On the Convergence of Multi-Block Alternating Direction Method of Multipliers and Block Coordinate Descent Method

Caihua Chen* Min Li† Xin Liu‡ Yinyu Ye§

This version: September 9, 2015

Abstract

The paper answers several open questions of the alternating direction method of multipliers (ADMM) and the block coordinate descent (BCD) method that are now wildly used to solve large scale convex optimization problems in many fields. For ADMM, it is still lack of theoretical understanding of the algorithm when the objective function is not separable across the variables. In this paper, we analyze the convergence of the 2-block ADMM for solving the linearly constrained convex optimization with coupled quadratic objective, and show that the classical ADMM point-wisely converges to a primal-dual solution pair of this problem. Moreover, we propose to use the randomly permuted ADMM (RPADMM) to solve the nonseparable multi-block convex optimization, and prove its expected convergence while applied to solve a class of quadratic programming problems. When the linear constraint vanishes, the 2-block ADMM and the randomly permuted ADMM reduce to the 2-block cyclic BCD method (also known as the alternating minimization method) and the EPOCHS[1]. Interestingly, our study provides the first iterate convergence result of the 2-block cyclic BCD method without assuming the boundedness of the iterates. Under the same setting, the sublinear convergence rate of the function values can also be verified. Moreover, we also theoretically establish the expected iterate convergence result of the multi-block EPOCHS for convex quadratic optimization which can be regarded as one of the first iterate convergence analysis of EPOCHS. Last but not least, although random permutation is indeed to make multi-block ADMM and BCD more robust, we theoretically demonstrate that EPOCHS has a worse convergence rate than the cyclic BCD does in solving 2-block convex quadratic minimization problems. Therefore, EPOCHS should be applied in caution when solving general optimization problems.

Keywords. Alternating direction method of multipliers, Block coordinate descent method, Iterate convergence, Random permutation, Large scale optimization

*International Center of Management Science and Engineering, School of Management and Engineering, Nanjing University, China. Email: chchen@nju.edu.cn.

†School of Economics and Management, Southeast University, China. Email: limin@seu.edu.cn.

‡State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China. Email: liuxin@lsec.cc.ac.cn.

§Department of Management Science and Engineering, School of Engineering, Stanford University, USA; and International Center of Management Science and Engineering, School of Management and Engineering, Nanjing University. Email: yyye@stanford.edu.

[1] The randomly permuted cyclic BCD method, also known as the "sampling without replacement" variant of randomized BCD. For simplicity, we call it EPOCHS as suggested in the survey paper [49].

1 Introduction

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

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

where $\theta_i : \mathbb{R}^{d_i} \mapsto (-\infty, +\infty]$ 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 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 KKT point of (1), if it satisfies

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

The set consisting of the KKT points of (1) is assumed to be nonempty. The problem (1) has found many applications in signal and imaging processing, machine learning, statistics and engineering computation, see e.g., [1, 14, 18, 28, 39, 40].

The augmented Lagrangian function of (1) is

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

where $\mu \in \mathbb{R}^m$ is the Lagrangian multiplier and $\beta > 0$ is the penalty parameter. To solve problem (1), we consider an n-block alternating direction method of multiplier (ADMM), 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.,

$$
\left\{
\begin{array}{l}
x_1^{k+1} := \arg\min_{x_1} \mathcal{L}_\beta(x_1, x_2^k, \ldots, x_n^k; \mu^k), \\
x_2^{k+1} := \arg\min_{x_2} \mathcal{L}_\beta(x_1^{k+1}, x_2, x_3^k, \ldots, x_n^k; \mu^k), \\
\cdots\cdots \\
x_n^{k+1} := \arg\min_{x_n} \mathcal{L}_\beta(x_1^{k+1}, x_2^{k+1}, \ldots, x_{n-1}^{k+1}, x_n; \mu^k), \\
\mu^{k+1} := \mu^k - \beta(\sum_{i=1}^n A_i x_i^{k+1} - b).
\end{array}
\right.
\tag{4}
$$

Notice that the algorithmic scheme (4) reduces to the classical ADMM when there are only two blocks ($n = 2$) and the coupled objective vanishes ($H = 0$ and $g = 0$). The ADMM was originally introduced in early 1970s [19, 22] and its convergence has been studied extensively in the literature [6, 15–17, 21, 27, 38]. Possibly due to its wide versatility and applicability in multifarious areas, the ADMM has gained great popularity in solving nowadays optimization problems especially in modern big data related problems; we refer to [8] for a survey on the modern applications of ADMM.

For the case of $n \geq 3$, lots of research efforts have been devoted to analyzing the convergence of the 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 solving a nonsingular square system of linear equations. Various methods have been proposed to overcome the divergence issue of the multi-block ADMM. One typical solution is to combine correction steps with the output of the n-block ADMM (4), see e.g., [24–26]. If at least $n-2$ functions in the objective are strongly convex, [9, 11, 23, 36, 50] proved the global convergence of (4) provided that the penalty parameter β is restricted to a specific range. Without strong convexity, it was shown in [29] that the n-block ADMM with a small dual stepsize, where the multiplier update (4) is replaced by

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

is linearly convergent provided that the objective function satisfies certain error bound condition. Some very recent works [34, 35] showed the convergence of the multi-block ADMM under some other kinds of conditions, and [12, 33, 45] put forward some convergent proximal variants of the multi-block ADMM for solving convex linear/quadratic conic programming problems. On the other hand, recent paper [46] proposed a randomly modified variant of the multi-block ADMM (4), called the randomly permuted ADMM (RPADMM). At each step, the RPADMM draws a permutation of $\{1, 2, \dots, n\}$ uniformly at random (or block sampling without replacement), and update the primal variables x_i $(i = 1, 2, \dots, n)$ in the order of the chosen permutation followed by the regular multiplier update. Surprisingly, it is proved in [46] that the RPADMM is convergent in expectation for solving any nonsingular square system of linear equations.

In contrast to the separable case, there are very limited works revealing the convergence properties of the n-block ADMM for (1) with nonseparable objective even for $n = 2$. In [28], the authors demonstrated that when the problem (1) is convex but not necessarily separable[2], and certain error bound condition is satisfied, then the ADMM iteration converges to some primal-dual optimal solution, provided that a sufficiently small stepsize is attached in the update of the multiplier. Despite of the conservativeness, the stepsize usually depends on some unknown parameters associated with the error bound and thus might not be easy 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 variant with a large stepsize $\tau \geq 1$ in practice. However, as mentioned in [30], **"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 another line, paper [14] investigated the convergence of a majorized ADMM for the convex optimization problem with a coupled smooth function in the objective, which includes the 2-block ADMM (4) for (1) as a special case. The convergence is established when the subproblems of the ADMM admit unique solutions and H, A_1 and A_2 satisfy some additional restrictions; see Remark 4.2 in [14] for details. Very recently, paper [20] 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 derived results in [20] are not applicable to the scheme (4) for problem (1).

In this present paper, we analyze the iterate convergence of the ADMM (4) and its randomly permuted version for solving the nonseparable convex optimization problem (1). The main contributions of our paper are threefold. First, we prove that the 2-block ADMM is convergent

[2]The models considered in [28, 30] are more general than problem (1) as the authors of [28, 30] actually allow generally nonseparable smooth function in the objective, but in (1) the coupled objective is a quadratic function.

for (1) only under a condition to ensure the unique solutions of the subproblems. Our condition is the weakest condition to expect the iterate convergence for the ADMM since, as we will see in Section 2, it is not only sufficient but also necessary for the convergence of the ADMM applied to some special problems. Our analysis partially answers the open question mentioned in [30] on the convergence of ADMM for nonseparable convex optimization problems. Second, we extend the RPADMM proposed in [46] 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 [46]. Significantly, unlike the non-singular system case, the objective is nonseparable and its optimal solution set is no longer a singleton in our setting.

Probably more importantly, when restricted to the unconstrained case, that is, A_i $(i = 1, \cdots, n)$ and b are absent, the ADMM and RPADMM reduce to the cyclic block coordinate descent (BCD) method (also known as the alternating minimization method) and the randomly permuted cyclic BCD method. An implication of our work is the iterate convergence of the 2-block cyclic BCD method for the whole sequence and, especially, the expected convergence randomly permuted multi-block BCD. Although the literature of BCD type methods is quite vast (see [3–5, 37, 41, 43, 44, 47, 48] for example), there are very few results on the iterate convergence of the BCD type methods. As mentioned in [7], **"in all these works (BCD or its proximal variants) only convergence of the subsequences can be established".** Upon assuming that the Kurdyka-Łojasiewicz property holds on the objective and the iterates are bounded, paper [2] and [7] established the iterate convergence of the proximal BCD and proximal alternating linearized minimization, respectively. To the best of our knowledge, our convergence of the 2-block BCD method is the first iterate convergence result for original BCD without assuming the boundedness of the iterates and only under the unique solutions type condition of subproblems. Moreover, the sublinear convergence rate of the function values can also be obtained with the absence of the assumption on the sequence boundedness.

For randomly permuted multi-block BCD, also known as the "sampling without replacement" variant of randomized BCD and called "EPOCHS" in a very recent survey [49], it is claimed in [49] that the EPOCHS tends to converge faster than the randomized BCD does and the classical cyclic version usually performs even worse. Also see [43] for some numerical advantages of the EPOCHS compared with the randomized BCD and cyclic BCD. In fact, it is stated in [49] that "this kind of randomization (the EPOCHS) has been shown in several contexts to be superior to the sampling with replacement scheme analyzed above, but a theoretical understanding this phenomenon remains elusive." Randomized BCD ("sampling with replacement") has already been extensively studied (see [42] for example), but its theoretical analysis does not apply on EPOCHS. Although the function value convergence results (see [3, 31, 47] for example) for cyclic or essential cyclic BCD can be simply extended to EPOCHS, as these analysis techniques are independent of permutation, there still lacks of direct theoretical analysis on the iterate convergence of EPOCHS. Our expected iterate convergence of EPOCHS for solving quadratic minimization problems can be regarded as the first direct analysis on the iterate convergence of the "sampling without replacement" variant of randomized BCD. On the other hand, we also prove that EPOCHS has a worse convergence rate than the cyclic BCD does in solving quadratic minimization problems. Thus, EPOCHS should be used in caution for solving general optimization problems.

The rest of our paper is organized as follows. In Section 2, we prove the iterate convergence of the 2-block ADMM and the cyclic BCD for solving the linearly constrained optimization problem with coupled quadratic objective (1) and its unconstrained variant, respectively. Section

3 illustrates the expected convergence of the RPADMM and the EPOCHS for solving a class of linear constrained quadratic optimization problems and its unconstrained variant, respectively. Finally, we conclude our paper in Section 4.

2 Convergence of 2-Block ADMM and Cyclic BCD

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

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

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

$$\phi(x_1, x_2) := \frac{1}{2} x_1^\top H_{11} x_1 + x_1^\top H_{12} x_2 + \frac{1}{2} x_2^\top H_{22} x_2 + g_1^\top x_1 + g_2^\top x_2. \tag{5}$$

As a result, the problem under consideration can be written as

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^d} & \theta(x) := \theta_1(x_1) + \theta_2(x_2) + \phi(x_1, x_2) \\ \text{s.t.} & A_1 x_1 + A_2 x_2 = b. \end{array} \tag{6}$$

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

$$(x_1 - \hat{x}_1)^\top (w_1 - \hat{w}_1) \geq \|x_1 - \hat{x}_1\|_{\Sigma_1}^2 \quad \forall\, x_1, \hat{x}_1 \in \operatorname{dom}(\theta_1),\ w_1 \in \partial\theta_1(x_1), \hat{w}_1 \in \partial\theta_1(\hat{x}_1) \tag{7}$$

and

$$(x_2 - \hat{x}_2)^\top (w_2 - \hat{w}_2) \geq \|x_2 - \hat{x}_2\|_{\Sigma_2}^2 \quad \forall\, x_2, \hat{x}_2 \in \operatorname{dom}(\theta_2),\ w_2 \in \partial\theta_2(x_2), \hat{w}_2 \in \partial\theta_2(\hat{x}_2), \tag{8}$$

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

$$x := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad w := \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad \text{and} \qquad \Sigma := \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{9}$$

for any $x, \hat{x} \in \operatorname{dom}(\theta_1) \times \operatorname{dom}(\theta_2)$, $w \in \partial\theta_1(x_1) \times \partial\theta_2(x_2)$ and $\hat{w} \in \partial\theta_1(\hat{x}_1) \times \partial\theta_2(\hat{x}_2)$, we have

$$(x - \hat{x})^\top (w - \hat{w}) \geq \|x - \hat{x}\|_\Sigma^2. \tag{10}$$

The following lemma presents some contract properties of the 2-block ADMM scheme (4), which plays an important role in the subsequent analysis.

Lemma 2.1. *Assume the ADMM scheme* (4) *with* $n = 2$ *for problem* (6) *is well defined. Let* $\{(x_1^k, x_2^k, \mu^k)\}$ *be a sequence generated by the* 2*-block ADMM, then the following statements hold.*

(i) *Let* $(\bar{x}_1, \bar{x}_2, \bar{\mu})$ *be any given KKT point of problem* (6)*, then we have*

$$\begin{aligned} &\Big(\frac{7}{8}\|x^k - \bar{x}\|_{H+\Sigma}^2 + \frac{1}{2}\|x_2^k - \bar{x}_2\|_{H_{22}+\Sigma_2+\beta A_2^\top A_2}^2 + \frac{1}{2\beta}\|\mu^k - \bar{\mu}\|^2\Big) \\ &-\Big(\frac{7}{8}\|x^{k+1} - \bar{x}\|_{H+\Sigma}^2 + \frac{1}{2}\|x_2^{k+1} - \bar{x}_2\|_{H_{22}+\Sigma_2+\beta A_2^\top A_2}^2 + \frac{1}{2\beta}\|\mu^{k+1} - \bar{\mu}\|^2\Big) \\ &\geq \frac{1}{16}\|x^{k+1} - x^k\|_{H+\Sigma}^2 + \frac{1}{6}\|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2+3\beta A_2^\top A_2}^2 + \frac{1}{2\beta}\|\mu^{k+1} - \mu^k\|^2. \end{aligned} \tag{11}$$

(ii) *It holds that*

$$\begin{cases} \lim\limits_{k\to\infty} d(0,\ \partial\theta_1(x_1^{k+1}) + \nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) - A_1^\top \mu^{k+1}) = 0, \\ \lim\limits_{k\to\infty} d(0,\ \partial\theta_2(x_2^{k+1}) + \nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) - A_2^\top \mu^{k+1}) = 0, \\ \lim\limits_{k\to\infty} \|A_1 x_1^{k+1} + A_2 x_2^{k+1} - b\| = 0, \end{cases} \tag{12}$$

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

Proof. (i) Notice that

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

where $\phi(\cdot,\cdot)$ is defined in (5). By using the definitions of ϕ and μ^{k+1}, from the above formula, we obtain that

$$\begin{cases} -\nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) + A_1^\top \mu^{k+1} + (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) \in \partial\theta_1(x_1^{k+1}), \\ -\nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) + A_2^\top \mu^{k+1} \in \partial\theta_2(x_2^{k+1}). \end{cases} \tag{13}$$

Since $(\bar{x}_1, \bar{x}_2, \bar{\mu})$ is a KKT point of (6), it then holds

$$\begin{cases} -\nabla_{x_1}\phi(\bar{x}_1, \bar{x}_2) + A_1^\top \bar{\mu} \in \partial\theta_1(\bar{x}_1), \\ -\nabla_{x_2}\phi(\bar{x}_1, \bar{x}_2) + A_2^\top \bar{\mu} \in \partial\theta_2(\bar{x}_2), \\ A_1\bar{x}_1 + A_2\bar{x}_2 = b. \end{cases} \tag{14}$$

From (10), (13) and (14), we obtain

$$\begin{aligned} &\|x^{k+1} - \bar{x}\|_\Sigma^2 \\ &\le (x_1^{k+1} - \bar{x}_1)^\top \Big\{ \big[-\nabla_{x_1}\phi(x_1^{k+1}, x_2^{k+1}) + A_1^\top \mu^{k+1} + (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) \big] \\ &\quad - \big[-\nabla_{x_1}\phi(\bar{x}_1, \bar{x}_2) + A_1^\top \bar{\mu} \big] \Big\} + (x_2^{k+1} - \bar{x}_2)^\top \Big\{ \big[-\nabla_{x_2}\phi(x_1^{k+1}, x_2^{k+1}) + A_2^\top \mu^{k+1} \big] \\ &\quad - \big[-\nabla_{x_2}\phi(\bar{x}_1, \bar{x}_2) + A_2^\top \bar{\mu} \big] \Big\} \\ &= -(x_1^{k+1} - \bar{x}_1)^\top A_1^\top (\bar{\mu} - \mu^{k+1}) - (x_2^{k+1} - \bar{x}_2)^\top A_2^\top (\bar{\mu} - \mu^{k+1}) \\ &\quad + (x_1^{k+1} - \bar{x}_1)^\top (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) - (x^{k+1} - \bar{x})^\top \big(\nabla\phi(x_1^{k+1}, x_2^{k+1}) - \nabla\phi(\bar{x}_1, \bar{x}_2) \big) \\ &= \frac{1}{\beta}(\mu^{k+1} - \mu^k)^\top (\bar{\mu} - \mu^{k+1}) + (x_1^{k+1} - \bar{x}_1)^\top (H_{12} + \beta A_1^\top A_2)(x_2^{k+1} - x_2^k) - \|x^{k+1} - \bar{x}\|_H^2. \end{aligned} \tag{15}$$

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

$$\begin{aligned} &\beta(x_1^{k+1} - \bar{x}_1)^\top A_1^\top A_2 (x_2^{k+1} - x_2^k) \\ = \ & -\beta(A_2 x_2^{k+1} - A_2\bar{x}_2)^\top (A_2 x_2^{k+1} - A_2 x_2^k) + \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b)^\top (A_2 x_2^{k+1} - A_2 x_2^k) \\ = \ & \frac{\beta}{2}(\|A_2 x_2^k - A_2\bar{x}_2\|^2 - \|A_2 x_2^{k+1} - A_2\bar{x}_2\|^2) - \frac{\beta}{2}\|A_2 x_2^{k+1} - A_2 x_2^k\|^2 \\ & \quad + \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b)^\top (A_2 x_2^{k+1} - A_2 x_2^k) \end{aligned} \tag{16}$$

and

$$\frac{1}{\beta}(\mu^{k+1}-\mu^k)^\top(\bar{\mu}-\mu^{k+1}) = \frac{1}{2\beta}(\|\mu^k-\bar{\mu}\|^2 - \|\mu^{k+1}-\bar{\mu}\|^2 - \|\mu^{k+1}-\mu^k\|^2). \tag{17}$$

On the other hand, it follows from (13) that

$$-\nabla_{x_2}\phi(x_1^{k+1},x_2^{k+1}) + A_2^\top\mu^{k+1} \in \partial\theta_2(x_2^{k+1}) \text{ and } -\nabla_{x_2}\phi(x_1^k,x_2^k) + A_2^\top\mu^k \in \partial\theta_2(x_2^k),$$

which together with (8) implies

$$(x_2^{k+1}-x_2^k)^\top\big(-\nabla_{x_2}\phi(x_1^{k+1},x_2^{k+1}) + A_2^\top\mu^{k+1} + \nabla_{x_2}\phi(x_1^k,x_2^k) - A_2^\top\mu^k\big) \geq \|x_2^{k+1}-x_2^k\|^2_{\Sigma_2}.$$

By using the definitions of ϕ and μ^{k+1}, we derive from the above inequality that

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

Substituting (16), (17) and the above inequality into (15), we further get

$$\begin{aligned}&\frac{1}{2\beta}\big(\|\mu^k-\bar{\mu}\|^2-\|\mu^{k+1}-\bar{\mu}\|^2\big) + \frac{\beta}{2}\big(\|A_2x_2^k-A_2\bar{x}_2\|^2-\|A_2x_2^{k+1}-A_2\bar{x}_2\|^2\big)\\ &\geq \|x^{k+1}-\bar{x}\|^2_{H+\Sigma} + \frac{1}{2\beta}\|\mu^{k+1}-\mu^k\|^2 + \frac{1}{2}\|x_2^{k+1}-x_2^k\|^2_{\beta A_2^\top A_2} - (x_2^{k+1}-x_2^k)^\top H_{12}^\top(x_1^{k+1}-\bar{x}_1)\\ &\quad+\|x_2^{k+1}-x_2^k\|^2_{H_{22}+\Sigma_2} - (x_2^{k+1}-x_2^k)^\top H_{12}^\top(x_1^k-x_1^{k+1}).\end{aligned} \tag{18}$$

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

$$\begin{aligned}&(x_2^{k+1}-x_2^k)^\top H_{12}^\top(x_1^{k+1}-\bar{x}_1) - \|x_2^{k+1}-x_2^k\|^2_{H_{22}+\Sigma_2} + (x_2^{k+1}-x_2^k)^\top H_{12}^\top(x_1^k-x_1^{k+1})\\ &= \begin{bmatrix} 0 \\ x_2^{k+1}-x_2^k\end{bmatrix}^\top (H+\Sigma)(x^k-\bar{x}) - (x_2^{k+1}-x_2^k)^\top(H_{22}+\Sigma_2)(x_2^{k+1}-\bar{x}_2)\\ &\leq \frac{3}{4}\|x^k-\bar{x}\|^2_{H+\Sigma} + \frac{1}{3}\|x_2^{k+1}-x_2^k\|^2_{H_{22}+\Sigma_2} - \frac{1}{2}\|x_2^{k+1}-x_2^k\|^2_{H_{22}+\Sigma_2}\\ &\quad+\frac{1}{2}\big(\|x_2^k-\bar{x}_2\|^2_{H_{22}+\Sigma_2} - \|x_2^{k+1}-\bar{x}_2\|^2_{H_{22}+\Sigma_2}\big)\\ &= \frac{3}{4}\|x^k-\bar{x}\|^2_{H+\Sigma} - \frac{1}{6}\|x_2^{k+1}-x_2^k\|^2_{H_{22}+\Sigma_2} + \frac{1}{2}\big(\|x_2^k-\bar{x}_2\|^2_{H_{22}+\Sigma_2} - \|x_2^{k+1}-\bar{x}_2\|^2_{H_{22}+\Sigma_2}\big).\end{aligned} \tag{19}$$

Using the elementary inequality $2(\|a\|^2_{H+\Sigma}+\|b\|^2_{H+\Sigma}) \geq \|a+b\|^2_{H+\Sigma}$, we obtain that

$$\begin{aligned}&\|x^{k+1}-\bar{x}\|^2_{H+\Sigma} - \frac{3}{4}\|x^k-\bar{x}\|^2_{H+\Sigma}\\ &= \frac{7}{8}\big(\|x^{k+1}-\bar{x}\|^2_{H+\Sigma} - \|x^k-\bar{x}\|^2_{H+\Sigma}\big) + \frac{1}{8}\big(\|x^{k+1}-\bar{x}\|^2_{H+\Sigma} + \|x^k-\bar{x}\|^2_{H+\Sigma}\big)\\ &\geq \frac{7}{8}\big(\|x^{k+1}-\bar{x}\|^2_{H+\Sigma} - \|x^k-\bar{x}\|^2_{H+\Sigma}\big) + \frac{1}{16}\|x^{k+1}-x^k\|^2_{H+\Sigma}.\end{aligned} \tag{20}$$

Substituting (19) and (20) into (18), we get (11).

(ii) From (11), we immediately see that

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

and it therefore holds

$$\lim_{k\to\infty} \|x^{k+1} - x^k\|_{H+\Sigma} = 0, \quad \lim_{k\to\infty} \|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2+3\beta A_2^\top A_2} = 0 \tag{21}$$

and

$$\lim_{k\to\infty} \|A_1 x_1^{k+1} + A_2 x_2^{k+1} - b\| = \frac{1}{\beta} \lim_{k\to\infty} \|\mu^{k+1} - \mu^k\| = 0. \tag{22}$$

Since $H + \Sigma$ and $H_{22} + \Sigma_2$ are positive semidefinite matrices, we deduce from (21) that

$$\begin{cases} \lim_{k\to\infty} (H + \Sigma)(x^{k+1} - x^k) = 0, \\ \lim_{k\to\infty} (H_{11} + \Sigma_1)(x_1^{k+1} - x_1^k) + H_{12}(x_2^{k+1} - x_2^k) = 0, \\ \lim_{k\to\infty} \|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2} = 0, \\ \lim_{k\to\infty} \|A_2(x_2^{k+1} - x_2^k)\| = 0, \end{cases}$$

and hence

$$\begin{cases} \lim_{k\to\infty} \|x_1^{k+1} - x_1^k\|_{H_{11}+\Sigma_1} \le \lim_{k\to\infty} \left(\|x^{k+1} - x^k\|_{H+\Sigma} + \|x_2^{k+1} - x_2^k\|_{H_{22}+\Sigma_2}\right) = 0, \\ \lim_{k\to\infty} H_{12}(x_2^{k+1} - x_2^k) = -\lim_{k\to\infty} (H_{11} + \Sigma_1)(x_1^{k+1} - x_1^k) = 0, \\ \lim_{k\to\infty} A_2(x_2^{k+1} - x_2^k) = 0. \end{cases}$$

This, together with (13) and (22), proves the assertion (12). □

For establishing the iterate convergence of the ADMM, we make the following assumption:

Assumption 2.1. *Symmetric matrix $H \succeq 0$ and*

$$\begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} A_1^\top A_1 & 0 \\ 0 & A_2^\top A_2 \end{bmatrix} \succ 0. \tag{23}$$

It is worthy of emphasizing that Assumption 2.1 means that the subproblems of the 2-block ADMM admit unique solutions since Assumption 2.1 holds if and only if

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

for any $\beta > 0$. However, the optimal solution of original problem (6) is not necessarily unique. Now, we are ready to present the iterate convergence result of the 2-block ADMM for the nonseparable convex optimization model (6).

Theorem 2.1. *Suppose Assumption 2.1 holds. Let $\{(x_1^k, x_2^k, \mu^k)\}$ be generated by the ADMM (4) with $n = 2$ for solving problem (6). Then the sequence $\{(x_1^k, x_2^k, \mu^k)\}$ converges to a KKT point of (6).*

Proof. It follows from the inequality (11) that the sequences $\{(H + \Sigma)x^{k+1}\}$, $\{(H_{22} + \Sigma_2 + \beta A_2^\top A_2)x_2^{k+1}\}$ and $\{\mu^{k+1}\}$ are all bounded. Together with the positiveness of $H_{22} + \Sigma_2 +$

$\beta A_2^\top A_2 \succ 0$, we directly obtain the boundedness of $\{x_2^{k+1}\}$. Note that $A_1\bar{x}_1 + A_2\bar{x}_2 = b$. By using the triangle inequalities

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

and

$$\|x_1^{k+1} - \bar{x}_1\|_{H_{11}+\Sigma_1} \leq \|x^{k+1} - \bar{x}\|_{H+\Sigma} + \|x_2^{k+1} - \bar{x}_2\|_{H_{22}+\Sigma_2},$$

we further obtain the boundedness of the sequences $\{A_1x_1^{k+1}\}$ and $\{(H_{11} + \Sigma_1)x_1^{k+1}\}$, and hence $\{(H_{11} + \Sigma_1 + \beta A_1^\top A_1)x_1^{k+1}\}$ is bounded. This, together with the positive definiteness of $H_{11} + \Sigma_1 + \beta A_1^\top A_1$, 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

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

Then, by setting $k = k_i - 1$ and invoking the upper semicontinuous of $\partial\theta_1$ and $\partial\theta_2$ in (12), we obtain that

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

which means $(x_1^\infty, x_2^\infty, \mu^\infty)$ is a KKT point of (6). Hence (11) 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 $k \geq k_i$ that

$$\begin{aligned} &\frac{7}{8}\|x^{k+1} - x^\infty\|_{H+\Sigma}^2 + \frac{1}{2}\|x_2^{k+1} - x_2^\infty\|_{H_{22}+\Sigma_2+\beta A_2^\top A_2}^2 + \frac{1}{2\beta}\|\mu^{k+1} - \mu^\infty\|^2 \\ &\leq \frac{7}{8}\|x^{k_i} - x^\infty\|_{H+\Sigma}^2 + \frac{1}{2}\|x_2^{k_i} - x_2^\infty\|_{H_{22}+\Sigma_2+\beta A_2^\top A_2}^2 + \frac{1}{2\beta}\|\mu^{k_i} - \mu^\infty\|^2. \end{aligned} \tag{24}$$

Note that

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

one can deduce from (24) that

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

which implies

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

and

$$\lim_{k\to\infty} \|x^{k+1} - x^\infty\|_{H+\Sigma}^2 = 0. \tag{25}$$

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

$$\lim_{k\to\infty} x_2^{k+1} = x_2^\infty. \tag{26}$$

On the other hand, it is easily seen that

$$\begin{aligned}\|A_1(x_1^{k+1}-x_1^\infty)\| &\le \|A_1x_1^{k+1}+A_2x_2^{k+1}-(A_1x_1^\infty+A_2x_2^\infty)\|+\|A_2(x_2^{k+1}-x_2^\infty)\| \\ &= \|A_1x_1^{k+1}+A_2x_2^{k+1}-b\|+\|A_2(x_2^{k+1}-x_2^\infty)\|\to 0, \qquad (27)\end{aligned}$$

as $k\to\infty$, where (27) follows from (12) and (26). Then, we obtain

$$\begin{aligned}&\|x_1^{k+1}-x_1^\infty\|^2_{H_{11}+\Sigma_1+\beta A_1^\top A_1} = \|x_1^{k+1}-x_1^\infty\|^2_{H_{11}+\Sigma_1}+\beta\|A_1(x_1^{k+1}-x_1^\infty)\|^2 \\ &\le (\|x^{k+1}-x^\infty\|_{H+\Sigma}+\|x_2^{k+1}-x_2^\infty\|_{H_{22}+\Sigma_2})^2+\beta\|A_1(x_1^{k+1}-x_1^\infty)\|^2,\end{aligned}$$

where the "$\le$" follows the triangle inequality of norms. This, together with (25), (26), (27) and the positive definiteness of $H_{11}+\Sigma_1+\beta A_1^\top A_1$, 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 $(x_1^\infty,x_2^\infty,\mu^\infty)$. The proof is completed. $\square$

Remark 2.1. *In fact, the iterate convergence of* 2*-block ADMM can also be guaranteed if there is a fixed stepsize* $\gamma\in(0,(1+\sqrt{5})/2)$ *for the dual update. Namely, the classical ADMM is extended as following*

$$\begin{cases} x_1^{k+1} := \arg\min\limits_{x_1}\ \mathcal{L}_\beta(x_1,x_2^k;\mu^k), \\ x_2^{k+1} := \arg\min\limits_{x_2}\ \mathcal{L}_\beta(x_1^{k+1},x_2;\mu^k), \\ \mu^{k+1} := \mu^k-\gamma\beta(A_1x_1^{k+1}+A_2x_2^{k+1}-b), \end{cases} \qquad (28)$$

where $\beta>0$ *and* $\gamma\in(0,(1+\sqrt{5})/2)$*. Under the same conditions of Theorem* 2.1*, we can similarly prove the global iterate convergence of* (28)*. For brevity, we omit the details here.*

We remark that Assumption 2.1 actually acts as the weakest condition to guarantee the iterate convergence of the ADMM for solving problem (6) in some sense. Firstly, if Assumption 2.1 is violated, the solution sets of subproblems in (4) might be empty and the 2-block ADMM scheme is not well defined, see [13] for an illustration. Secondly, it can be seen from the following corollary that Assumption 2.1 is not only a sufficient but also necessary condition for the iterate convergence of the 2-block ADMM for solving problem (6). For example, even for solving a coupled quadratic minimization problem, the conditions we proposed are already tight.

Corollary 2.1. *Assume that* $H\succeq 0$ *and problem* (6) *is a quadratic programming, that is* $\theta_1(x_1)\equiv 0$ *and* $\theta_2(x_2)\equiv 0$*. Then, any sequence generated by the* 2*-block ADMM is convergent if and only if Assumption* 2.1 *holds.*

Proof. The "if" part follows from Theorem 2.1 immediately. For the "only if" part, we prove that if Assumption 2.1 fails to hold, there must exist some sequence generated by the 2-block ADMM which is divergent. Indeed, let $\{(x_1^k,x_2^k,\mu^k)\}$ be a sequence generated by the 2-block ADMM, i.e.,

$$\begin{cases} x_1^{k+1} \in \arg\min\limits_{x_1}\ \mathcal{L}_\beta(x_1,x_2^k;\mu^k), \\ x_2^{k+1} \in \arg\min\limits_{x_2}\ \mathcal{L}_\beta(x_1^{k+1},x_2;\mu^k), \\ \mu^{k+1} = \mu^k-\beta(A_1x_1^{k+1}+A_2x_2^{k+1}-b). \end{cases} \qquad (29)$$

If the sequence is divergent, then the "only if" part of this corollary holds. Thus we only need to consider the case where $\{(x_1^k, x_2^k, \mu^k)\}$ converges. Since $H_{ii} + \beta A_i^\top A_i\,(i = 1, 2)$ are not positive definite, there exists a nonzero vector $(\bar{y}_1, \bar{y}_2)$ such that

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

or equivalently,

$$H_{ii}\bar{y}_i = 0 \qquad \text{and} \qquad A_i\bar{y}_i = 0 \quad \forall\, i = 1, 2. \tag{30}$$

By using the fact that $0 \preceq H \preceq 2\begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix}$, we have $H\bar{y} = 0$, and hence it holds that

$$H_{12}\bar{y}_2 = 0 \qquad \text{and} \qquad H_{12}^\top \bar{y}_1 = 0. \tag{31}$$

By (29), (30) and (31), it is easily seen for any $k \geq 1$ that

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

and

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

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 ADMM. Thus, for the coupled convex quadratic programming, Assumption 2.1 is also necessary for the iterate convergence. This completes the proof. $\square$

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

Corollary 2.2. *Assume $H \succeq 0$ and $H_{ii} + \Sigma_i \succ 0$ for $i = 1, 2$. Let $\{(x_1^k, x_2^k)\}$ be generated by the 2-block cyclic BCD method for solving the following unconstrained optimization problem:*

$$\min_{x \in \mathbb{R}^d}\ \theta_1(x_1) + \theta_2(x_2) + \frac{1}{2}x^\top H x + g^\top x. \tag{32}$$

Then the whole sequence $\{(x_1^k, x_2^k)\}$ converges to an optimal solution of (32).

Despite the fact that various sublinear convergence rates have been established for BCD type methods (see [3,4,41,44] for example), none of them can be directly applied to our 2-block cyclic BCD for problem (32) since its level set is not necessarily bounded which violates the common assumption in the above mentioned analysis. Note that under the condition in Corollary 2.2, the sequence generated by the BCD method is bounded and then we can obtain the $O(1/k)$ global convergence rate of the method by using the main techniques developed in [4]. Specifically, we assume

$$\|x_1^k - \bar{x}_1\| \le R_1 \quad \text{and} \quad \|x_2^k - \bar{x}_2\| \le R_2,$$

where $(\bar{x}_1, \bar{x}_2)$ is an optimal solution of (32). Let L_1 and L_2 be the largest eigenvalue of H_{11} and H_{22}, respectively. Then, by invoking [4, Theorem 3.9], we immediate have the following corollary.

Corollary 2.3. *Assume $H \succeq 0$ and $H_{ii} + \Sigma_i \succ 0$ for $i = 1, 2$. Let $\{(x_1^k, x_2^k)\}$ be generated by the 2-block cyclic BCD method for solving problem (32). Then it holds that*

$$\theta(x^k) - \theta^* \le \max\left\{\left(\frac{1}{2}\right)^{\frac{k-1}{2}}(\theta(x_0) - \theta^*), \frac{8\min\{L_1, L_2\}(R_1^2 + R_2^2)}{k-1}\right\}$$

where θ^* is the function value of (32).

3 Convergence of Multi-block RPADMM and EPOCHS

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

$$\begin{cases} x_{\sigma(1)}^{k+1} := \arg\min_{x_{\sigma(1)}} \mathcal{L}_\beta(x_{\sigma(1)}, x_{\sigma(2)}^k, \dots, x_{\sigma(n)}^k; \mu^k), \\ x_{\sigma(2)}^{k+1} := \arg\min_{x_{\sigma(2)}} \mathcal{L}_\beta(x_{\sigma(1)}^{k+1}, x_{\sigma(2)}, x_{\sigma(3)}^k, \dots, x_{\sigma(n)}^k; \mu^k), \\ \cdots\cdots \\ x_{\sigma(n)}^{k+1} := \arg\min_{x_{\sigma(n)}} \mathcal{L}_\beta(x_{\sigma(1)}^{k+1}, x_{\sigma(2)}^{k+1}, \dots, x_{\sigma(n-1)}^{k+1}, x_{\sigma(n)}; \mu^k), \\ \mu^{k+1} := \mu^k - \beta(\sum_{i=1}^n A_i x_i^{k+1} - b), \end{cases} \tag{33}$$

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

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

It was shown in [46] that the 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 or not the RPADMM is even convergent in expectation for solving the following simpler linearly constrained quadratic minimization problem,

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \theta(x) := \frac{1}{2} x^\top H x + g^\top x \\ \text{s.t.} \quad & \sum_{i=1}^n A_i x_i = b, \end{aligned} \tag{34}$$

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

Assumption 3.1. *Assume $H \succeq 0$, and*

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

Although our current result is restricted for nonseparable quadratic minimization, a special case of (1), it servers as a good indicator of the expected convergence of the RPADMM for more general case. It is noteworthy that our result is a non-trivial extension of the result in [46] since in our setting the problem under consideration is more general than that considered in [46]. For example, the optimal solution set of (34) may be no longer a singleton, in which case the spectral radius of the algorithm mapping may not be strictly less than 1. But this fact plays a key role in establishing their result.

3.1 Proof Outline and Preliminaries

For convenience, we follow the notation in [46], and give the iterative scheme of RPADMM in a matrix form. Let $L_\sigma \in \mathbb{R}^{d \times d}$ be a $n \times n$ block matrix defined by

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

and R_σ is defined as

$$R_\sigma := L_\sigma - (H + \beta A^\top A) := L_\sigma - S. \tag{35}$$

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

$$z^{k+1} := M_\sigma z^k + \bar{L}_\sigma^{-1} \bar{b}, \tag{36}$$

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

$$Q := E_\sigma(L_\sigma^{-1}) = \frac{1}{n!} \sum_{\sigma \in \Gamma} L_\sigma^{-1} \tag{37}$$

and M by

$$M := E_\sigma(M_\sigma) = \frac{1}{n!} \sum_{\sigma \in \Gamma} M_\sigma, \tag{38}$$

where Γ is the set of all permutations of $\{1, 2, \ldots, n\}$. By direct computation, it is easily seen that

$$M := \begin{bmatrix} I - QS & QA^\top \\ -\beta A + \beta AQS & I - \beta AQA^\top \end{bmatrix}. \tag{39}$$

To prove the expected convergence of the randomly permuted ADMM (33) for problem (34) under Assumption 3.1, we will use a similar, but not the same manner as introduced in [46], which can be concluded as following:

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

(2) Let λ be any eigenvalue of M, $\mathrm{eig}(QS) \subset \left[0, \frac{4}{3}\right)$ 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.

In order to prove the above items, we need the following lemmas whose proof can be found in the Appendix.

Lemma 3.1. *Suppose that Assumption* 3.1 *holds,* $S \in \mathbb{R}^{d\times d}$ *is a symmetric matrix defined by* (35) *annd* Q *is defined by* (37)*. Then, the matrix* Q *is positive definite and all the eigenvalues of* QS *lie in* $\left[0, \frac{4}{3}\right)$*, i.e.,*

$$\mathrm{eig}(QS) \subset \left[0, \frac{4}{3}\right). \tag{40}$$

Lemma 3.2. *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*

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

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

Lemma 3.3. *Suppose* $S \in \mathbb{R}^{d\times d}$ *is a symmetric matrix defined by* (35) *and* $\beta > 0$*, then*

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

where $\mathrm{Rank}(\cdot)$ *denotes the rank of some matrix.*

Lemma 3.1 is an enhanced version of Lemma 2 in [46] to be compatible with problem (34). The proofs of Steps (2) and (3), which reveal the essential of this extension and hence are the key contribution here, will be presented in Subsections 3.2. The proof for Step (4) is given in Subsection 3.3.

3.2 The 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 3.4. *Suppose that Assumption* 3.1 *holds and* $S \in \mathbb{R}^{d\times d}$ *is a symmetric matrix defined by* (35)*. Let* λ *be any eigenvalue of* M*, then we have either* $|\lambda| < 1$ *or* $\lambda = 1$*.*

Proof. We introduce the following notation:

$$\gamma(u) = \frac{\beta u^* A^\top A u}{u^* S u} \quad \forall\, u \in \mathbb{C}^n \qquad \text{and} \qquad Su \neq 0, \tag{41}$$

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

$$0 \leq \gamma(u) \leq 1 \quad \forall\, u \in \mathbb{C}^n \qquad \text{and} \qquad Su \neq 0. \tag{42}$$

Similarly, we define

$$\kappa(u) = \frac{u^* Q^{-1} u}{u^* S u} \quad \forall\, u \in \mathbb{C}^n \qquad \text{and} \qquad Su \neq 0. \tag{43}$$

Note that $\operatorname{eig}(QS) < \frac{4}{3}$ by Lemma 3.1, we know that $\frac{4}{3}Q^{-1} - S \succeq 0$, and it therefore holds

$$0 < \kappa(u)^{-1} < \frac{4}{3} \quad \forall\, u \in \mathbb{C}^n \qquad \text{and} \qquad Su \neq 0. \tag{44}$$

We notice that matrix M can be factorized as

$$M = \begin{bmatrix} I & 0 \\ -\beta A & I \end{bmatrix} \begin{bmatrix} I - QS & QA^\top \\ 0 & I \end{bmatrix}. \tag{45}$$

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

$$M' := \begin{bmatrix} I - QS & QA^\top \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -\beta A & I \end{bmatrix} = \begin{bmatrix} I - QS - \beta Q A^\top A & QA^\top \\ -\beta A & I \end{bmatrix}. \tag{46}$$

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

Now, we assume that $\left(\lambda, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right)$ is an eigen-pair of M', namely,

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

which implies

$$(I - QS - \beta Q A^\top A) v_1 + Q A^\top v_2 = \lambda v_1; \tag{47}$$

$$-\beta A v_1 + v_2 = \lambda v_2. \tag{48}$$

The equality (48) gives that

$$(1 - \lambda) v_2 = \beta A v_1. \tag{49}$$

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 (47), we have

$$QSv_1 = (1-\lambda)v_1 + \frac{\lambda\beta}{1-\lambda} QA^\top Av_1.$$

By using the nonsingularity of Q, the above equality can be written as

$$Sv_1 = (1-\lambda)Q^{-1}v_1 + \frac{\lambda\beta}{1-\lambda} A^\top Av_1.$$

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

$$v_1^* Sv_1 = (1-\lambda)v_1^* Q^{-1}v_1 + \frac{\lambda\beta}{1-\lambda} v_1^* A^\top Av_1, \tag{50}$$

We claim that $v_1^* Sv_1 \neq 0$. Otherwise, it implies $v_1^* A^\top Av_1 = 0$ and therefore $\lambda = 1$ by the inequality $v_1^* Q^{-1}v_1 > 0$ and (50). This contradicts with our assumption $\lambda \neq 1$. Multiplying both sides of (50) by $(v_1^* Sv_1)^{-1}$ and substituting the definitions (41) and (43) into the above relation, we can 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

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

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

$$\lambda^2 + \left(\kappa(v_1)^{-1}(\gamma(v_1)+1) - 2\right)\lambda + \left(1 - \kappa(v_1)^{-1}\right) = 0. \tag{51}$$

The discriminant of the quadratic equation (51) is

$$\begin{aligned} \Delta &= \left(\kappa(v_1)^{-1}(\gamma(v_1)+1) - 2\right)^2 - 4\left(1 - \kappa(v_1)^{-1}\right) \\ &= \kappa(v_1)^{-1}\left(\kappa(v_1)^{-1}(\gamma(v_1)+1)^2 - 4\gamma(v_1)\right). \end{aligned} \tag{52}$$

We notice that

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

holds, which results from (42). Recalling (44), 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 (51) satisfies that

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

Case 2: $\frac{4\gamma(v_1)}{(\gamma(v_1)+1)^2} \leq \kappa(v_1)^{-1} < \frac{4}{3}$. This means the discriminant $\Delta \geq 0$, 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 (42) and (44), we know that

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

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

To sum up, it can be concluded that either $\lambda = 1$ or $|\lambda| < 1$ holds. $\square$

Now, we pay attention to the case where M has an eigenvalue 1 and show that it has a complete set of eigenvectors.

Lemma 3.5. *Suppose that Assumption* 3.1 *holds, and* $M \in \mathbb{R}^{(m+d)\times(m+d)}$ *is a matrix defined by* (39). *Suppose* 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

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

This, together with Lemma 3.2, shows that the algebraic multiplicity of 1 for M equals

$$\begin{aligned}
&m - d + 2d - \mathrm{Rank}(Q^{1/2}\beta A^\top AQ^{1/2}) - \mathrm{Rank}(Q^{1/2}(\beta A^\top A + H)Q^{1/2}) \\
&= m + d - \mathrm{Rank}(\beta A^\top A) - \mathrm{Rank}(\beta A^\top A + H),
\end{aligned} \tag{53}$$

where the equality follows from $Q \succ 0$ by Lemma 3.1. On the other hand, the geometric multiplicity of 1 for M is identical to the following quantity:

$$\begin{aligned}
&m + n - \mathrm{Rank}(I - M) \\
&= m + n - \mathrm{Rank}\begin{bmatrix} QS & -QA^\top \\ \beta A - \beta AQS & \beta AQA^\top \end{bmatrix} \\
&= m + n - \mathrm{Rank}\begin{bmatrix} QS & -QA^\top \\ \beta A & 0 \end{bmatrix} \\
&= m + n - \mathrm{Rank}\begin{bmatrix} S & -A^\top \\ \beta A & 0 \end{bmatrix},
\end{aligned} \tag{54}$$

where the second equality follows from the rank invariant property under elementary transformation, and the last equality is due to $Q \succ 0$ by Lemma 3.1. Combing (53), (54), Lemma 3.3 and the definition of S, we derive the desired conclusion. $\square$

3.3 Expected Convergence

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

Theorem 3.1. *Assume Assumption* 3.1 *holds. Suppose the RPADMM* (33) *is employed to solve the nonseparable quadratic programming* (34)*, then the expected iterate sequence converges to some KKT point to* (34)*.*

Proof. Let $(\bar{x}, \bar{\mu})$ be a KKT point of (34), i.e.,

$$\begin{bmatrix} H & -A^\top \\ \beta A & 0 \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{\mu} \end{bmatrix} = \begin{bmatrix} -g \\ \beta b \end{bmatrix}. \tag{55}$$

Denote (x^k, μ^k) by kth iterate of the algorithm. It follow from (36)and (55) that

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

By Lemma 3.4, we know $\rho(M) \le 1$. We proceed to this 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 3.1 is valid.

Case 2: $\rho(M) = 1$. By Lemmas 3.4 and 3.5, 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 non-singular matrix and $|\rho_i| < 1$ for all $i = 1, \dots, t$. 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 by manipulation shows that

$$\begin{bmatrix} H & -A^\top \\ \beta A & 0 \end{bmatrix} \begin{bmatrix} x^0 \\ \mu^0 \end{bmatrix} = 0. \tag{56}$$

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

$$\begin{bmatrix} H & -A^\top \\ \beta A & 0 \end{bmatrix} \begin{bmatrix} \bar{x} + x^0 \\ \bar{\mu} + \mu^0 \end{bmatrix} = \begin{bmatrix} -g \\ \beta b \end{bmatrix}. \tag{57}$$

This means that $(\bar{x} + x^0, \bar{\mu} + \mu^0)$ is a KKT point of (34).

The proof is completed. □

A byproduct of Theorem 3.1 is the expected convergence result of the EPOCHS (the "sampling without replacement" version of the randomized BCD method) for solving convex quadratic optimization. To the best of our knowledge, this is the first expected iterate convergence result of EPOCHS.

Corollary 3.1. *Assume $H \succeq 0$ and $H_{ii} \succ 0$ for $i = 1, 2, \ldots, n$. Suppose the EPOCHS is used to solve the following unconstrained quadratic programming:*

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top H x + g^\top x, \tag{58}$$

then the expected iterate sequence converges to an optimal solution of (58).

3.4 Convergence Rate Comparsison to Cyclic BCD

There seems a common perception that EPOCHS is dominating cyclic BCD (see [49] for example) in performance. In this subsection, we theoretically show that this finding may not be true in general. Consider the quadratic programming problem (58), where x is splitted 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 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

$$M_1 = \begin{bmatrix} H_{11} & 0 \\ H_{12}^\top & H_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & -H_{12} \\ 0 & 0 \end{bmatrix} \text{ and } M_2 = \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ -H_{12}^\top & 0 \end{bmatrix}. \tag{59}$$

The asymptotic convergence rates of these two iterative schemes are $\rho(M_1)$ and $\rho(M_2)$, respectively. The expected asymptotic convergence rate of EPOCHS in this case is $\rho((M_1 + M_2)/2)$. We have the following proposition which reveals the relationship among these rates.

Proposition 3.1. *Suppose $H \succeq 0, H_{11} \succ 0$ and $H_{22} \succ 0$. Let M_1 and M_2 be defined by* (59), *and $M_3 = (M_1 + M_2)/2$, then it holds that*

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

Proof. Without loss of generality, we only need to 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, we obtain by simple calculation that

$$M_1 = \begin{bmatrix} 0 & -H_{12} \\ 0 & H_{12}^\top H_{12} \end{bmatrix} \qquad \text{and} \qquad M_2 = \begin{bmatrix} H_{12}H_{12}^\top & 0 \\ -H_{12}^\top & 0 \end{bmatrix}. \tag{60}$$

Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{d_2}$ 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 get $\sigma_i \in [0, 1]$, $i = 1, \ldots, d_2$ and obtain from (60) 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 then the eigenvelues of M_3 are 0 (multiplicty $= d_1 - d_2$) and $\frac{\sigma_i \pm \sqrt{\sigma_i}}{2}$ for $i = 1, 2, \ldots, d_2$. Note that $\sigma_1 \in [0, 1]$, we further have

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

The proof is completed. □

Therefore, while random permutation is indeed to make multi-block ADMM and BCD more robust, especially against those "bad" or diverging problems; cyclic ADMM or BCD may still perform well or even better for solving "nice" problems.

4 Concluding Remarks

In this paper, we prove the point-wise or iterate convergence of the classical 2-block ADMM for solving convex optimization with coupled quadratic objective under a mild assumption. Such assumption becomes necessary and sufficient for the global convergence of the ADMM when the objective is a quadratic function. The result partially answers affirmatively the open question arising in [30] on the convergence of ADMM for nonseparable optimization problems. On the other hand, we show the expected convergence of RPADMM in solving linearly constrained coupled quadratic optimization problems. This is a non-trivial extension of the convergence analysis shown in [46], which is only for solving nonsingular linear systems. When the linear constraint is absent, the ADMM and RPADMM reduce to the cyclic BCD and EPOCHS. Our study thus provides new convergence results of the BCD type methods. In particular, we establish the first iterate convergence result for the 2-block cyclic BCD without assuming the boundedness of the iterates and the expected iterate convergence for EPOCHS of multi-block convex quadratic optimization. At the same time, we also theoretically demonstrate that EPOCHS does not necessarily dominate cyclic BCD in general so that it should be used in caution. Two challenging open questions are to extend our convergence results of RPADMM and EPOCHS to solve more general convex optimization problems, and to explore the global convergence rate of RPADMM and EPOCHS. In particular, which problems would be more suitable to apply RPADMM or EPOCHS to solve?

Acknowledgements. Caihua Chen was supported by the Natural Science Foundation of Jiangsu Province [Grant No. BK20130550] and the National Natural Science Foundation of China [Grant No. 11371192]. Min Li was supported by the National Natural Science Foundation of China [Grant No. 71390335, 11001053], Program for New Century Excellent Talents in University [Grant No. NCET-12-0111], and Qing Lan Project. Xin Liu was supported by the National Natural Science Foundation of China [Grant No. 11331012, 11471325, 11461161005], China 863 Program [Grant No. 2013AA122902], and the National Center for Mathematics and Interdisciplinary 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.

Appendix.

Appendix A. Proof of Lemma 3.1. The proof Lemma 3.1 is similar to the proof of Lemma 2 in [46]. The only difference is that we need to do mathematical induction for the results $Q \succ 0$ and $\mathrm{eig}(QS) \subset \left[0, \frac{4}{3}\right)$ together instead of only inducting on the eigenvalues of $\mathrm{eig}(QS)$ since S is only positive semidefinite. The induction can be achieved in the similar manner, though slightly more involved. For the brevity, we omit the details here. $\square$

Appendix B. Proof of Lemma 3.2. For convenience, we use the notation that

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

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 3.2 holds in this case. Suppose this Lemma is valid for $d = k-1$, we turn our attention to the case where $d = k$.

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

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

thus Lemma 3.2 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 $s_i > 0$. Let $W = P^\top T P \succeq 0$, then

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

Next, we proceed to our proof by consider the following two sub-cases.

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 sub-matrices of D and W deleting the first row and column. Since Lemma 3.2 holds for $d = k - 1$ by our assumption, there exists a polynomial $p(x)$ such that

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

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

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

which implies that Lemma 3.2 is true for $d = k$ in this sub-case.

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

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

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

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

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

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

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

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

and then Lemma 3.2 holds for this sub-case.

The proof is completed. $\square$

Appendix C. Proof of Lemma 3.3. It is easily seen that

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

and therefore we only need to prove that

$$\mathrm{Rank}\begin{bmatrix} S & -A^\top \\ \beta A & 0 \end{bmatrix} = \mathrm{Rank}\begin{bmatrix} S & 0 \\ 0 & \beta AA^\top \end{bmatrix}. \tag{61}$$

Indeed, consider the following linear system that

$$\begin{bmatrix} S & -A^\top \\ \beta A & 0 \end{bmatrix}\begin{bmatrix} x \\ \mu \end{bmatrix} = 0, \tag{62}$$

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 $A^\top \mu = 0$ since $S = H + \beta A^\top A$ is positive semidefinite. This means that

$$\begin{bmatrix} S & 0 \\ 0 & \beta AA^\top \end{bmatrix}\begin{bmatrix} x \\ \mu \end{bmatrix} = 0. \tag{63}$$

On the other hand, it is not difficult to verify that any solution of (63) is the solution of (62), or equivalently, the two linear systems (62) and (63) are equivalent. As a result, the rank equality (61) holds, which completes the proof of Lemma 3.3. $\square$

References

[1] Agarwal A, Negahban S, Wainwright MJ (1997) Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2): 1171–1197.

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

[3] Beck A, Tetruashvili L (2013) On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(2): 2037–2060.

[4] Beck A (2015) On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1): 185–209.

[5] Bertsekas DP (1999) Nonlinear Programming. *Athena-Scientific*, 2nd ed.

[6] Bertsekas DP, Tsitsiklis JN (1997) A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Parallel and Distributed Computation: Numerical Methods*, 2nd ed. Athena Scientific, Belmont, MA.

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

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

[9] Cai X, Han D, Yuan X (2014) The direct extension of ADMM for three-block separable convex minimization models is convergent when one function is strongly convex. Optimization Online.

[10] Chen C, He B, Yuan X, Ye Y (2014) The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, DOI: 10.1007/s10107-014-0826-5.

[11] Chen C, Shen Y, You Y (2013) 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 (2015) An efficient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming. arXiv:1506.00741v1.

[13] Chen L, Sun D, Toh K-C (2015) A note on the convergence of ADMM for linearly constrained convex optimization problems. arXiv:1507.02051v1.

[14] Cui Y, Li X, Sun D, Toh K-C (2015) On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions. arXiv:1502.00098v1.

[15] Davis D, Yin W (2014) Convergence rate analysis of several splitting schemes. *UCLA CAM Report*, 14-51.

[16] Deng W, Yin W (2015) On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, DOI: 10.1007/s10915-015-0048-x.

[17] Eckstein J, Bertsekas DP (1992) On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1): 293–318.

[18] Feng C, Xu H, Li BC (2014) An alternating direction method approach to traffic management. arXiv:1407.8309.

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

[20] Gao X, Zhang S (2015) First-order algorithms for convex optimization with nonseparable objective and coupled constraints. *Optimization online.*

[21] Glowinski R (1984) Numerical methods for nonlinear variational problems. *Springer-Verlag*, New York.

[22] Glowinski R, Marroco A (1975) 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.

[23] Han D, Yuan X (2012) A note on the alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 155(1): 227–238.

[24] Han D, Yuan X, Zhang W, Cai X (2013) An ADM-based splitting method for separable convex programming. *Computational Optimization and Applications*, 54: 343–369.

[25] He B, Tao M, Yuan X (2015) A splitting method for separable convex programming. *IMA Journal of Numerical Analysis*, 35: 394–426.

[26] He B, Tao M, Yuan X (2012) Alternating direction method with Gaussian back substitution for separable convex programming. *SIAM Journal on Optimization*, 22: 313–340.

[27] He B, Yuan X (2012) On the O($1/n$) convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2): 700–709.

[28] Hong M, Chang T, Wang X, Razaviyayn M, Ma S, Luo Z (2014) A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. arXiv:1401.7079.

[29] Hong M, Luo Z (2012) On the linear convergence of the alternating direction method of multipliers. arXiv:1208.3922.

[30] Hong M, Luo Z, Razaviyayn M (2014) Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. arXiv:1410.1390v1.

[31] Hong M, Wang X, Razaviyayn M, Luo Z (2014) Iterations complexity analysis for block coordinate descent method. arXiv:1310.6957v2.

[32] Li M, Sun D, Toh K-C (2015) A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. *Asia Pacific Journal of Operational Research*, 32: 1550024.

[33] Li X, Sun D, Toh K-C (2014) A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. *Mathematical Programming*, DOI: 10.1007/s10107-014-0850-5.

[34] Lin T, Ma S, Zhang S (2015) On the global linear convergence of the ADMM with multi-block variables. *SIAM Journal on Optimization*, 25: 1478–1497.

[35] Lin T, Ma S, Zhang S (2015) Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. arXiv:1504.03087.

[36] Lin T, Ma S, Zhang S (2015) On the sublinear convergence rate of multi-block ADMM. *To appear in Journal of the Operations Research Society of China.*

[37] Lu Z, Xiao L (2015) On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152: 615–642.

[38] Monteiro R, Svaiter B (2013) Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1): 475–507.

[39] Mota JFC, Xavier JMF, Aguiar PMF, Puschel M (2015) Distributed optimization with local domains: Applications in MPC and network flows. *IEEE Transactions on Automatic Control*, to appear.

[40] Peng YG, Ganesh A, Wright J, Xu WL, Ma Y (2012) Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34: 2233–2246.

[41] Razaviyayn M, Hong M, Luo Z (2013) A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2): 1126–1153.

[42] Richtárik P, Takáč M (2014) Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(2): 1–38.

[43] Shalev-Shwartz S, Zhang T (2013) Stochastic dual coordinate ascent methods for regularized loss. *Journal Of Machine Learning Research*, 14: 567–599.

[44] Shefi R, Teboulle M (2015) On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. *EURO Journal on Computational Optimization*, DOI: 10.1007/s13675-015-0048-5.

[45] Sun D, Toh K-C, Yang L (2015) A convergent 3-block semi-proximal alternating direction method of multipliers for conic programming with 4-block constraints. *SIAM Journal on Optimization*, 25(2): 882–915.

[46] Sun R, Luo Z, Ye Y (2015) On the expected convergence of randomly permuted ADMM. arXiv:1503.06387v1.

[47] Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109: 475–494.

[48] Tseng P, Yun S (2009) A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117: 387–423.

[49] Wright S (2015) Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34.

[50] Zhang Y (2010) Convergence of a class of stationary iterative methods for saddle point problems. *Rice University Technique Report*, TR10-24.