequence generated by the 2-block proximal ADMM that is divergent. Indeed, let $\{(x_1^k,x_2^k,\mu^k)\}$ be a sequence generated by the 2-block proximal ADMM, i.e.,

$$
\begin{cases}
x_1^{k+1}\in\arg\min\limits_{x_1\in\mathbb{R}^{d_1}}\Big\{\mathcal{L}_\beta(x_1,x_2^k;\mu^k)+\frac{1}{2}\|x_1-x_1^k\|_{R_1}^2\Big\},\\
x_2^{k+1}\in\arg\min\limits_{x_2\in\mathbb{R}^{d_2}}\Big\{\mathcal{L}_\beta(x_1^{k+1},x_2;\mu^k)+\frac{1}{2}\|x_2-x_2^k\|_{R_2}^2\Big\},\\
\mu^{k+1}=\mu^k-\beta(A_1x_1^{k+1}+A_2x_2^{k+1}-b).
\end{cases}
\tag{42}
$$

If the sequence is divergent, then the “only if” part of this corollary holds. Thus we need only consider the case where $\{(x_1^k,x_2^k,\mu^k)\}$ converges. Because either $H_{11}+\beta A_1^\top A_1+R_1$ or $H_{22}+\beta A_2^\top A_2+R_2$ is not positive definite, there exists a nonzero vector $(\bar{y}_1,\bar{y}_2)$ such that

$$(H_{ii}+\beta A_i^\top A_i+R_i)\bar{y}_i=0\quad\forall i=1,2,$$

or equivalently,

$$H_{ii}\bar{y}_i=0,\qquad A_i\bar{y}_i=0\qquad\text{and}\qquad R_i\bar{y}_i=0\quad\forall i=1,2,\tag{43}$$

since H and R are positive semidefinite. Using the fact that $0\preceq H\preceq 2\begin{bmatrix}H_{11}&0\\0&H_{22}\end{bmatrix}$, we have $H\bar{y}=0$. Hence, it holds that

$$H_{12}\bar{y}_2 = 0 \qquad \text{and} \qquad H_{12}^\top \bar{y}_1 = 0. \tag{44}$$

By (42), (43) and (44), it can easily be seen that, for any $k \geq 1$,

$$\begin{cases} x_1^{2k} + \bar{y}_1 \in \arg\min\limits_{x_1} \left\{ \mathcal{L}_\beta(x_1, x_2^{2k-1}; \mu^{2k-1}) + \frac{1}{2}\|x_1 - x_1^{2k-1}\|_{R_1}^2 \right\}, \\ x_2^{2k} + \bar{y}_2 \in \arg\min\limits_{x_2} \left\{ \mathcal{L}_\beta(x_1^{2k} + \bar{y}_1, x_2; \mu^{2k-1}) + \frac{1}{2}\|x_2 - x_2^{2k-1}\|_{R_2}^2 \right\}, \\ \mu^{2k} = \mu^{2k-1} - \beta\big(A_1(x_1^{2k} + \bar{y}_1) + A_2(x_2^{2k} + \bar{y}_2) - b\big) \end{cases}$$

and

$$\begin{cases} x_1^{2k+1} \in \arg\min\limits_{x_1} \left\{ \mathcal{L}_\beta(x_1, x_2^{2k} + \bar{y}_2; \mu^{2k}) + \frac{1}{2}\|x_1 - (x_1^{2k} + \bar{y}_1)\|_{R_1}^2 \right\}, \\ x_2^{2k+1} \in \arg\min\limits_{x_2} \left\{ \mathcal{L}_\beta(x_1^{2k+1}, x_2; \mu^{2k}) + \frac{1}{2}\|x_2 - (x_2^{2k} + \bar{y}_2)\|_{R_2}^2 \right\}, \\ \mu^{2k+1} = \mu^{2k} - \beta\big(A_1 x_1^{2k+1} + A_2 x_2^{2k+1} - b\big). \end{cases}$$

This means that the divergent sequence $(x_1^1, x_2^1, \mu^1) \to (x_1^2 + \bar{y}_1, x_2^2 + \bar{y}_2, \mu^2) \to (x_1^3, x_2^3, \mu^3) \to (x_1^4 + \bar{y}_1, x_2^4 + \bar{y}_2, \mu^4) \to \ldots$ could be generated by the 2-block proximal ADMM. Thus, Assumption 1 is also necessary for the iterate convergence. This completes the proof. □

When restricted to the case that A_i $(i = 1, 2)$ and b are absent, the 2-block proximal ADMM reduces to the 2-block cyclic proximal BCD method. Our analysis of proximal ADMM provides an iterate convergence result for the 2-block cyclic proximal BCD method without assuming the boundedness of the iterates, but only requiring a condition to ensure the uniqueness of the subproblem solutions. This result is an important supplement to traditional studies on BCD, which have mainly focused on subsequence convergence and the complexity of the function values, and enables a better understanding of the performance of this method.

Corollary 2 *Assume $H_{ii} + \Sigma_i + R_i \succ 0$. Let $\{(x_1^k, x_2^k)\}$ be generated by the cyclic proximal BCD (5) with $n = 2$ to solve the following unconstrained optimization problem:*

$$\min_{x \in \mathbb{R}^d} \ \theta_1(x_1) + \theta_2(x_2) + \frac{1}{2}x^\top Hx + g^\top x. \tag{45}$$

Then the whole sequence $\{(x_1^k, x_2^k)\}$ converges to an optimal solution of (45).

Remark 3 Similar to the proximal ADMM, the proximal BCD includes BCD and its linearized version (also know as BCPG) as special cases. When $R_1 = 0$ and $R_2 = 0$, the proximal BCD reduces to BCD and, according to Theorem 1, its convergence can be established under the condition that

$$\begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \succ 0$$

The BCPG is a combination of the proximal gradient method and BCD, which can be easily implemented when θ_i have closed-form proximal operators. Specifically, it takes the form that

$$\begin{cases} x_1^{k+1/2} := \left[(r_1 I - H_{11})x_1^k - H_{12}x_2^k - g_1\right]/r_1, \\ x_1^{k+1} := \operatorname*{arg\,min}_{x_1 \in \mathbb{R}^{d_1}} \left\{\theta_1(x_1) + \frac{r_1}{2}\|x_1 - x_1^{k+1/2}\|^2\right\}, \\ x_2^{k+1/2} := \left[(r_2 I - H_{22})x_2^k - H_{12}^\top x_1^{k+1} - g_2\right]/r_2, \\ x_2^{k+1} := \operatorname*{arg\,min}_{x_2 \in \mathbb{R}^{d_2}} \left\{\theta_2(x_2) + \frac{r_2}{2}\|x_2 - x_2^{k+1/2}\|^2\right\}, \end{cases}$$

which is equivalent to the proximal BCD with $R_1 = r_1 I - H_{11}$ and $R_2 = r_2 I - H_{22}$. Thus the iterate convergence of linearized ADMM can be guranteed under the condition that

$$r_i \geq \max_{1 \leq j \leq d_i} \lambda_j(H_{ii}), \quad i = 1, 2.$$

3 Convergence of multi-block RPADMM and RPBCD

As shown in [10], the convergence result for 2-block ADMM obtained in the previous section cannot be extended to the multi-block case, i.e., $n \geq 3$. To remove the possibility of divergence, we use randomly permuted ADMM (RPADMM) to solve the nonseparable optimization problem (1). Specifically, RPADMM first picks a permutation σ of $\{1, \ldots, n\}$ uniformly at random, and then iterates as follows:

$$\begin{cases} x_{\sigma(1)}^{k+1} := \operatorname*{arg\,min}_{x_{\sigma(1)}} \left\{\mathcal{L}_\beta(x_{\sigma(1)}, x_{\sigma(2)}^k, \ldots, x_{\sigma(n)}^k; \mu^k)\right\}, \\ x_{\sigma(2)}^{k+1} := \operatorname*{arg\,min}_{x_{\sigma(2)}} \left\{\mathcal{L}_\beta(x_{\sigma(1)}^{k+1}, x_{\sigma(2)}, x_{\sigma(3)}^k, \ldots, x_{\sigma(n)}^k; \mu^k)\right\}, \\ \cdots\cdots \\ x_{\sigma(n)}^{k+1} := \operatorname*{arg\,min}_{x_{\sigma(n)}} \left\{\mathcal{L}_\beta(x_{\sigma(1)}^{k+1}, x_{\sigma(2)}^{k+1}, \ldots, x_{\sigma(n-1)}^{k+1}, x_{\sigma(n)}; \mu^k)\right\}, \\ \mu^{k+1} := \mu^k - \beta(\sum_{i=1}^n A_i x_i^{k+1} - b), \end{cases} \tag{46}$$

where the permuted augmented Lagrangian function $\mathcal{L}_\beta(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}; \mu)$ is defined by

$$\mathcal{L}_\beta(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}; \mu) := \mathcal{L}_\beta(x_1, x_2, \ldots, x_n; \mu).$$

It has been shown in [48] that RPADMM is convergent in expectation for solving the nonsingular square system of linear equations. To extend their result to the nonseparable convex optimization model (1), it is natural to first study whether RPADMM is even convergent in expectation for solving the following simpler linearly constrained quadratic minimization problem

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & \theta(x) := \frac{1}{2}x^\top Hx + g^\top x \\ \text{s.t.} \quad & \sum_{i=1}^{n} A_i x_i = b, \end{aligned} \tag{47}$$

where H can be partitioned into $n \times n$ blocks $H_{ij} \in \mathbb{R}^{d_i \times d_j}$ $(1 \leq i, j \leq n)$ accordingly. In this section, we provide an affirmative answer to the above question under the following assumption.

Assumption 2 Assume

$$\begin{bmatrix} H_{11} & 0 & \cdots & 0 \\ 0 & H_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H_{nn} \end{bmatrix} + \begin{bmatrix} A_1^\top A_1 & 0 & \cdots & 0 \\ 0 & A_2^\top A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n^\top A_n \end{bmatrix} \succ 0.$$

Although our current result is restricted for nonseparable quadratic minimization, a special case of (1), it serves as a good indicator of the expected convergence of RPADMM in more general cases. It is noteworthy that our result is a non-trivial extension of the result in [48], because, in our setting, the problem under consideration is more general. For example, the optimal solution set of (47) is not necessarily a singleton, in which case the spectral radius of the algorithm mapping may not be strictly less than 1, although this fact played a key role in establishing their result.

3.1 Proof outline and preliminaries

For convenience, we follow the notation in [48], and describe the iterative scheme of RPADMM in a matrix form. Let $L_\sigma \in \mathbb{R}^{d\times d}$ be an $n \times n$ block matrix defined by

$$(L_\sigma)_{\sigma(i),\sigma(j)} := \begin{cases} H_{\sigma(i)\sigma(j)} + \beta A_{\sigma(i)}^\top A_{\sigma(j)}, & \text{if } i \geq j, \\ 0, & \text{otherwise,} \end{cases}$$

and R_σ be defined as

$$R_\sigma := L_\sigma - (H + \beta A^\top A) := L_\sigma - S. \tag{48}$$

By setting $z := (x, \mu)$, the randomly permuted ADMM can be viewed as a fixed point iteration

$$z^{k+1} := M_\sigma z^k + \bar{L}_\sigma^{-1}\bar{b}, \tag{49}$$

where

$$M_\sigma := \bar{L}_\sigma^{-1}\bar{R}_\sigma, \quad \bar{L}_\sigma := \begin{bmatrix} L_\sigma & 0 \\ \beta A & I \end{bmatrix}, \quad \bar{R}_\sigma := \begin{bmatrix} R_\sigma & A^\top \\ 0 & I \end{bmatrix}, \quad \bar{b} := \begin{bmatrix} -g + \beta A^\top b \\ \beta b \end{bmatrix}.$$

Define the matrix Q by

$$Q := E_\sigma(L_\sigma^{-1}) = \frac{1}{n!}\sum_{\sigma\in\Gamma} L_\sigma^{-1} \tag{50}$$

and M by

$$M := E_\sigma(M_\sigma) = \frac{1}{n!}\sum_{\sigma\in\Gamma} M_\sigma, \tag{51}$$

where Γ is the set of all permutations of $\{1, 2, \dots, n\}$. By direct computation, we can easily see that

$$M := \begin{bmatrix} I - QS & QA^\top \\ -\beta A + \beta AQS & I - \beta AQA^\top \end{bmatrix}. \tag{52}$$

To prove the expected convergence of the RPADMM (46) for problem (47) under Assumption 2, we will use a similar, but not identical, structure as that introduced in [48], which consists of the following main steps:

(1) $\operatorname{eig}(QS) \subset \left[0, \frac{4}{3}\right)$;
(2) For any eigenvalue λ of M, $\operatorname{eig}(QS) \subset \left[0, \frac{4}{3}\right)$ implies that $|\lambda| < 1$ or $\lambda = 1$;
(3) If 1 is an eigenvalue of M, then the eigenvalue 1 has a complete set of eigenvectors;
(4) Items (2) and (3) imply the convergence in expectation of the RPADMM.

To prove the above items, we need the following linear algebra lemmas, whose proofs can be found in the Appendix.

Lemma 2 *Suppose that Assumption 2 holds, $S \in \mathbb{R}^{d\times d}$ is a symmetric matrix defined by (48) and Q is defined by (50). Then, the matrix Q is positive definite and all the eigenvalues of QS lie in $\left[0, \frac{4}{3}\right)$, i.e.,*

$$\operatorname{eig}(QS) \subset \left[0, \frac{4}{3}\right). \tag{53}$$

Lemma 3 *Let S and T be two symmetric positive semidefinite matrices in $\mathbb{R}^{d\times d}$. Then, there exists a polynomial $p(x)$ such that*

$$\det\left((\lambda-1)^2 I + (2\lambda-1)S + (\lambda-1)T\right) = (\lambda-1)^l p(\lambda)$$

and $p(1) > 0$, where $\det(\cdot)$ denotes the determinant of some matrix, $l = 2d - \operatorname{Rank}(S) - \operatorname{Rank}(S+T)$ and $\operatorname{Rank}(\cdot)$ denotes the rank of some matrix.

Lemma 4 *Suppose $S \in \mathbb{R}^{d\times d}$ is a symmetric matrix defined by (48) and $\beta > 0$, then*

$$\operatorname{Rank}\begin{bmatrix} S & -A^\top \\ \beta A & 0 \end{bmatrix} = \operatorname{Rank}(S) + \operatorname{Rank}(\beta A^\top A).$$

Here, Lemma 2, Step (1) of the proof structure, is an enhanced version of Lemma 2 in [48] that is compatible with problem (47). The proofs of Steps (2) and (3), which reveal the essential nature of this extension and are hence the key contributions here, will be presented in Sect. 3.2. The proof for Step (4) is given in Sect. 3.3.

3.2 Eigenvalues of the expected update matrix

One of the main differences between the nonsingular linear system case and that of the extended case is reflected in the following lemma, where 1 can be an eigenvalue of the expected update matrix M.

Lemma 5 *Suppose that Assumption 2 holds and $S \in \mathbb{R}^{d\times d}$ is a symmetric matrix defined by (48). Let λ be any eigenvalue of M, then we have either $|\lambda| < 1$ or $\lambda = 1$.*

Proof We introduce the following notation:

$$\gamma(u) = \frac{\beta u^* A^\top A u}{u^* S u} \quad \text{for all } u \in \mathbb{C}^n \text{ such that} \quad Su \neq 0, \tag{54}$$

where u^* is the complex conjugate of u. Recalling that $S = H + \beta A^\top A$, we know

$$0 \leq \gamma(u) \leq 1 \quad \text{for all } u \in \mathbb{C}^n \text{ such that} \quad Su \neq 0. \tag{55}$$

Similarly, we define

$$\kappa(u) = \frac{u^* Q^{-1} u}{u^* S u} \quad \text{for all } u \in \mathbb{C}^n \text{ such that} \quad Su \neq 0. \tag{56}$$

Note that $\text{eig}(QS) < \frac{4}{3}$ by Lemma 2. Thus, we know that $\frac{4}{3}Q^{-1} - S \succeq 0$, and therefore

$$0 < \kappa(u)^{-1} < \frac{4}{3} \quad \text{for all } u \in \mathbb{C}^n \text{ such that} \quad Su \neq 0. \tag{57}$$

Note that M can be factorized as

$$M = \begin{bmatrix} I & 0 \\ -\beta A & I \end{bmatrix} \begin{bmatrix} I - QS & QA^\top \\ 0 & I \end{bmatrix}. \tag{58}$$

Switching the order of the products, we obtain a new matrix

$$M' := \begin{bmatrix} I - QS & QA^\top \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -\beta A & I \end{bmatrix} = \begin{bmatrix} I - QS - \beta QA^\top A & QA^\top \\ -\beta A & I \end{bmatrix}. \tag{59}$$

Note that $\text{eig}(M) = \text{eig}(M')$. Thus, it suffices to show either $\rho(M') < 1$ or 1 is the eigenvalue of M'.

Let $\left(\lambda, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right)$ be an eigenpair of M', namely,

$$\begin{bmatrix} I - QS - \beta QA^\top A & QA^\top \\ -\beta A & I \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

which implies

$$(I - QS - \beta QA^\top A)v_1 + QA^\top v_2 = \lambda v_1; \tag{60}$$

$$-\beta A v_1 + v_2 = \lambda v_2. \tag{61}$$

Equality (61) gives

$$(1 - \lambda)v_2 = \beta A v_1. \tag{62}$$

Suppose $\lambda \neq 1$. Hence, it holds that

$$v_2 = \frac{\beta}{1 - \lambda} A v_1.$$

Clearly, this relation implies that $v_1 \neq 0$. Substituting the above relation into (60), we have

$$QSv_1 = (1 - \lambda)v_1 + \frac{\lambda\beta}{1 - \lambda} QA^\top A v_1.$$

Using the nonsingularity of Q, the above equality can be written as

$$Sv_1 = (1 - \lambda)Q^{-1}v_1 + \frac{\lambda\beta}{1 - \lambda} A^\top A v_1.$$

Multiplying both sides of the above equality by v_1^*, we arrive at

$$v_1^* S v_1 = (1 - \lambda)v_1^* Q^{-1} v_1 + \frac{\lambda\beta}{1 - \lambda} v_1^* A^\top A v_1, \tag{63}$$

We claim that $v_1^* S v_1 \neq 0$. Otherwise, $v_1^* A^\top A v_1 = 0$ and therefore $\lambda = 1$ from the inequality $v_1^* Q^{-1} v_1 > 0$ and (63). This contradicts our assumption that $\lambda \neq 1$. Multiplying both sides of (63) by $(v_1^* S v_1)^{-1}$ and substituting the definitions (54) and (56) into the above relation, we obtain the following key equality with respect to λ

$$1 = (1 - \lambda)\kappa(v_1) + \frac{\lambda}{1 - \lambda}\gamma(v_1),$$

which can be further reformulated as

$$\kappa(v_1)\lambda^2 - (2\kappa(v_1) - \gamma(v_1) - 1)\lambda + \kappa(v_1) - 1 = 0.$$

Because $\kappa(v_1)$ is positive, we have

$$\lambda^2 + \left(\kappa(v_1)^{-1}(\gamma(v_1)+1) - 2\right)\lambda + \left(1 - \kappa(v_1)^{-1}\right) = 0. \tag{64}$$

The discriminant of the quadratic equation in (64) is

$$\begin{aligned} \Delta &= \left(\kappa(v_1)^{-1}(\gamma(v_1)+1) - 2\right)^2 - 4\left(1 - \kappa(v_1)^{-1}\right) \\ &= \kappa(v_1)^{-1}\left(\kappa(v_1)^{-1}(\gamma(v_1)+1)^2 - 4\gamma(v_1)\right). \end{aligned} \tag{65}$$

Note that

$$0 \le \frac{4\gamma(v_1)}{(\gamma(v_1)+1)^2} \le 1$$

holds as a result of (55). Recalling (57), we consider the following two cases.

Case 1: $0 < \kappa(v_1)^{-1} < \frac{4\gamma(v_1)}{(\gamma(v_1)+1)^2}$. This means the discriminant $\Delta < 0$, and the two solutions of (64) satisfy

$$|\lambda_{1,2}| = \sqrt{\lambda_1 * \lambda_2} = \sqrt{1 - \kappa(v_1)^{-1}} < 1.$$

Case 2: $\frac{4\gamma(v_1)}{(\gamma(v_1)+1)^2} \le \kappa(v_1)^{-1} < \frac{4}{3}$. This means the discriminant $\Delta \ge 0$, and the two solutions are real. Let

$$f(\lambda) := \lambda^2 + \left(\kappa(v_1)^{-1}(\gamma(v_1)+1) - 2\right)\lambda + \left(1 - \kappa(v_1)^{-1}\right).$$

By (55) and (57), we know that

$$\begin{cases} f(1) = \frac{\gamma(v_1)}{\kappa(v_1)} \ge 0, \\ f(-1) = 4 - \frac{\gamma(v_1)+2}{\kappa(v_1)} > 0, \\ \lambda_1 + \lambda_2 = 2 - \frac{\gamma(v_1)+1}{\kappa(v_1)} \in (-2,\ 2), \end{cases}$$

which together with $\lambda \ne 1$, establishes that $|\lambda| < 1$.

Thus, it can be concluded that either $\lambda = 1$ or $|\lambda| < 1$ holds. □

We now consider the case where M has an eigenvalue equal to 1 and show that it has a complete set of eigenvectors.

Lemma 6 *Suppose that Assumption 2 holds, and $M \in \mathbb{R}^{(m+d)\times(m+d)}$ is a matrix defined by (52). Suppose that 1 is an eigenvalue of M, then the algebraic multiplicity of 1 for M equals its geometric multiplicity. Namely, the eigenvalue 1 has a complete set of eigenvectors.*

Proof By direct computation, it holds that

$$\begin{aligned}
\det(\lambda I - M) &= \det\begin{bmatrix} (\lambda-1)I + QS & -QA^\top \\ \beta A - \beta AQS & (\lambda-1)I + \beta AQA^\top \end{bmatrix} \\
&= \det\begin{bmatrix} (\lambda-1)I + QS & -QA^\top \\ \lambda\beta A & (\lambda-1)I \end{bmatrix} \\
&= \det\begin{bmatrix} (\lambda-1)I + QS + \dfrac{\lambda\beta}{\lambda-1}QA^\top A & -QA^\top \\ 0 & (\lambda-1)I \end{bmatrix} \\
&= (\lambda-1)^{m-d}\det\Big[(\lambda-1)^2 I + (2\lambda-1)\beta QA^\top A + (\lambda-1)QH\Big] \\
&= (\lambda-1)^{m-d}\det\Big[(\lambda-1)^2 I + (2\lambda-1)\beta Q^{1/2}A^\top AQ^{1/2} + (\lambda-1)Q^{1/2}HQ^{1/2}\Big].
\end{aligned}$$

This, together with Lemma 3, shows that the algebraic multiplicity of 1 for M equals

$$\begin{aligned}
&m - d + 2d - \mathrm{Rank}(Q^{1/2}\beta A^\top AQ^{1/2}) - \mathrm{Rank}(Q^{1/2}(\beta A^\top A + H)Q^{1/2}) \\
&= m + d - \mathrm{Rank}(\beta A^\top A) - \mathrm{Rank}(\beta A^\top A + H), \qquad (66)
\end{aligned}$$

where the equality follows from $Q \succ 0$ by Lemma 2. In addition, the geometric multiplicity of 1 for M is identical to the following quantity:

$$\begin{aligned}
&m + d - \mathrm{Rank}(I - M) \\
&= m + d - \mathrm{Rank}\begin{bmatrix} QS & -QA^\top \\ \beta A - \beta AQS & \beta AQA^\top \end{bmatrix} \\
&= m + d - \mathrm{Rank}\begin{bmatrix} QS & -QA^\top \\ \beta A & 0 \end{bmatrix} \\
&= m + d - \mathrm{Rank}\begin{bmatrix} S & -A^\top \\ \beta A & 0 \end{bmatrix}, \qquad (67)
\end{aligned}$$

where the second equality follows from the rank invariant property under elementary transformation, and the final equality holds because $Q \succ 0$ by Lemma 2. Combining (66), (67), Lemma 4, and the definition of S, we derive the desired conclusion. □

3.3 Expected convergence

Step (4) can be formulated as the following theorem.

Theorem 3 *Assume Assumption 2 holds. Suppose RPADMM (46) is employed to solve the nonseparable quadratic programming (47). Then, the expected iterative sequence converges to some KKT point of (47).*

Proof Let $(\bar{x}, \bar{\mu})$ be a KKT point of (47), i.e.,

$$\begin{bmatrix} H & -A^\top \\ \beta A & 0 \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{\mu} \end{bmatrix} = \begin{bmatrix} -g \\ \beta b \end{bmatrix}. \tag{68}$$

Denote $(x^k,\ \mu^k)$ by the kth iterate of the algorithm. It follows from (49) and (68) that

$$E_\sigma[x^{k+1} - \bar{x}; \mu^{k+1} - \bar{\mu}] = M E_\sigma[x^k - \bar{x}; \mu^k - \bar{\mu}].$$

By Lemma 5, we know that $\rho(M) \le 1$. We proceed with the proof by considering the following two cases.

Case 1: $\rho(M) < 1$. It holds that $E_\sigma x^k \to \bar{x}$ and $E_\sigma \mu^k \to \bar{\mu}$ as $k \to \infty$. Theorem 3 is valid.

Case 2: $\rho(M) = 1$. By Lemmas 5 and 6, we know that all eigenvalues of M with modulus 1 must be 1, which has a complete set of eigenvectors. As a result, M admits the following Jordan decomposition:

$$M = P^{-1} \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \rho_1 & * & \\ & & & & \ddots & * \\ & & & & & \rho_t \end{bmatrix} P,$$

where P is a nonsingular matrix and $|\rho_i| < 1$ for all $i = 1, \ldots, t$. It is easily verified that

$$M^k \to P^{-1} \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} P$$

as $k \to \infty$, and therefore the sequence $\{E[x^{k+1} - \bar{x}; \mu^{k+1} - \bar{\mu}]\}$ converges to an eigenvector of M associated with the eigenvalue 1, say $[x^0; \mu^0]$. Then

$$(I - M)[x^0; \mu^0] = 0,$$

which, after some manipulation, shows that

$$\begin{bmatrix} H & -A^\top \\ \beta A & 0 \end{bmatrix} \begin{bmatrix} x^0 \\ \mu^0 \end{bmatrix} = 0. \tag{69}$$

Therefore, $Ex^k \to \bar{x} + x^0$ and $E\mu^k \to \bar{\mu} + \mu^0$ with

$$\begin{bmatrix} H & -A^\top \\ \beta A & 0 \end{bmatrix} \begin{bmatrix} \bar{x} + x^0 \\ \bar{\mu} + \mu^0 \end{bmatrix} = \begin{bmatrix} -g \\ \beta b \end{bmatrix}. \tag{70}$$

This means that $(\bar{x} + x^0, \bar{\mu} + \mu^0)$ is a KKT point of (47).

This completes the proof. □

One byproduct of Theorem 3 is the expected convergence result for RPBCD when applied to convex quadratic optimization. To the best of our knowledge, this is the first expected iterate convergence result of RPBCD.

Corollary 3 *Assume $H_{ii} \succ 0$ for $i = 1, 2, \dots, n$. If RPBCD is used to solve the unconstrained quadratic programming problem*

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top H x + g^\top x, \tag{71}$$

then the expected iterative sequence converges to an optimal solution of (71).

3.4 Convergence rate comparison to cyclic BCD

There is a common perception that RPBCD dominates cyclic BCD in terms of performance (see [51], for example). In this subsection, we theoretically show that this is not generally true. Consider the quadratic programming problem (71), where x is split into two blocks (x_1, x_2) with $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$, and $d = d_1 + d_2$. Accordingly, we denote

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^\top & H_{22} \end{bmatrix}.$$

By applying different minimizaing orders to the variables, the cyclic BCD (Gauss-Seidel method) has the following two iterative schemes:

$$x^{k+1} = M_1 x^k - \begin{bmatrix} H_{11} & 0 \\ H_{12}^\top & H_{22} \end{bmatrix}^{-1} b$$

and

$$x^{k+1} = M_2 x^k - \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix}^{-1} b,$$

where

$$M_1 = \begin{bmatrix} H_{11} & 0 \\ H_{12}^\top & H_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & -H_{12} \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad M_2 = \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ -H_{12}^\top & 0 \end{bmatrix}. \tag{72}$$

The asymptotic convergence rates of these two iterative schemes are $\rho(M_1)$ and $\rho(M_2)$, respectively. In this case, the expected asymptotic convergence rate of RPBCD is $\rho((M_1 + M_2)/2)$. The following proposition reveals the relationship between these rates.

Proposition 2 *Suppose $H_{11} \succ 0$ and $H_{22} \succ 0$. Let M_1 and M_2 be defined by (72), and $M_3 = (M_1 + M_2)/2$. Then, it holds that*

$$\rho(M_1) = \rho(M_2) \leq \rho(M_3).$$

Proof Without loss of generality, we need only consider the situation where $H_{ii} = I_{d_i}$ for $i = 1, 2$ and $d_1 \geq d_2$ because the similarity transformation $M \mapsto PMP^{-1}$ does not change the spectrum of M, where $P = \begin{bmatrix} H_{11}^{\frac{1}{2}} & 0 \\ 0 & H_{22}^{\frac{1}{2}} \end{bmatrix}$. In this case, a simple calculation yields

$$M_1 = \begin{bmatrix} 0 & -H_{12} \\ 0 & H_{12}^\top H_{12} \end{bmatrix} \qquad \text{and} \qquad M_2 = \begin{bmatrix} H_{12} H_{12}^\top & 0 \\ -H_{12}^\top & 0 \end{bmatrix}. \tag{73}$$

Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{d_2}$ be the eigenvalues of $H_{12}^\top H_{12}$. Recall that $H \succeq 0$ and $H_{ii} = I_{d_i}$ for $i = 1, 2$. Then, we have that $\sigma_i \in [0, 1]$, $i = 1, \ldots, d_2$, and obtain from (73) that

$$\rho(M_1) = \rho(M_2) = \sigma_1.$$

Clearly,

$$M_3 = \frac{1}{2} \begin{bmatrix} H_{12} H_{12}^\top & -H_{12} \\ -H_{12}^\top & H_{12}^\top H_{12} \end{bmatrix}.$$

By direct computation, it holds that

$$\begin{aligned} \det(\lambda I - M_3) &= \left(\frac{1}{2}\right)^d \det \begin{bmatrix} 2\lambda I - H_{12} H_{12}^\top & H_{12} \\ H_{12}^\top & 2\lambda I - H_{12}^\top H_{12} \end{bmatrix} \\ &= \left(\frac{1}{2}\right)^d (2\lambda)^{d_1 - d_2} \det \left[4\lambda^2 I - (4\lambda + 1) H_{12}^\top H_{12} + (H_{12}^\top H_{12})^2 \right] \\ &= \left(\frac{1}{2}\right)^d (2\lambda)^{d_1 - d_2} \prod_{i=1}^{d_2} (4\lambda^2 - (4\lambda + 1)\sigma_i + \sigma_i^2) \end{aligned}$$

and so the eigenvelues of M_3 are 0 (multiplicty $= d_1 - d_2$) and $\frac{\sigma_i \pm \sqrt{\sigma_i}}{2}$ for $i = 1, 2, \ldots, d_2$. Because $\sigma_1 \in [0, 1]$, we have that

$$\rho(M_3) = \frac{\sigma_1 + \sqrt{\sigma_1}}{2} \geq \sigma_1.$$

This completes the proof. □

Therefore, although random permutation does indeed make multi-block ADMM and BCD more robust, especially for “bad” or diverging problems, cyclic ADMM or BCD may still perform well, or even better, for solving “nice” problems with a few blocks.

4 Concluding remarks

In this paper, we have demonstrated the point-wise or iterate convergence of the classical 2-block ADMM for solving convex optimization problems with coupled quadratic objective functions under a mild assumption. This assumption becomes necessary and sufficient for the global convergence of the ADMM when the objective is a quadratic function. This result partially answers, in the affirmative, the open question arising in [31] on the convergence of ADMM for nonseparable optimization problems. We also derived the expected convergence of RPADMM in solving linearly constrained coupled quadratic optimization problems. This is a non-trivial extension of the convergence analysis given in [48], which is only applicable to nonsingular linear systems. When the linear constraint is absent, the proximal ADMM and RPADMM reduce to the cyclic proximal BCD and RPBCD. Thus, this study has provided new convergence results for BCD-type methods. In particular, we have established the first iterate convergence result for 2-block cyclic proximal BCD without assuming the boundedness of the iterates and the expected iterate convergence of RPBCD for multi-block convex quadratic optimization. We also theoretically demonstrated that RPBCD does not necessarily dominate cyclic BCD. Although the results for RPADMM and RPBCD are restricted to quadratic minimization models, they provide some interesting insights on the use of these methods: 1) random permutation makes multi-block ADMM and BCD more robust for multi-block convex minimization problems; 2) cyclic BCD may outperform RPBCD for “nice” problems, and therefore RPBCD should be applied with caution when solving general convex optimization problems especially with a few blocks.

Two challenging open questions concern the extension of our convergence results for RPADMM and RPBCD to more general convex optimization problems, and an exploration of the global convergence rate of RPADMM and RPBCD. In particular, it would be interesting to know which problems are better suited to RPADMM or RPBCD.

Acknowledgements Caihua Chen was supported by the National Natural Science Foundation of China [Grant No. 11401300, 71732003, 71673130]. Min Li was supported by the National Natural Science Foundation of China [Grant No.11771078, 71390335, 71661147004]. Xin Liu was supported by the National Natural Science Foundation of China [Grant No. 11622112, 11471325, 91530204, 11331012, 11461161005, 11688101], the National Center for Mathematics and Interdisciplinary Sciences, CAS, the Youth Innovation Promotion Association, CAS, and Key Research Program of Frontier Sciences, CAS. Yinyu Ye was supported by the AFOSR Grant [Grant No. FA9550-12-1-0396]. The authors would like to thank Dr. Ji Liu from University of Rochester and Dr. Ruoyu Sun from Stanford University for the helpful discussions on the block coordinate descent method. The authors would also like to thank the associate editor and two anonymous referees for their detailed and valuable comments and suggestions.

Appendix A.

The proof of Lemma 2 is similar to, but not exactly the same as, that of [48, Lemma 2]. Since S is allowed to be singular here, we need also show the positive definiteness of Q by mathematical induction. For completeness, we will provide a concise proof here. Interested readers are referred to [48] for the motivation and other details of this proof.

Lemma 2 actually reveals a linear algebra property, and is essentially not related with H, A and β if we define L_σ directly by S. For brevity, we restate the main assertion to be proved as following:

$$\mathrm{eig}(QS) \subset \left[0, \frac{4}{3}\right), \tag{74}$$

where $S \in \mathbb{R}^{d\times d}$ is positive semidefinite, $S_{ii} \in \mathbb{R}^{d_i \times d_i}$ $(i = 1, \ldots, n)$ is positive definite,

$$(L_\sigma)_{\sigma(i),\sigma(j)} := \begin{cases} S_{\sigma(i)\sigma(j)}, & \text{if } \ 1 \le j \le i \le n, \\ 0, & \text{otherwise}, \end{cases} \qquad Q := \frac{1}{n!}\sum_{\sigma\in\Gamma} L_\sigma^{-1}, \tag{75}$$

and Γ is a set consisting of all permutations of $(1, \ldots, n)$.

For the brevity of notation, we define the block permutation matrix P_k as following:

$$(P_k)_{ij} := \begin{cases} I_{d_i}, & \text{if}\, 1 \le i = j \le k-1; \\ I_{d_i}, & \text{if}\, k+1 \le i = j+1 \le n; \\ I_{d_i}, & \text{if}\, i = k,\ j = n; \\ 0_{d_i\times d_j}, & \text{if}\, 1 \le j \le k-1,\ i \ne j; \\ 0_{d_i\times d_{j+1}}, & \text{if}\, k \le j \le n-1,\ i \ne j+1; \\ 0_{d_i\times d_k}, & \text{otherwise.} \end{cases} \tag{76}$$

It can be easily verified that $P_k^\top = P_k^{-1}$, and $P_n = I_d$. For $k \in \{1, \ldots, n\}$, we define $\Gamma_k := \{\sigma' \mid \sigma'\text{is a permutation of}\, \{1, \ldots, k-1, k+1, \ldots, n)\}$. For any $\sigma' \in \Gamma_k$, we define $L_{\sigma'} \in \mathbb{R}^{(d-d_k)\times(d-d_k)}$ as the following

$$(L_{\sigma'})_{\sigma'(i),\sigma'(j)} := \begin{cases} S_{\sigma'(i)\sigma'(j)}, & \text{if } \ 1 \le j \le i \le n-1, \\ 0, & \text{otherwise.} \end{cases} \tag{77}$$

We define $\hat{Q}_k \in \mathbb{R}^{(n-d_k)\times(n-d_k)}$ by

$$\hat{Q}_k := \frac{1}{|\Gamma_k|}\sum_{\sigma'\in\Gamma_k} L_{\sigma'}^{-1}, \qquad k=1,\dots,n, \tag{78}$$

and W_k as the k-th block-column of S excluding the block S_{kk}, i.e.

$$W_k = [S_{k1},\dots,S_{kn}]^\top. \tag{79}$$

Due to the positive semi-definiteness of S, and by a slight abuse of the notation A, there exists $A \in \mathbb{R}^{d\times d}$ satisfying

$$S = A^\top A. \tag{80}$$

Let $A_i \in \mathbb{R}^{d\times d_i}$ $(i = 1,\dots,n)$ be the column blocks of A, and it is clear that $S_{ij} = A_i^\top A_j$ for all $1 \le i, j \le n$. For convenience, we define

$$\hat{A}_k := [A_1,\dots,A_{k-1},A_{k+1},\dots,A_n], \tag{81}$$

we have $AP_k = [\hat{A}_k, A_k]$.

For the clearness of the proof structure, we introduce the following two lemmas.

Lemma 7 *Let $S \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix L_σ, Q, $\hat{Q}^k$ and P_k be defined by* (48)*,* (50)*,* (78) *and* (76)*. It holds that*

$$Q = \frac{1}{n}\sum_{k=1}^{n} P_k Q_k P_k^\top, \tag{82}$$

where

$$Q_k := \begin{bmatrix} \hat{Q}_k & -\frac{1}{2}\hat{Q}_k W_k \\ -\frac{1}{2}W_k^\top \hat{Q}_k & I_{d_k} \end{bmatrix}. \tag{83}$$

Proof Let $\sigma' \in \Gamma_k$, we can partition $L_{\sigma'}$ as following

$$L_{\sigma'} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \tag{84}$$

Here the sizes of Z_{11} and Z_{22} are $(d_1+\cdots+d_{k-1})\times(d_1+\cdots+d_{k-1})$ and $(d_{k+1}+\cdots+d_n)\times(d_{k+1}+\cdots+d_n)$, respectively. The sizes of Z_{12} and Z_{21} can be determined accordingly. We denote

$$U_k = (A_1,\dots,A_{k-1}), \qquad V_k = (A_{k+1},\dots,A_n),$$

which implies

$$W_k = [U_k, V_k]^\top A_k = \begin{bmatrix} U_k^\top A_k \\ V_k^\top A_k \end{bmatrix}. \tag{85}$$

It is then easy to verify that

$$L_{(\sigma',k)} = \begin{bmatrix} Z_{11} & U_k^\top A_k & Z_{12} \\ 0 & I_{d_k} & 0 \\ Z_{21} & V_k^\top A_k & Z_{22} \end{bmatrix}.$$

Left and right multiplying both sides of the above relationship by $P_k^\top$ and P_k, respectively, we obtain

$$P_k^\top L_{(\sigma',k)} P_k = P_k^\top \begin{bmatrix} Z_{11} & Z_{12} & U_k^\top A_k \\ 0 & 0 & I_{d_k} \\ Z_{21} & Z_{22} & V_k^\top A_k \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} & U_k^\top A_k \\ Z_{21} & Z_{22} & V_k^\top A_k \\ 0 & 0 & I_{d_k} \end{bmatrix} = \begin{bmatrix} L_{\sigma'} & W_k \\ 0 & I_{d_k} \end{bmatrix}. \tag{86}$$

Taking the inverse of both sides of (86), we obtain

$$P_k^\top L_{(\sigma',k)}^{-1} P_k = \begin{bmatrix} L_{\sigma'}^{-1} & -L_{\sigma'}^{-1} W_k \\ 0 & I_{d_k} \end{bmatrix}. \tag{87}$$

Summing up (87) for all $\sigma' \in \Gamma_k$ and dividing by $|\Gamma_k|$, we get

$$\begin{aligned} \frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} P_k^\top L_{(\sigma',k)}^{-1} P_k &= \begin{bmatrix} \frac{1}{|\Gamma_k|} \sum\limits_{\sigma' \in \Gamma_k} L_{(\sigma')}^{-1} & -\frac{1}{|\Gamma_k|} \sum\limits_{\sigma' \in \Gamma_k} L_{(\sigma')}^{-1} W_k \\ 0 & I_{d_k} \end{bmatrix} \\ &= \begin{bmatrix} \hat{Q}_k & -\hat{Q}_k W_k \\ 0 & I_{d_k} \end{bmatrix}. \end{aligned} \tag{88}$$

Here, the last equality follows from (78). By the definition of L_σ, it is easy to verify that $L_\sigma^\top = L_{\bar{\sigma}}$, where $\bar{\sigma}$ is a "reverse permutation" of σ that satisfies $\bar{\sigma}(i) = \sigma(n+1-i)$ $(i = 1, \ldots, n)$. Thus we have $L_{(\sigma',k)} = L_{(k,\bar{\sigma}')}^\top$, where $\bar{\sigma}'$ is a reverse permutation of σ'. Summing over all σ', we get

$$\sum_{\sigma' \in \Gamma_k} L_{(\sigma',k)}^{-1} = \sum_{\sigma' \in \Gamma_k} L_{(k,\bar{\sigma}')}^{-\top} = \sum_{\sigma' \in \Gamma_k} L_{(k,\sigma')}^{-\top},$$

where the last equality follows from the fact that the summing over $\bar{\sigma}'$ is the same as summing over σ'. Thus, we have

$$\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} P_k^\top L_{(k,\sigma')}^{-1} P_k = \left(\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} P_k^\top L_{(\sigma',k)}^{-1} P_k \right)^\top = \begin{bmatrix} \hat{Q}_k & 0 \\ -W_k^\top \hat{Q}_k & I_{d_k} \end{bmatrix}.$$

Here, the last equality uses the symmetry of $\hat{Q}_k$. Combining the above relation, (88) and the definition of Q_k, we have

$$\frac{1}{2|\Gamma_k|} P_k^\top \sum_{\sigma' \in \Gamma_k} \left(L_{(k,\sigma')}^{-1} + L_{(\sigma',k)}^{-1} \right) P_k = \begin{bmatrix} \hat{Q}_k & -\frac{1}{2}\hat{Q}_k W_k \\ -\frac{1}{2} W_k^\top \hat{Q}_k & I_{d_k} \end{bmatrix} = Q_k. \quad (89)$$

Using the definition of P_k and the fact that $|\Gamma_k| = (n-1)!$, we can rewrite (89) as

$$S_k Q_k S_k^\top = \frac{1}{2(n-1)!} \sum_{\sigma' \in \Gamma_k} \left(L_{(k,\sigma')}^{-1} + L_{(\sigma',k)}^{-1} \right).$$

Summing up the above relation for $k = 1, \dots, n$ and then dividing by n, we immediately arrive at (82). □

Lemma 8 *Let Q, $\hat{Q}_k$, Q_k, A, $\hat{A}_n$ and W_n be defined by* (50), (78), (83), (80), (81) *and* (79). *Suppose $\hat{Q}_n \succ 0$ and*

$$\mathrm{eig}(\hat{Q}_n \hat{A}_n^\top \hat{A}_n) \subset \left[0, \frac{4}{3}\right). \quad (90)$$

It holds that

$$\mathrm{eig}(A Q_n A^\top) \subset \left[0, \frac{4}{3}\right). \quad (91)$$

Proof For simplicity, we use W, $\hat{Q}$ and $\hat{A}$ to take the place W_n, $\hat{Q}_n$ and $\hat{A}_n$, respectively.

It is implied by assumptions $\hat{Q} \succ 0$ and (90) that $\Theta := W^\top \hat{Q} W \succeq 0$. Recall that $S_{nn} = A_n^\top A_n = I_{d_n}$, we have

$$\begin{aligned} \rho(\Theta) &= \max_{v \in \mathbb{R}^{d_n},\, ||v||=1} \; v^\top A_n^\top \hat{A}^\top \hat{Q} \hat{A} A_n v \\ &\le \rho(\hat{A}\hat{Q}\hat{A}) \max_{v \in \mathbb{R}^{d_n},\, ||v||=1} ||A_n v||_2^2 < \frac{4}{3} ||A_n||_\mathrm{F}^2 = \frac{4}{3}. \end{aligned} \quad (92)$$

Hence, we obtain

$$0 \preceq \Theta \prec \frac{4}{3} I_{d_n}. \quad (93)$$

Recall the definition (83), we have

$$Q_n = \begin{bmatrix} I_{d-d_n} & 0 \\ -\frac{1}{2}W^\top & I_{d_n} \end{bmatrix} \begin{bmatrix} \hat{Q} & 0 \\ 0 & I_{d_n} - \frac{1}{4}W^\top \hat{Q} W \end{bmatrix} \begin{bmatrix} I & -\frac{1}{2}W \\ 0 & I_{d_n} \end{bmatrix} = J \begin{bmatrix} \hat{Q} & 0 \\ 0 & C \end{bmatrix} J^\top, \quad (94)$$

where $J := \begin{bmatrix} I_{d-d_n} & 0 \\ -\frac{1}{2}W^\top & I_{d_n} \end{bmatrix}$ and $C := I_{d_n} - \frac{1}{4}W^\top \hat{Q} W$. Apparently, we have $C \succ 0$. Together with $\hat{Q} \succ 0$, it implies $Q_n \succ 0$. Thus, we directly obtain $\mathrm{eig}(A Q_n A^\top) \subset [0, \infty)$. It remains to show

$$\rho(AQ_nA^\top) < \frac{4}{3}. \tag{95}$$

Denote $\hat{B} := \hat{A}^\top \hat{A}$, then we can write S as

$$S = A^\top A = \begin{bmatrix} \hat{B} & W \\ W^\top & I_{d_n} \end{bmatrix}.$$

We can reformulate $\rho(AQ_nA^\top)$ as follows:

$$\rho(AQ_nA^\top) = \rho\left(AJ\begin{bmatrix} \hat{Q} & 0 \\ 0 & C \end{bmatrix} J^\top A^\top\right) = \rho\left(\begin{bmatrix} \hat{Q} & 0 \\ 0 & C \end{bmatrix} J^\top A^\top AJ\right). \tag{96}$$

It is easy to verify that

$$J^\top A^\top AJ = \begin{bmatrix} I_{d-d_n} & -\frac{1}{2}W \\ 0 & I_{d_n} \end{bmatrix} \begin{bmatrix} \hat{B} & W \\ W^\top & I \end{bmatrix} \begin{bmatrix} I_{d-d_n} & 0 \\ -\frac{1}{2}W^\top & I_{d_n} \end{bmatrix} = \begin{bmatrix} \hat{B} - \frac{3}{4}WW^\top & \frac{1}{2}W \\ \frac{1}{2}W^\top & I_{d_n} \end{bmatrix}.$$

Thus,

$$Z := \begin{bmatrix} \hat{Q} & 0 \\ 0 & C \end{bmatrix} J^\top A^\top AJ = \begin{bmatrix} \hat{Q}\hat{B} - \frac{3}{4}\hat{Q}WW^\top & \frac{1}{2}\hat{Q}W \\ \frac{1}{2}CW^\top & C \end{bmatrix}. \tag{97}$$

According to (96), it suffices to prove $\rho(Z) < \frac{4}{3}$. Suppose λ is an arbitrary eigenvalue of Z, and $v \in \mathbb{R}^d$ is one of its associate eigenvector. In the rest, we only need to show

$$\lambda < \frac{4}{3} \tag{98}$$

holds. Then, using its arbitrariness, we have $\rho(Z) < \frac{4}{3}$ which implies (95), and then (91) holds.

Partition v into $v = \begin{bmatrix} v_1 \\ v_0 \end{bmatrix}$, where $v_1 \in \mathbb{R}^{d-d_n}$, $v_0 \in \mathbb{R}^{d_n}$. Then, $Zv = \lambda v$ implies that

$$\left(\hat{Q}\hat{B} - \frac{3}{4}\hat{Q}WW^\top\right)v_1 + \frac{1}{2}\hat{Q}Wv_0 = \lambda v_1, \tag{99}$$

$$\frac{1}{2}CW^\top v_1 + Cv_0 = \lambda v_0. \tag{100}$$

If $\lambda I_{d_n} - C$ is singular, i.e. λ is an eigenvalue of C. By the definition of C and (93), we have $\frac{2}{3}I_{d_n} \prec C = I_{d_n} - \frac{1}{4}\Theta \preceq I_{d_n}$, which implies that $\lambda \leq 1$, thus inequality (98) holds. In the following, we assume $\lambda I_{d_n} - C$ is nonsingular. An immediate consequence is $v_1 \neq 0$.

By (100), we obtain $v_0 = \frac{1}{2}(\lambda I_{d_n} - C)^{-1}CW^\top v_1$. Substituting this explicit formula into (99), we obtain

$$\begin{aligned}\lambda v_1 &= \left(\hat{Q}\hat{B} - \frac{3}{4}\hat{Q}WW^\top\right)v_1 + \frac{1}{4}\hat{Q}W(\lambda I_{d_n} - C)^{-1}CW^\top v_1\\ &= (\hat{Q}\hat{B} + \hat{Q}W\Phi W^\top)v_1, \qquad (101)\end{aligned}$$

where $\Phi := -I_{d_n} + \lambda[(4\lambda - 4)I_{d_n} + \Theta]^{-1}$. Since Θ is a symmetric matrix, Φ is also symmetric.

Suppose $\lambda_{\max}(\Phi) > 0$, the definition of Φ gives us

$$\theta \in \mathrm{eig}(\Theta) \Leftrightarrow -1 + \frac{\lambda}{(4\lambda - 4) + \theta} \in \mathrm{eig}(\Phi).$$

Together with $\lambda_{\max}(\Phi) > 0$, there exists $\theta \in \mathrm{eig}(\Theta)$ such that $-1 + \frac{\lambda}{(4\lambda-4)+\theta}$. If $\lambda \le 1$, (98) already holds. Otherwise, $\lambda > 1$, which implies $1 < \frac{\lambda}{(4\lambda-4)+\theta} \le \frac{\lambda}{4\lambda-4}$, and then (98) holds.

Now we assume $\lambda_{\max}(\Phi) \le 0$, i.e. $\Phi \preceq 0$. By the induction, we have $\hat{\lambda} := \rho(\hat{Q}\hat{B}) = \rho(\hat{Q}\hat{A}^\top\hat{A}) \subset \left[0, \frac{4}{3}\right)$. Due to the positive definiteness of $\hat{Q}$, there exists nonsingular $U \in \mathbb{R}^{(d-d_n)\times(d-d_n)}$ such that $\hat{Q} = U^\top U$. Let $Y := UW\Phi W^\top U^\top \in \mathbb{R}^{(d-d_n)\times(d-d_n)}$.

We have $v^\top Yv = v^\top UW\Phi W^\top U^\top v = (W^\top U^\top v)^\top \Phi(W^\top U^\top v) \le 0$ holds for all $v \in \mathbb{R}^{d-d_n}$, where the last inequality follows from $\Phi \preceq 0$. Thus, $Y \preceq 0$. Pick up arbitrary g satisfying $g > \rho(Y)$. Then, it holds that

$$\rho(gI_{d-d_n} + Y) \le g. \qquad (102)$$

From (101), we can conclude that $(g + \lambda)v_1 = (\hat{Q}\hat{B} + \hat{Q}W\Phi W^\top + gI_{d-d_n})v_1$. Consequently,

$$g + \lambda \in \mathrm{eig}(\hat{Q}\hat{B} + \hat{Q}W\Phi W^\top + gI_{d-d_n}) = \mathrm{eig}(U\hat{B}U^\top + UW\Phi W^\top U^\top + gI_{d-d_n}),$$

which implies

$$\begin{aligned}g + \lambda &\le \rho(U\hat{B}U^\top + Y + gI) \le \rho(U\hat{B}U^\top) + \rho(Y + gI)\\ &= \hat{\lambda} + \rho(Y + gI) \le \hat{\lambda} + g, \qquad (103)\end{aligned}$$

where the last inequality follows from (102). The relation (103) directly gives us that $\lambda \le \hat{\lambda} < \frac{4}{3}$. Namely, (98) also holds in this case.

We have completed the proof. □

Now we are ready to present the main proof of Lemma 2.

Proof of Lemma 2 Without loss of generality, we assume $S_{ii} = I_{d_i}$ $(i = 1, \ldots, n)$. Otherwise, we denote

$$D := \mathrm{Diag}\left(S_{11}^{-\frac{1}{2}}, \ldots, S_{nn}^{-\frac{1}{2}}\right).$$

It is easy to verify that $\tilde{Q} = D^{-1}QD^{-1}$, if $\tilde{S} = DSD$, and $\tilde{L}_\sigma$ and $\tilde{Q}$ are defined by (75) with $\tilde{S}$. It holds that

$$\mathrm{eig}(\tilde{Q}\tilde{S}) = \mathrm{eig}(D^{-1}QD^{-1}DSD) = \mathrm{eig}(D^{-1}QSD) = \mathrm{eig}(QS),$$

and $\tilde{S}_{ii} = I_{d_i}$ $(i = 1, \ldots, n)$.

It follows from the definition of A, (80), that $\mathrm{eig}(QS) = \mathrm{eig}(AQA^\top)$. Now we use mathematical induction to prove this lemma. Firstly, the assertion (74) and $Q \succ 0$ hold when $n = 1$, as $QS = I$ in this case. Next, we will prove the lemma for any $n \geq 2$ given that the assertion (74) and $Q \succ 0$ hold for $n - 1$.

By using Lemma 7, it directly follows from (82) that $AQA^\top = \frac{1}{n}\sum\limits_{k=1}^{n} AP_kQ_kP_k^\top A^\top$. Consequently,

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\lambda_{\min}(AP_kQ_kP_k^\top A^\top) &\leq \lambda_{\min}(AQA^\top) \leq \lambda_{\max}(AQA^\top)\\ &\leq \frac{1}{n}\sum_{k=1}^{n}\lambda_{\max}(AP_kQ_kP_k^\top A^\top). \end{aligned} \tag{104}$$

By the induction assumptions and Lemma 8, we obtain the relationship (91). Together with the similarity among the blocks, the relationship (91) implies

$$\mathrm{eig}(AP_kQ_kP_k^\top A^\top) \subset \left[0, \frac{4}{3}\right), \qquad \text{for all } k = 1, \ldots, n. \tag{105}$$

Substituting (105) into (104), we prove the assertion (74) for n, and hence complete the proof of Lemma 2.

Appendix B

Proof of Lemma 3 For convenience, we use the notation

$$g(\lambda; S, T) := \det\big[(\lambda - 1)^2 I + (2\lambda - 1)S + (\lambda - 1)T\big].$$

We prove this lemma by mathematical induction on the dimension d. When $d = 1$, it is easily seen that

$$g(\lambda; S, T) = \begin{cases} (\lambda-1)^0[(\lambda-1)^2 + (2\lambda-1)S + (\lambda-1)T] & \text{if } S \neq 0, \\ (\lambda-1)^1(\lambda-1+T) & \text{if } S = 0,\ T \neq 0, \\ (\lambda-1)^2 \cdot 1 & \text{if } S = 0,\ T = 0, \end{cases}$$

which means that Lemma 3 holds in this case. Suppose this lemma is valid for $d \leq k-1$. Consider the case where $d = k$.

Case 1: $S \succ 0$. In this case, $\text{Rank}(S) = \text{Rank}(S+T) = k$ and then $l = 0$. Because

$$g(\lambda; S, T) = (\lambda-1)^l g(\lambda; S, T) \quad \text{and} \quad g(1; S, T) = \det(S) > 0,$$

Lemma 3 holds in this case.

Case 2: $S \succeq 0$ but not positive definite. Let S admit the following eigenvalue decomposition

$$P^\top S P = \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & s_1 & & \\ & & & & \ddots & \\ & & & & & s_t \end{bmatrix} := D,$$

where P is a orthogonal matrix and $s_i > 0$. If we let $W = P^\top T P \succeq 0$, then

$$g(\lambda; S, T) = g(\lambda; D, W).$$

The proof proceeds by considering the following two subcases.

Case 2.1: $W_{11} = 0$. Since W is positive semidefinite, then $W_{1i} = W_{i1} = 0$ for $i = 1, 2, \ldots, k$. Note that

$$g(\lambda; D, W) = (\lambda-1)^2 g(\lambda; D', W')$$

where D' and W' are the submatrices of D and W obtained by deleting the first row and column. As we have assumed that Lemma 3 holds for $d = k-1$, there exists a polynomial $p(x)$ such that

$$g(\lambda; D, W) = (\lambda-1)^2(\lambda-1)^{2k-2-\text{Rank}D'-\text{Rank}(D'+W')} p(\lambda).$$

Note that $\text{Rank}(D') = \text{Rank}(D) = \text{Rank}(S)$ and $\text{Rank}(D'+W') = \text{Rank}(D+W) = \text{Rank}(S+T)$. Thus, we have

$$g(\lambda; S, T) = (\lambda-1)^{2k-\text{Rank}(S)-\text{Rank}(S+T)},$$

which implies that Lemma 3 is true for $d = k$ in this subcase.

Case 2.2: $W_{11} \neq 0$. Without loss of generality, assume $W_{11} = 1$. Let $w^\top = [W_{12}, \ldots, W_{1k}]$. By direct calculation, we obtain

$$g(\lambda; D, W) = (\lambda - 1)^2 g(\lambda; D', W') + (\lambda - 1) g(\lambda; D', W' - ww^\top).$$

Since $\text{Rank}(D' + W') \leq \text{Rank}(D + W) = \text{Rank}(S + T)$, there exists a polynomial $p_1(x)$ such that

$$g(\lambda; D', W') = (\lambda - 1)^{2k-2-\text{Rank}(S)-\text{Rank}(S+T)} p_1(\lambda),$$

where $p_1(1) \geq 0$. On the other hand, since $\text{Rank}(D' + W' - ww^\top) = \text{Rank}(D + W) - 1 = \text{Rank}(S + T) - 1$, there exists a polynomial $p_2(x)$ such that

$$g(\lambda; D', W' - ww^\top) = (\lambda - 1)^{2k-1-\text{Rank}(S)-\text{Rank}(S+T)} p_2(\lambda),$$

where $p_2(1) > 0$. Therefore,

$$g(\lambda; S, T) = (\lambda - 1)^{2k-\text{Rank}(S)-\text{Rank}(S+T)} (p_1(\lambda) + p_2(\lambda))$$

and then Lemma 3 holds for this subcase.

This completes the proof. □

Appendix C

Proof of Lemma 4 It is easily seen that

$$\text{Rank}(S) + \text{Rank}(\beta A^\top A) = \text{Rank} \begin{bmatrix} S & 0 \\ 0 & \beta A A^\top \end{bmatrix},$$

and therefore we need only prove that

$$\text{Rank} \begin{bmatrix} S & -A^\top \\ \beta A & 0 \end{bmatrix} = \text{Rank} \begin{bmatrix} S & 0 \\ 0 & \beta A A^\top \end{bmatrix}. \tag{106}$$

Indeed, consider the following linear system

$$\begin{bmatrix} S & -A^\top \\ \beta A & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = 0, \tag{107}$$

which is equivalent to

$$\begin{cases} Sx - A^\top \mu = 0, \\ Ax = 0. \end{cases}$$

It then holds that

$$x^\top Sx = x^\top A^\top \mu = (Ax)^\top \mu = 0,$$

and therefore $Sx = 0$ and $A^\top \mu = 0$, because $S = H + \beta A^\top A$ is positive semidefinite. This means that

$$\begin{bmatrix} S & 0 \\ 0 & \beta A A^\top \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = 0. \tag{108}$$

On the other hand, it is not difficult to verify that any solution of (108) is the solution of (107), in other words, linear systems (107) and (108) are equivalent. As a result, the rank equality (106) holds, which completes the proof. □

References

1. Agarwal, A., Negahban, S., Wainwright, M.J.: Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. Ann. Stat. **40**(2), 1171–1197 (1997)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**, 438–457 (2010)
3. Beck, A.: On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. SIAM J. Optim. **25**(1), 185–209 (2015)
4. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM J. Optim. **23**(2), 2037–2060 (2013)
5. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena-Scientific, Belmont (1999)
6. Bertsekas, D.P., Tsitsiklis, J.N.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. In: Parallel and Distributed Computation: Numerical Methods, 2nd ed. Athena Scientific, Belmont, MA (1997)
7. Bolte, J., Sabach, S.Y., Teboulle, M.: Proximal alternating linearized minimization nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014)
8. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
9. Cai, X., Han, D., Yuan, X.: On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. Comput. Optim. Appl. **66**(1), 39–73 (2017)
10. Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math. Program. **155**(1), 57–79 (2016)
11. Chen, C., Shen, Y., You, Y.: On the convergence analysis of the alternating direction method of multipliers with three blocks. *Abstract and Applied Analysis*, 2013, Article ID 183961, 7 pages
12. Chen, L., Sun, D., Toh, K.-C.: A note on the convergence of ADMM for linearly constrained convex optimization problems. Comput. Optim. Appl. **66**(2), 327–343 (2017)
13. Chen, L., Sun, D., Toh, K.-C.: An efficient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming. Math. Program. **161**(1), 237–270 (2017)
14. Cui, Y., Li, X., Sun, D., Toh, K.-C.: On the convergence properties of a majorized alternating direction method of multipliers for linearly constrained convex optimization problems with coupled objective functions. J. Optim. Theory Appl. **169**(3), 1013–1041 (2016)
15. Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. *UCLA CAM Report*, 14–51 (2014)
16. Deng, W., Lai, M., Peng, Z., Yin, W.: Parallel multi-block ADMM with $o(1/k)$ convergence. J. Sci. Comput. **71**(2), 712–736 (2017)
17. Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3), 889–916 (2016)
18. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1), 293–318 (1992)

19. Feng, C., Xu, H., Li, B.C.: An alternating direction method approach to cloud traffic management. IEEE Trans. Parallel Distrib. Syst. **28**(8), 2145–2158 (2017)
20. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**, 17–40 (1976)
21. Gao, X., Zhang, S.: First-order algorithms for convex optimization with nonseparable objective and coupled constraints. J. Oper. Res. Soc. China **5**(2), 131–159 (2017)
22. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, New York (1984)
23. Glowinski, R., Marroco, A.: Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualite, d'une classe de problemes de dirichlet non lineares. Revue Franqaise d'Automatique, Informatique et Recherche Opirationelle **9**, 41–76 (1975)
24. Han, D., Yuan, X.: A note on the alternating direction method of multipliers. J. Optim. Theory Appl. **155**(1), 227–238 (2012)
25. Han, D., Yuan, X., Zhang, W., Cai, X.: An ADM-based splitting method for separable convex programming. Comput. Optim. Appl. **54**, 343–369 (2013)
26. He, B., Tao, M., Yuan, X.: A splitting method for separable convex programming. IMA J. Numer. Anal. **35**, 394–426 (2015)
27. He, B., Tao, M., Yuan, X.: Alternating direction method with Gaussian back substitution for separable convex programming. SIAM J. Optim. **22**, 313–340 (2012)
28. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal. **50**(2), 700–709 (2012)
29. Hong, M., Chang, T., Wang, X., Razaviyayn, M., Ma, S., Luo, Z.: A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. arXiv:1401.7079 (2014)
30. Hong, M., Luo, Z.: On the linear convergence of the alternating direction method of multipliers. Math. Program. **162**, 165–199 (2017)
31. Hong, M., Luo, Z., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. **26**(1), 337–364 (2016)
32. Hong, M., Wang, X., Razaviyayn, M., Luo, Z.: Iteration complexity analysis of block coordinate descent methods. arXiv:1310.6957v2 (2014)
33. Li, M., Sun, D., Toh, K.-C.: A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. Asia Pac. J Oper. Res. **32**, 1550024 (2015)
34. Li, M., Sun, D., Toh, K.-C.: A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. SIAM J. Optim. **26**(2), 922–950 (2016)
35. Li, X., Sun, D., Toh, K.-C.: A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. Math. Program. **155**(1), 333–373 (2016)
36. Lin, T., Ma, S., Zhang, S.: Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. J. Sci. Comput. **69**(1), 52–81 (2016)
37. Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the ADMM with multi-block variables. SIAM J. Optim. **25**, 1478–1497 (2015)
38. Lin, T., Ma, S., Zhang, S.: On the sublinear convergence rate of multi-block ADMM. J. Oper. Res. Soc. China **3**(3), 251–274 (2015)
39. Lu, Z., Xiao, L.: On the complexity analysis of randomized block-coordinate descent methods. Math. Program. **152**, 615–642 (2015)
40. Monteiro, R., Svaiter, B.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**(1), 475–507 (2013)
41. Mota, J.F.C., Xavier, J.M.F., Aguiar, P.M.F., Puschel, M.: Distributed optimization with local domains: Applications in MPC and network flows. IEEE Trans. Autom. Control **60**(7), 2004–2009 (2015)
42. Peng, Y.G., Ganesh, A., Wright, J., Xu, W.L., Ma, Y.: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 2233–2246 (2012)
43. 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)
44. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**(2), 1–38 (2014)
45. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss. J. Mach. Learn. Res. **14**, 567–599 (2013)

46. Shefi, R., Teboulle, M.: On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. EURO J. Comput. Optim. **4**(1), 27–46 (2016)
47. Sun, D., Toh, K.-C., Yang, L.: A convergent 3-block semi-proximal alternating direction method of multipliers for conic programming with 4-block constraints. SIAM J. Optim. **25**(2), 882–915 (2015)
48. Sun, R., Luo, Z., Ye, Y.: On the expected convergence of randomly permuted ADMM. arXiv:1503.06387v1 (2015)
49. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**, 475–494 (2001)
50. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**, 387–423 (2009)
51. Wright, S.: Coordinate descent algorithms. Math. Program. **151**(1), 3–34 (2015)
52. Zhang Y (2010) Convergence of a class of stationary iterative methods for saddle point problems. *Rice University Technique Report*, TR10-24