A NON-MONOTONE ALTERNATING UPDATING METHOD FOR A CLASS OF MATRIX FACTORIZATION PROBLEMS

LEI YANG*, TING KEI PONG*, AND XIAOJUN CHEN*

Abstract. In this paper we consider a general matrix factorization model which covers a large class of existing models with many applications in areas such as machine learning and imaging sciences. To solve this possibly nonconvex, nonsmooth and non-Lipschitz problem, we develop a non-monotone alternating updating method based on a potential function. Our method essentially updates two blocks of variables in turn by inexactly minimizing this potential function, and updates another auxiliary block of variables using an explicit formula. The special structure of our potential function allows us to take advantage of efficient computational strategies for non-negative matrix factorization to perform the alternating minimization over the two blocks of variables. A suitable line search criterion is also incorporated to improve the numerical performance. Under some mild conditions, we show that the line search criterion is well defined, and establish that the sequence generated is bounded and any cluster point of the sequence is a stationary point. Finally, we conduct some numerical experiments using real datasets to compare our method with some existing efficient methods for non-negative matrix factorization and matrix completion. The numerical results show that our method can outperform these methods for these specific applications.

Key words. Matrix factorization; non-monotone line search; stationary point; alternating updating.

AMS subject classifications. 90C26, 90C30, 90C90, 65K05

1. Introduction.

In this paper we consider a class of matrix factorization problems, which can be modeled as

$$\min_{X,Y}\ \mathcal{F}(X,Y) := \Psi(X) + \Phi(Y) + \frac{1}{2}\left\|\mathcal{A}(XY^\top) - \boldsymbol{b}\right\|^2, \tag{1.1}$$

where $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{n\times r}$ are decision variables with $r \leq \min\{m, n\}$, the functions $\Psi : \mathbb{R}^{m\times r} \to \mathbb{R}\cup\{\infty\}$ and $\Phi : \mathbb{R}^{n\times r} \to \mathbb{R}\cup\{\infty\}$ are proper closed but possibly *nonconvex*, *nonsmooth* and *non-Lipschitz*, $\boldsymbol{b} \in \mathbb{R}^q$ is a given vector and $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^q$ is a linear map with $q \leq mn$ and $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ ($\mathcal{I}_q$ denotes the identity map from $\mathbb{R}^q$ to $\mathbb{R}^q$). Model (1.1) covers many existing widely-studied models in many application areas such as machine learning [35] and imaging sciences [44]. In particular, $\Psi(X)$ and $\Phi(Y)$ can be various regularizers for inducing desired structures, and $\mathcal{A}$ can be suitably chosen to model different scenarios. For example, when $\Psi(X)$ and $\Phi(Y)$ are chosen as the indicator functions (see the next section for notation and definitions) for $\mathcal{X} = \{X \in \mathbb{R}^{m\times r} : X \geq 0\}$ and $\mathcal{Y} = \{Y \in \mathbb{R}^{n\times r} : Y \geq 0\}$, respectively, and $\mathcal{A}$ is the identity map, (1.1) reduces to the non-negative matrix factorization (NMF) problem, which has been widely used in data mining applications to provide interpretable decompositions of data. NMF was first introduced by Paatero and Tapper [25], and then popularized by Lee and Seung [17]. The basic task of NMF is to find two nonnegative matrices $X \in \mathbb{R}^{m\times r}_+$ and $Y \in \mathbb{R}^{n\times r}_+$ such that $M \approx XY^\top$ for a given nonnegative data matrix $M \in \mathbb{R}^{m\times n}_+$. We refer readers to [2, 9, 10, 18, 37] for more information on NMF and its variants. Another example of (1.1) arises in recent models of the matrix completion (MC) problem (see [30, 31, 32]), where $\Psi(X)$

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, P.R. China. (lei.yang@connect.polyu.hk, tk.pong@polyu.edu.hk, xiaojun.chen@polyu.edu.hk). The second author's work was supported in part by Hong Kong Research Grants Council PolyU153085/16p. The third author's work was supported in part by Hong Kong Research Grants Council PolyU153000/15p.

This manuscript is for review purposes only.

and $\Phi(Y)$ are chosen as the Schatten-p_1 quasi-norm and the Schatten-p_2 quasi-norm for suitable $p_1, p_2 > 0$, respectively, and $\mathcal{A}$ is the sampling map. The MC problem aims to recover an unknown low rank matrix from a sample of its entries and arises in various applications (see, for example, [3, 22, 27, 33]). Many widely-studied models for MC are based on nuclear-norm minimization [5, 6, 26], or, more generally, Schatten-p $(0 < p \leq 1)$ (quasi$-$)norm minimization [16, 23, 42]. Recently, models based on low-rank matrix factorization such as (1.1) have become popular because singular value decompositions or eigenvalue decompositions of huge $(m \times n)$ matrices are not required for solving these models (see, for example, [15, 30, 31, 32, 34, 38]). More examples of (1.1) can be found in recent surveys [35, 44].

Problem (1.1) is in general nonconvex (even when Ψ, Φ are convex) and NP-hard[1]. Therefore, in this paper, we focus on finding a *stationary point* of the objective $\mathcal{F}$ in (1.1). Note that $\mathcal{F}$ involves two blocks of variables. This kind of structure has been widely studied in the literature; see, for example, [1, 4, 13, 14, 40, 41, 43]. One popular class of methods for tackling this kind of problems is the alternating direction method of multipliers (ADMM) (see, for example, [41, 43]), in which each iteration consists of an alternating minimization of an augmented Lagrangian function that involves X, Y and some auxiliary variables, followed by updates of the associated multipliers. However, the conditions presented in [41, 43] that guarantee convergence of the ADMM are too restrictive. Moreover, updating the auxiliary variables and the multipliers can be expensive for large-scale problems. Another class of methods for (1.1) is the alternating-minimization-based (or block-coordinate-descent-type) methods (see [1, 4, 8, 11, 20, 21, 40]), which alternately (exactly or inexactly) minimizes $\mathcal{F}(X, Y)$ over each block of variables and converges under some mild conditions. When $\mathcal{A}$ is not the identity map, the majorization technique can be used to simplify the subproblems. Some representative algorithms of this class are proximal alternating linearized minimization (PALM) [4], hierarchical alternating least squares (HALS) (for NMF only; see [8, 11, 20, 21]) and block coordinate descent (BCD) [40]. Comparing with ADMM, it was reported in [40] that BCD outperforms ADMM in both CPU time and solution quality for NMF.

PALM, HALS and BCD are currently the state-of-the-art algorithms for solving problems of the form (1.1). In this paper, we develop a new iterative method for (1.1), which, according to our numerical experiments in Section 6, outperforms HALS and BCD for NMF, and PALM for MC. Our method is based on the following potential function (specifically constructed for $\mathcal{F}$ in (1.1)):

$$\Theta_{\alpha,\beta}(X, Y, Z) := \Psi(X) + \Phi(Y) + \frac{\alpha}{2}\|XY^\top - Z\|_F^2 + \frac{\beta}{2}\|\mathcal{A}(Z) - \boldsymbol{b}\|^2, \tag{1.2}$$

where α and β are real numbers. Instead of alternately (exactly or inexactly) minimizing $\mathcal{F}(X, Y)$ or the augmented Lagrangian function, our method alternately updates X and Y by inexactly minimizing $\Theta_{\alpha,\beta}(X, Y, Z)$ over X and Y, and then updates Z by an *explicit formula*. Note that the coupled variables $XY^\top$ is now separated from $\mathcal{A}$ in our potential function. Thus, one can readily take advantage of efficient computational strategies for NMF, such as those used in HALS (see the "hierarchical-prox" updating strategy in Section 4), for inexactly minimizing $\Theta_{\alpha,\beta}(X, Y, Z)$ over X or Y. Furthermore, our method can be implemented for NMF and MC without explicitly forming the huge $(m \times n)$ matrix Z (see (6.3) and (6.5)) in each iteration. This signifi-

[1] Problem (1.1) is NP-hard because it contains NMF as a special case, which is NP-hard in general [36].

cantly reduces the computational cost per iteration. Finally, a suitable non-monotone line search criterion, which is motivated by recent studies on non-monotone algorithms (see, for example, [7, 12, 39]), is also incorporated to improve the numerical performance.

In the rest of this paper, we first present notation and preliminaries in Section 2. We then study the properties of our potential function $\Theta_{\alpha,\beta}$ in Section 3. Specifically, if $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and α, β are chosen such that $\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succ 0$ and $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, then the problem $\min\limits_{X,Y,Z} \{\Theta_{\alpha,\beta}(X, Y, Z)\}$ is equivalent to (1.1) (see Theorem 3.2). Furthermore, under the weaker conditions that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, we can show that (i) a stationary point of $\Theta_{\alpha,\beta}$ gives a stationary point of $\mathcal{F}$; (ii) a stationary point of $\mathcal{F}$ can be used to construct a stationary point of $\Theta_{\alpha,\beta}$ (see Theorem 3.3). Thus, one can find a stationary point of $\mathcal{F}$ by finding a stationary point of $\Theta_{\alpha,\beta}$. In Section 4, we develop a non-monotone alternating updating method to find a stationary point of $\Theta_{\alpha,\beta}$, and hence of $\mathcal{F}$. The convergence analysis of our method is presented in Section 5. We show that our non-monotone line search criterion is well defined and any cluster point of the sequence generated by our method is a stationary point of $\mathcal{F}$ under some mild conditions. Section 6 gives numerical experiments to evaluate the performance of our method for NMF and MC on real datasets. Our computational results illustrate the efficiency of our method. Finally, some concluding remarks are given in Section 7.

2. Notation and preliminaries. In this paper, for a vector $\boldsymbol{x} \in \mathbb{R}^m$, x_i denotes its i-th entry, $\|\boldsymbol{x}\|$ denotes the Euclidean norm of $\boldsymbol{x}$ and $\mathrm{Diag}(\boldsymbol{x})$ denotes the diagonal matrix whose i-th diagonal element is x_i. For a matrix $X \in \mathbb{R}^{m\times n}$, x_{ij} denotes the ij-th entry of X, $\boldsymbol{x}_j$ denotes the j-th column of X and $\mathrm{tr}(X)$ denotes the trace of X. The Schatten-p (quasi-)norm ($0 < p < \infty$) of X is defined as $\|X\|_{S_p} = \left(\sum_{i=1}^{\min(m,n)} \varsigma_i^p(X)\right)^{\frac{1}{p}}$, where $\varsigma_i(X)$ is the i-th singular value of X. For $p = 2$, the Schatten-2 norm reduces to the Frobenius norm $\|X\|_F$, and for $p = 1$, the Schatten-1 norm reduces to the nuclear norm $\|X\|_*$. Moreover, the spectral norm is denoted by $\|X\|$, which is the largest singular value of X; and the ℓ_1-norm and ℓ_p-quasi-norm ($0 < p < 1$) of X are given by $\|X\|_1 := \sum_{i=1}^m \sum_{j=1}^n |x_{ij}|$ and $\|X\|_p := \left(\sum_{i=1}^m \sum_{j=1}^n |x_{ij}|^p\right)^{\frac{1}{p}}$, respectively. For two matrices X and Y of the same size, we denote their trace inner product by $\langle X, Y\rangle := \sum_{i=1}^m \sum_{j=1}^n x_{ij}y_{ij}$. We also use $X \leq Y$ (resp., $X \geq Y$) to denote $x_{ij} \leq y_{ij}$ (resp., $x_{ij} \geq y_{ij}$) for all (i, j). Furthermore, for a linear map $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^q$, $\mathcal{A}^*$ denotes the adjoint linear map and $\|\mathcal{A}\|$ denotes the induced operator norm of $\mathcal{A}$, i.e., $\|\mathcal{A}\| = \sup\{\|\mathcal{A}(X)\| : \|X\|_F \leq 1\}$. A linear self-map $\mathcal{T}$ is said to be symmetric if $\mathcal{T} = \mathcal{T}^*$. For a symmetric linear self-map $\mathcal{T} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$, we say that $\mathcal{T}$ is positive definite, denoted by $\mathcal{T} \succ 0$, if $\langle X, \mathcal{T}(X)\rangle > 0$ for all $X \neq 0$. The identity map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m\times n}$ is denoted by $\mathcal{I}$ and the identity map from $\mathbb{R}^q$ to $\mathbb{R}^q$ is denoted by $\mathcal{I}_q$. Finally, for a nonempty closed set $\mathcal{C} \subseteq \mathbb{R}^{m\times n}$, its indicator function $\delta_{\mathcal{C}}$ is defined by

$$\delta_{\mathcal{C}}(X) = \begin{cases} 0 & \text{if } X \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

For an extended-real-valued function $f : \mathbb{R}^{m\times n} \to [-\infty, \infty]$, we say that it is *proper* if $f(X) > -\infty$ for all $X \in \mathbb{R}^{m\times n}$ and its domain $\mathrm{dom} f := \{X \in \mathbb{R}^{m\times n} : f(X) < \infty\}$ is nonempty. A function $f : \mathbb{R}^{m\times n} \to [-\infty, \infty]$ is level-bounded [28, Definition 1.8] if for every $\alpha \in \mathbb{R}$, the set $\{X \in \mathbb{R}^{m\times n} : f(X) \leq \alpha\}$ is bounded (possibly empty). For a proper function $f : \mathbb{R}^{m\times n} \to (-\infty, \infty]$, we use the notation

$Y \xrightarrow{f} X$ to denote $Y \to X$ (i.e., $\|Y - X\|_F \to 0$) and $f(Y) \to f(X)$. The (limiting) subdifferential [28, Definition 8.3] of f at $X \in \mathrm{dom} f$ used in this paper, denoted by $\partial f(X)$, is defined as

$$\partial f(X) := \left\{ D \in \mathbb{R}^{m \times n} : \exists\, X^k \xrightarrow{f} X \text{ and } D^k \to D \text{ with } D^k \in \widehat{\partial} f(X^k) \text{ for all } k \right\},$$

where $\widehat{\partial} f(\widetilde{Y})$ denotes the Fréchet subdifferential of f at $\widetilde{Y} \in \mathrm{dom} f$, which is the set of all $D \in \mathbb{R}^{m \times n}$ satisfying

$$\liminf_{Y \neq \widetilde{Y}, Y \to \widetilde{Y}} \frac{f(Y) - f(\widetilde{Y}) - \langle D,\, Y - \widetilde{Y} \rangle}{\|Y - \widetilde{Y}\|_F} \geq 0.$$

From the above definition, we can easily observe (see, for example, [28, Proposition 8.7]) that

$$\left\{ D \in \mathbb{R}^{m \times n} : \exists\, X^k \xrightarrow{f} X,\ D^k \to D,\ D^k \in \partial f(X^k) \right\} \subseteq \partial f(X). \tag{2.1}$$

When f is continuously differentiable or convex, the above subdifferential coincides with the classical concept of derivative or convex subdifferential of f; see, for example, [28, Exercise 8.8] and [28, Proposition 8.12]. In this paper, we say that X^* is *stationary point* of f if $0 \in \partial f(X^*)$.

For a proper closed function $g : \mathbb{R}^m \to (-\infty, \infty]$, the proximal mapping $\mathrm{Prox}_g : \mathbb{R}^m \to \mathbb{R}^m$ of g is defined by $\mathrm{Prox}_g(\boldsymbol{z}) := \mathop{\mathrm{Argmin}}\limits_{\boldsymbol{x} \in \mathbb{R}^m} \left\{ g(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 \right\}$. For any $\nu > 0$, the matrix shrinkage operator $\mathcal{S}_\nu : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ is defined by

$$\mathcal{S}_\nu(X) := U \mathrm{Diag}(\bar{\boldsymbol{s}}) V^\top \text{ with } \bar{s}_i = \begin{cases} s_i - \nu, & \text{if } s_i - \nu > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $U \in \mathbb{R}^{m \times t}$, $\boldsymbol{s} \in \mathbb{R}^t_+$ and $V \in \mathbb{R}^{n \times t}$ are given by the singular value decomposition of X, i.e, $X = U \mathrm{Diag}(\boldsymbol{s}) V^\top$.

We now present two propositions, which will be useful for developing our method in Section 4.

PROPOSITION 2.1. *Suppose that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and $\alpha(\alpha + \beta) \neq 0$. Then, $\alpha \mathcal{I} + \beta \mathcal{A}^* \mathcal{A}$ is invertible and its inverse is given by $\frac{1}{\alpha}\mathcal{I} - \frac{\beta}{\alpha(\alpha+\beta)}\mathcal{A}^*\mathcal{A}$.*

Proof. It is easy to check that $\frac{1}{\alpha}\mathcal{I} - \frac{\beta}{\alpha(\alpha+\beta)}\mathcal{A}^*\mathcal{A}$ is well defined since $\alpha(\alpha+\beta) \neq 0$, and that $\left(\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A}\right)\left(\frac{1}{\alpha}\mathcal{I} - \frac{\beta}{\alpha(\alpha+\beta)}\mathcal{A}^*\mathcal{A}\right) = \mathcal{I}$. This completes the proof. ∎

PROPOSITION 2.2. *Let $\psi : \mathbb{R}^m \to (-\infty, \infty]$ and $\phi : \mathbb{R}^n \to (-\infty, \infty]$ be proper closed functions. Given $P, Q \in \mathbb{R}^{m \times n}$ and $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^m$ with $\|\boldsymbol{a}\| \neq 0$, $\|\boldsymbol{b}\| \neq 0$, the following statements hold.*

(i) *The problem $\min\limits_{\boldsymbol{x} \in \mathbb{R}^m} \left\{ \psi(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}\boldsymbol{a}^\top - P\|_F^2 \right\}$ is equivalent to*

$$\min_{\boldsymbol{x} \in \mathbb{R}^m} \left\{ \psi(\boldsymbol{x}) + \frac{\|\boldsymbol{a}\|^2}{2} \left\| \boldsymbol{x} - \frac{P\boldsymbol{a}}{\|\boldsymbol{a}\|^2} \right\|^2 \right\};$$

(ii) *The problem $\min\limits_{\boldsymbol{y} \in \mathbb{R}^n} \left\{ \phi(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{b}\boldsymbol{y}^\top - Q\|_F^2 \right\}$ is equivalent to*

$$\min_{\boldsymbol{y} \in \mathbb{R}^n} \left\{ \phi(\boldsymbol{y}) + \frac{\|\boldsymbol{b}\|^2}{2} \left\| \boldsymbol{y} - \frac{Q^\top\boldsymbol{b}}{\|\boldsymbol{b}\|^2} \right\|^2 \right\}.$$

Proof. Statement (i) can be easily proved by noticing that

$$\begin{aligned}\|\boldsymbol{x}\boldsymbol{a}^\top - P\|_F^2 &= \|\boldsymbol{x}\boldsymbol{a}^\top\|_F^2 - 2\langle \boldsymbol{x}\boldsymbol{a}^\top, P\rangle + \|P\|_F^2 = \|\boldsymbol{a}\|^2\|\boldsymbol{x}\|^2 - 2\langle \boldsymbol{x}, P\boldsymbol{a}\rangle + \|P\|_F^2 \\ &= \|\boldsymbol{a}\|^2 \left\|\boldsymbol{x} - P\boldsymbol{a}/\|\boldsymbol{a}\|^2\right\|^2 - \|P\boldsymbol{a}\|^2/\|\boldsymbol{a}\|^2 + \|P\|_F^2.\end{aligned}$$

Then, statement (ii) can be easily proved by using statement (i) and $\|\boldsymbol{b}\boldsymbol{y}^\top - Q\|_F^2 = \|\boldsymbol{y}\boldsymbol{b}^\top - Q^\top\|_F^2$. ∎

Before ending this section, we discuss the first-order necessary conditions for (1.1). First, from [28, Exercise 8.8] and [28, Proposition 10.5], we see that

$$\partial\mathcal{F}(X,\,Y) = \begin{pmatrix} \partial\Psi(X) + \mathcal{A}^*\left(\mathcal{A}(XY^\top) - \boldsymbol{b}\right)Y \\ \partial\Phi(Y) + \left(\mathcal{A}^*\left(\mathcal{A}(XY^\top) - \boldsymbol{b}\right)\right)^\top X \end{pmatrix}.$$

Then, it follows from the generalized Fermat's rule [28, Theorem 10.1] that any local minimizer $(\overline{X}, \overline{Y})$ of (1.1) satisfies $0 \in \partial\mathcal{F}(\overline{X}, \overline{Y})$, i.e.,

$$\begin{cases} 0 \in \partial\Psi(\overline{X}) + \mathcal{A}^*(\mathcal{A}(\overline{X}\,\overline{Y}^\top) - \boldsymbol{b})\overline{Y}, \\ 0 \in \partial\Phi(\overline{Y}) + (\mathcal{A}^*(\mathcal{A}(\overline{X}\,\overline{Y}^\top) - \boldsymbol{b}))^\top\overline{X}, \end{cases} \tag{2.2}$$

which implies that $(\overline{X}, \overline{Y})$ is a stationary point of $\mathcal{F}$. In this paper, we focus on finding a stationary point (X^*, Y^*) of $\mathcal{F}$, i.e., (X^*, Y^*) satisfies (2.2) in place of $(\overline{X}, \overline{Y})$.

3. The potential function for $\mathcal{F}$.

In this section, we analyze the relation between $\mathcal{F}$ and its potential function $\Theta_{\alpha,\beta}$ defined in (1.2). Intuitively, $\Theta_{\alpha,\beta}$ originates from $\mathcal{F}$ by separating the coupled variables $XY^\top$ from the linear mapping $\mathcal{A}$ via introducing an auxiliary variable Z and penalizing $XY^\top = Z$. We will see later that the stationary point of $\mathcal{F}$ can be characterized by the stationary point of $\Theta_{\alpha,\beta}$. Before proceeding, we prove the following technical lemma.

LEMMA 3.1. *Suppose that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Then, for any (X, Y, Z) satisfying*

$$Z = \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)\left(XY^\top\right) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b}), \tag{3.1}$$

we have $\mathcal{F}(X, Y) = \Theta_{\alpha,\beta}(X, Y, Z)$.

Proof. First, from (3.1), we have

$$XY^\top - Z = \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\mathcal{A}(XY^\top) - \boldsymbol{b}) \tag{3.2}$$

$$\begin{aligned}\mathcal{A}(Z) - \boldsymbol{b} &= \mathcal{A}\left(XY^\top - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}(XY^\top) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b})\right) - \boldsymbol{b} \\ &= \mathcal{A}(XY^\top) - \tfrac{\beta}{\alpha+\beta}\mathcal{A}\mathcal{A}^*\mathcal{A}(XY^\top) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}\mathcal{A}^*(\boldsymbol{b}) - \boldsymbol{b} = \tfrac{\alpha}{\alpha+\beta}\left(\mathcal{A}(XY^\top) - \boldsymbol{b}\right),\end{aligned} \tag{3.3}$$

where the last equality follows from $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$. Then, we see that

$$\begin{aligned}&\tfrac{\alpha}{2}\|XY^\top - Z\|_F^2 + \tfrac{\beta}{2}\left\|\mathcal{A}(Z) - \boldsymbol{b}\right\|^2 \\ &= \tfrac{\alpha}{2}\left\|\tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\mathcal{A}(XY^\top) - \boldsymbol{b})\right\|_F^2 + \tfrac{\beta}{2}\left\|\tfrac{\alpha}{\alpha+\beta}\left(\mathcal{A}(XY^\top) - \boldsymbol{b}\right)\right\|^2 \\ &= \tfrac{\alpha\beta^2}{(\alpha+\beta)^2}\cdot\tfrac{1}{2}\left\|\mathcal{A}^*(\mathcal{A}(XY^\top) - \boldsymbol{b})\right\|_F^2 + \tfrac{\alpha^2\beta}{(\alpha+\beta)^2}\cdot\tfrac{1}{2}\left\|\mathcal{A}(XY^\top) - \boldsymbol{b}\right\|^2 \\ &= \tfrac{\alpha\beta^2}{(\alpha+\beta)^2}\cdot\tfrac{1}{2}\left\|\mathcal{A}(XY^\top) - \boldsymbol{b}\right\|^2 + \tfrac{\alpha^2\beta}{(\alpha+\beta)^2}\cdot\tfrac{1}{2}\left\|\mathcal{A}(XY^\top) - \boldsymbol{b}\right\|^2 \\ &= \tfrac{\alpha\beta}{\alpha+\beta}\cdot\tfrac{1}{2}\left\|\mathcal{A}(XY^\top) - \boldsymbol{b}\right\|^2,\end{aligned}$$

where the first equality follows from (3.2) and (3.3); and the third equality follows from $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$. This, together with $\frac{1}{\alpha} + \frac{1}{\beta} = 1$ and the definitions of $\mathcal{F}$ and $\Theta_{\alpha,\beta}$ completes the proof. ◻

Based on the above lemma, we now establish the following property of $\Theta_{\alpha,\beta}$.

THEOREM 3.2. *Suppose that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$. If α and β are chosen such that $\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succ 0$ and $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, then the problem $\min\limits_{X,Y,Z} \{\Theta_{\alpha,\beta}(X,Y,Z)\}$ is equivalent to (1.1).*

Proof. First, it is easy to see from $\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succ 0$ that the function $Z \longmapsto \Theta_{\alpha,\beta}(X,Y,Z)$ is strongly convex. Thus, for any fixed X and Y, the optimal solution Z^* to the problem $\min\limits_{Z} \{\Theta_{\alpha,\beta}(X,Y,Z)\}$ exists and is unique, and can be obtained explicitly. Indeed, from the optimality condition, we have

$$\alpha(Z^* - XY^\top) + \beta\mathcal{A}^*(\mathcal{A}(Z^*) - \boldsymbol{b}) = 0.$$

Then, since $\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A}$ is invertible (as $\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succ 0$), we see that

$$\begin{aligned}
Z^* &= (\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A})^{-1}\left[\alpha XY^\top + \beta\mathcal{A}^*(\boldsymbol{b})\right] \\
&= \left[\tfrac{1}{\alpha}\mathcal{I} - \tfrac{\beta}{\alpha(\alpha+\beta)}\mathcal{A}^*\mathcal{A}\right]\left[\alpha XY^\top + \beta\mathcal{A}^*(\boldsymbol{b})\right] \\
&= \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)(XY^\top) + \left[\tfrac{\beta}{\alpha}\mathcal{A}^*(\boldsymbol{b}) - \tfrac{\beta^2}{\alpha(\alpha+\beta)}\mathcal{A}^*\mathcal{A}\mathcal{A}^*(\boldsymbol{b})\right] \\
&= \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)(XY^\top) + \left[\tfrac{\beta}{\alpha} - \tfrac{\beta^2}{\alpha(\alpha+\beta)}\right]\mathcal{A}^*(\boldsymbol{b}) \\
&= \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)(XY^\top) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b}),
\end{aligned}$$

where the second equality follows from Proposition 2.1 and the fourth equality follows from $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$. This, together with Lemma 3.1, implies that $\mathcal{F}(X,Y) = \Theta_{\alpha,\beta}(X,Y,Z^*)$. Then, we have that

$$\begin{aligned}
\min_{X,Y,Z} \{\Theta_{\alpha,\beta}(X,Y,Z)\} &= \min_{X,Y}\left\{\min_{Z}\{\Theta_{\alpha,\beta}(X,Y,Z)\}\right\} = \min_{X,Y}\{\Theta_{\alpha,\beta}(X,Y,Z^*)\} \\
&= \min_{X,Y}\{\mathcal{F}(X,Y)\}.
\end{aligned}$$

This completes the proof. ◻

REMARK 3.1. *From the proof of Lemma 3.1, we see that if Φ and Ψ are the indicator functions of some nonempty closed sets, then $\mathcal{F}(X,Y) = \left(\frac{1}{\alpha} + \frac{1}{\beta}\right)\Theta_{\alpha,\beta}(X,Y,Z)$ holds with the special choice of Z in (3.1) whenever $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and $\frac{1}{\alpha} + \frac{1}{\beta} > 0$. Thus, the result in Theorem 3.2 remains valid whenever $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and α, β are chosen such that $\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succ 0$ and $\frac{1}{\alpha} + \frac{1}{\beta} > 0$.*

We see from Theorem 3.2 that (1.1) is equivalent to minimizing $\Theta_{\alpha,\beta}$ with some suitable choices of α and β. On the other hand, we can also characterize the relation between the stationary points of $\mathcal{F}$ and $\Theta_{\alpha,\beta}$ under weaker conditions on α and β.

THEOREM 3.3. *Suppose that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and α, β are chosen such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Then, the following statements hold.*

(i) *If (X^*, Y^*, Z^*) is a stationary point of $\Theta_{\alpha,\beta}$, then (X^*, Y^*) is a stationary point of $\mathcal{F}$;*

(ii) *If (X^*, Y^*) is a stationary point of $\mathcal{F}$, then (X^*, Y^*, Z^*) is a stationary point of $\Theta_{\alpha,\beta}$, where Z^* is given by*

$$Z^* = \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)\left(X^*(Y^*)^\top\right) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b}). \tag{3.4}$$

Proof. First, if (X^*,Y^*,Z^*) is a stationary point of $\Theta_{\alpha,\beta}$, then we have $0 \in \partial\Theta_{\alpha,\beta}(X^*,Y^*,Z^*)$, i.e.,

$$\begin{cases} 0 \in \partial\Psi(X^*) + \alpha(X^*(Y^*)^\top - Z^*)Y^*, & \text{(3.5a)}\\ 0 \in \partial\Phi(Y^*) + \alpha(X^*(Y^*)^\top - Z^*)^\top X^*, & \text{(3.5b)}\\ 0 = \alpha(Z^* - X^*(Y^*)^\top) + \beta\mathcal{A}^*(\mathcal{A}(Z^*) - \boldsymbol{b}). & \text{(3.5c)} \end{cases}$$

Since $\frac{1}{\alpha}+\frac{1}{\beta} = 1$, we have $\alpha(\alpha+\beta) \neq 0$ and hence $\alpha\mathcal{I}+\beta\mathcal{A}^*\mathcal{A}$ is invertible from Lemma 2.1. Then, using the same arguments in the proof of Theorem 3.2, we see from (3.5c) that (X^*,Y^*,Z^*) satisfies (3.4). Moreover, using (3.4) and the same arguments in (3.2) and (3.3), we have

$$X^*(Y^*)^\top - Z^* = \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\mathcal{A}(X^*(Y^*)^\top) - \boldsymbol{b}), \tag{3.6}$$

$$\mathcal{A}(Z^*) - \boldsymbol{b} = \tfrac{\alpha}{\alpha+\beta}\left(\mathcal{A}(X^*(Y^*)^\top) - \boldsymbol{b}\right). \tag{3.7}$$

Thus, substituting (3.6) into (3.5a) and (3.5b), we see that

$$\begin{cases} 0 \in \partial\Psi(X^*) + \frac{\alpha\beta}{\alpha+\beta}\mathcal{A}^*(\mathcal{A}(X^*(Y^*)^\top) - \boldsymbol{b})Y^*, \\ 0 \in \partial\Phi(Y^*) + \frac{\alpha\beta}{\alpha+\beta}\left(\mathcal{A}^*(\mathcal{A}(X^*(Y^*)^\top) - \boldsymbol{b})\right)^\top X^*. \end{cases} \tag{3.8}$$

This together with $\frac{1}{\alpha}+\frac{1}{\beta} = 1$ implies (X^*,Y^*) is a stationary point of $\mathcal{F}$. This proves statement (i).

We now prove statement (ii). First, if (X^*,Y^*) is a stationary point of $\mathcal{F}$, then invoking $\frac{1}{\alpha}+\frac{1}{\beta} = 1$ and (2.2), we have (3.8). Next, we consider (X^*,Y^*,Z^*) with Z^* given by (3.4). Then, (X^*,Y^*,Z^*) satisfies (3.6) and (3.7). Thus, substituting (3.6) into (3.8), we obtain (3.5a) and (3.5b). Moreover, we have from (3.6) and (3.7) that

$$\begin{aligned} &\alpha(Z^* - X^*(Y^*)^\top) + \beta\mathcal{A}^*(\mathcal{A}(Z^*) - \boldsymbol{b}) \\ &= -\tfrac{\alpha\beta}{\alpha+\beta}\mathcal{A}^*\left((\mathcal{A}(X^*(Y^*)^\top) - \boldsymbol{b}\right) + \beta\mathcal{A}^*\left(\tfrac{\alpha}{\alpha+\beta}\left(\mathcal{A}(X^*(Y^*)^\top) - \boldsymbol{b}\right)\right) = 0. \end{aligned} \tag{3.9}$$

This together with (3.5a) and (3.5b) implies that (X^*,Y^*,Z^*) is a stationary point of $\Theta_{\alpha,\beta}$. This proves statement (ii). ∎

REMARK 3.2. *From the proof of Theorem 3.3, one can see that if $\partial\Psi$ and $\partial\Phi$ are cones, Theorem 3.3 remains valid under the weaker conditions that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and $\frac{1}{\alpha}+\frac{1}{\beta} > 0$.*

From Theorem 3.3, we see that a stationary point of $\mathcal{F}$ can be obtained from a stationary point of $\Theta_{\alpha,\beta}$ with a suitable choice of α and β, i.e., $\frac{1}{\alpha}+\frac{1}{\beta} = 1$. Since the linear map $\mathcal{A}$ is no longer associated with the coupled variables $XY^\top$ in $\Theta_{\alpha,\beta}$, finding a stationary point of $\Theta_{\alpha,\beta}$ is conceivably easier. Thus, one can consider finding a stationary point of $\Theta_{\alpha,\beta}$ in order to find a stationary point of $\mathcal{F}$. Note that some existing alternating-minimization-based methods (see, for example, [1, 40]) can be used to find a stationary point of $\Theta_{\alpha,\beta}$, and hence of $\mathcal{F}$, under the conditions that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and α, β are chosen so that $\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succ 0$ and $\frac{1}{\alpha}+\frac{1}{\beta} = 1$. These conditions further imply that $\alpha > 1$ and $\beta = \frac{\alpha}{\alpha-1} > 1$. However, as we will see from our numerical results in Section 6, finding a stationary point of $\Theta_{\alpha,\beta}$ with $\alpha > 1$ can be slow. In view of this, in the next section, we develop a new non-monotone alternating updating method for finding a stationary of $\Theta_{\alpha,\beta}$ (and hence of $\mathcal{F}$) under the weaker conditions that $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and $\frac{1}{\alpha}+\frac{1}{\beta} = 1$. This allows more flexibilities in choosing α and β.

4. Non-monotone alternating updating method. In this section, we consider a non-monotone alternating updating method (NAUM) for finding a stationary point of $\Theta_{\alpha,\beta}$ with $\frac{1}{\alpha}+\frac{1}{\beta}=1$. Compared to existing alternating-minimization-based methods [1, 40] applied to $\Theta_{\alpha,\beta}$, which update X, Y, Z by alternately solving subproblems related to $\Theta_{\alpha,\beta}$, NAUM updates Z by an *explicit formula* (see (4.5)) and updates X, Y by solving subproblems related to $\Theta_{\alpha,\beta}$ in a Gauss-Seidel manner. Before presenting the complete algorithm, we first comment on the updates of X and Y.

Let (X^k, Y^k) denote the value of (X,Y) after the $(k-1)^{\text{th}}$ iteration, and let (U,V) denote the candidate for (X^{k+1}, Y^{k+1}) at the k-th iteration (we will set (X^{k+1}, Y^{k+1}) to be (U,V) if a line search criterion is satisfied; more details can be found in Algorithm 1). For notational simplicity, we also define

$$\mathcal{H}_\alpha(X,Y,Z) := \frac{\alpha}{2}\|XY^\top - Z\|_F^2$$

for any (X,Y,Z). Then, at the k-th iteration, we first compute Z^k by (4.5) and, in the line search loop, we compute U in one of the following 3 ways for a given $\mu_k > 0$:

- **Proximal**

$$U \in \operatorname*{Argmin}_{X}\ \Psi(X) + \mathcal{H}_\alpha(X, Y^k, Z^k) + \frac{\mu_k}{2}\|X - X^k\|_F^2. \tag{4.1a}$$

- **Prox-linear**

$$U \in \operatorname*{Argmin}_{X}\ \Psi(X) + \langle \nabla_X \mathcal{H}_\alpha(X^k, Y^k, Z^k),\, X - X^k\rangle + \frac{\mu_k}{2}\|X - X^k\|_F^2. \tag{4.1b}$$

- **Hierarchical-prox** If Ψ is column-wise separable, i.e., $\Psi(X) = \sum_{i=1}^r \psi_i(\boldsymbol{x}_i)$ for $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r] \in \mathbb{R}^{m\times r}$, we can update U column-by-column. Specifically, for $i = 1, 2, \cdots, r$, compute

$$\boldsymbol{u}_i \in \operatorname*{Argmin}_{\boldsymbol{x}_i}\ \psi_i(\boldsymbol{x}_i) + \mathcal{H}_\alpha(\boldsymbol{u}_{j<i}, \boldsymbol{x}_i, \boldsymbol{x}^k_{j>i}, Y^k, Z^k) + \frac{\mu_k}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_i^k\|^2, \tag{4.1c}$$

  where $\boldsymbol{u}_{j<i}$ denotes $(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{i-1})$ and $\boldsymbol{x}^k_{j>i}$ denotes $(\boldsymbol{x}^k_{i+1}, \cdots, \boldsymbol{x}^k_r)$.

After computing U, we compute V in one of the following 3 ways for a given $\sigma_k > 0$:

- **Proximal**

$$V \in \operatorname*{Argmin}_{Y}\ \Phi(Y) + \mathcal{H}_\alpha(U, Y, Z^k) + \frac{\sigma_k}{2}\|Y - Y^k\|_F^2. \tag{4.2a}$$

- **Prox-linear**

$$V \in \operatorname*{Argmin}_{Y}\ \Phi(Y) + \langle \nabla_Y \mathcal{H}_\alpha(U, Y^k, Z^k),\, Y - Y^k\rangle + \frac{\sigma_k}{2}\|Y - Y^k\|_F^2. \tag{4.2b}$$

- **Hierarchical-prox** If Φ is column-wise separable, i.e., $\Phi(Y) = \sum_{i=1}^r \phi_i(\boldsymbol{y}_i)$ for $Y = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_r] \in \mathbb{R}^{n\times r}$, we can update V column-by-column. Specifically, for $i = 1, 2, \cdots, r$, compute

$$\boldsymbol{v}_i \in \operatorname*{Argmin}_{\boldsymbol{y}_i}\ \phi_i(\boldsymbol{y}_i) + \mathcal{H}_\alpha(U, \boldsymbol{v}_{j<i}, \boldsymbol{y}_i, \boldsymbol{y}^k_{j>i}, Z^k) + \frac{\sigma_k}{2}\|\boldsymbol{y}_i - \boldsymbol{y}_i^k\|^2, \tag{4.2c}$$

  where $\boldsymbol{v}_{j<i}$ denotes $(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{i-1})$ and $\boldsymbol{y}^k_{j>i}$ denotes $(\boldsymbol{y}^k_{i+1}, \cdots, \boldsymbol{y}^k_r)$.

For notational simplicity, we further let

$$\rho := \left\| \mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A} \right\|^2 \tag{4.3}$$

and let $\gamma \geq 0$ be a nonnegative number satisfying

$$(\alpha+\gamma)\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succeq 0. \tag{4.4}$$

REMARK 4.1 (**Comments on "hierarchical-prox"**). *The hierarchical-prox updating scheme requires the column-wise separability of Ψ or Φ. This is satisfied for many common regularizers, for example, $\|\cdot\|_F^2$, $\|\cdot\|_1$, $\|\cdot\|_p^p$ $(0 < p < 1)$, and the indicator function of the nonnegativity (or box) constraint.*

REMARK 4.2 (**Comments on ρ and γ**). *Since $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$, we see that the eigenvalues of $\mathcal{A}^*\mathcal{A}$ are either 0 or 1. Then, the eigenvalues of $\mathcal{I} - \frac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}$ must be either 1 or $\frac{\alpha}{\alpha+\beta}$, and hence $\rho = \max\left\{1, \, \alpha^2/(\alpha+\beta)^2\right\}$. Similarly, the eigenvalues of $-(\alpha\mathcal{I} + \beta\mathcal{A}^*\mathcal{A})$ are either $-\alpha$ or $-(\alpha+\beta)$. Then, (4.4) is satisfied whenever $\gamma \geq \max\{0, \, -\alpha, \, -(\alpha+\beta)\}$.*

Now, we are ready to present NAUM as Algorithm 1.

Algorithm 1 NAUM for finding a stationary point of $\mathcal{F}$

Input: (X^0, Y^0), α and β such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, ρ as in (4.3), $\gamma \geq 0$ satisfying (4.4), $\tau > 1$, $c > 0$, $\mu^{\min} > 0$, $\sigma^{\max} > \sigma^{\min} > 0$, and an integer $N \geq 0$. Set $k = 0$.

while a termination criterion is not met, **do**

Step 1. Compute Z^k by

$$Z^k = \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)\left(X^k(Y^k)^\top\right) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b}). \tag{4.5}$$

Step 2. Choose $\mu_k^0 \geq \mu^{\min}$ and $\sigma_k^0 \in [\sigma^{\min}, \, \sigma^{\max}]$ arbitrarily. Set $\tilde{\mu}_k = \mu_k^0$, $\sigma_k = \sigma_k^0$ and $\mu_k^{\max} = (\alpha + 2\gamma\rho)\|Y^k\|^2 + c$.

(2a) Set $\mu_k \leftarrow \min\{\tilde{\mu}_k, \, \mu_k^{\max}\}$. Compute U by either (4.1a), (4.1b) or (4.1c).

(2b) Compute V by either (4.2a), (4.2b) or (4.2c).

(2c) If

$$\mathcal{F}(U,V) - \max_{[k-N]_+ \leq i \leq k} \mathcal{F}(X^i, Y^i) \leq -\frac{c}{2}\left(\|U - X^k\|_F^2 + \|V - Y^k\|_F^2\right), \tag{4.6}$$

then go to **Step 3**.

(2d) If $\mu_k = \mu_k^{\max}$, set $\sigma_k^{\max} = (\alpha + 2\gamma\rho)\|U\|^2 + c$, $\sigma_k \leftarrow \min\{\tau\sigma_k, \, \sigma_k^{\max}\}$ and then, go to step **(2b)**; otherwise, set $\tilde{\mu}_k \leftarrow \tau\mu_k$ and $\sigma_k \leftarrow \tau\sigma_k$ and then, go to step **(2a)**.

Step 3. Set $X^{k+1} \leftarrow U$, $Y^{k+1} \leftarrow V$, $\bar{\mu}_k \leftarrow \mu_k$, $\bar{\sigma}_k \leftarrow \sigma_k$, $k \leftarrow k+1$ and go to **Step 1**.

end while

Output: (X^k, Y^k)

In Algorithm 1, the update for Z^k is given explicitly. This is motivated by the condition on Z at a stationary point of $\Theta_{\alpha,\beta}$; see (3.5c). In fact, following the same arguments in (3.9), we see that (3.5c) always holds at (X^k, Y^k, Z^k) with Z^k

given in (4.5) when $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$ and $\frac{1}{\alpha}+\frac{1}{\beta}=1$. If, in addition, $\alpha\mathcal{I}+\beta\mathcal{A}^*\mathcal{A}\succ 0$ holds, one can show that Z^k is actually the optimal solution to the problem $\min_Z\left\{\Theta_{\alpha,\beta}(X^k,Y^k,Z)\right\}$. In this case, our NAUM with $N=0$ in (4.6) can be viewed as an alternating-minimization-based method (see, for example, [1, 40]) applied to the problem $\min_{X,Y,Z}\{\Theta_{\alpha,\beta}(X,Y,Z)\}$. However, if $\alpha\mathcal{I}+\beta\mathcal{A}^*\mathcal{A}\not\succeq 0$,[2] then the corresponding $\inf_Z\left\{\Theta_{\alpha,\beta}(X^k,Y^k,Z)\right\}=-\infty$ for all k, and Z^k is only a stationary point of $Z\mapsto\Theta_{\alpha,\beta}(X^k,Y^k,Z)$. In this case, the function value of $\Theta_{\alpha,\beta}$ may increase after updating Z by (4.5). Fortunately, as we shall see later in (5.8) and (5.9), as long as $\mathcal{A}\mathcal{A}^*=\mathcal{I}_q$ and $\frac{1}{\alpha}+\frac{1}{\beta}=1$, we still have $\Theta_{\alpha,\beta}(X^{k+1},Y^{k+1},Z^k)<\Theta_{\alpha,\beta}(X^k,Y^k,Z^k)$ by updating X^{k+1} and Y^{k+1} with properly chosen parameters μ_k and σ_k. Thus, if the possible increase in $\Theta_{\alpha,\beta}$ induced by the Z-update is not too large, one can still ensure $\Theta_{\alpha,\beta}(X^{k+1},Y^{k+1},Z^{k+1})<\Theta_{\alpha,\beta}(X^k,Y^k,Z^k)$. Moreover, it can be seen from Lemma 3.1 and (4.5) that $\mathcal{F}(X^k,Y^k)=\Theta_{\alpha,\beta}(X^k,Y^k,Z^k)$ and hence the decrease of $\Theta_{\alpha,\beta}$ translates to that of $\mathcal{F}$ (see Lemma 5.1 below). In view of this, $\Theta_{\alpha,\beta}$ is a valid potential function for minimizing $\mathcal{F}$ as long as $\mathcal{A}\mathcal{A}^*=\mathcal{I}_q$ and $\frac{1}{\alpha}+\frac{1}{\beta}=1$, even when $\beta<0$ or $\alpha<0$. Allowing negative α or β makes our NAUM (even with $N=0$ in (4.6)) different from the classical alternating minimization schemes.

Our NAUM also allows U and V to be updated in three different ways respectively, and hence there are 9 possible combinations. Thus, one can choose suitable updating schemes to fit different applications. In particular, if Ψ or Φ are column-wise separable, taking advantage of the structure of $\Theta_{\alpha,\beta}$ and the fact that $XY^\top$ can be written as $\sum_{i=1}^r \boldsymbol{x}_i\boldsymbol{y}_i^\top$ with $X=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_r]\in\mathbb{R}^{m\times r}$ and $Y=[\boldsymbol{y}_1,\cdots,\boldsymbol{y}_r]\in\mathbb{R}^{n\times r}$, one can update X or Y column-wise even when $\mathcal{A}\neq\mathcal{I}$. The motivation for updating X (or Y) column-wise rather than updating the whole X (or Y) is that the resulting subproblems (4.1c) (or (4.2c)) can be reduced to the computation of the proximal mapping of ψ_i (or ϕ_i), which is easy for many commonly used ψ_i (or ϕ_i). Indeed, from (4.1c) and (4.2c), $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are given by

$$\begin{cases}\boldsymbol{u}_i\in\operatorname*{Argmin}_{\boldsymbol{x}_i}\left\{\psi_i(\boldsymbol{x}_i)+\dfrac{\alpha}{2}\left\|\boldsymbol{x}_i(\boldsymbol{y}_i^k)^\top-P_i^k\right\|_F^2+\dfrac{\mu_k}{2}\|\boldsymbol{x}_i-\boldsymbol{x}_i^k\|^2\right\},\\ \boldsymbol{v}_i\in\operatorname*{Argmin}_{\boldsymbol{y}_i}\left\{\phi_i(\boldsymbol{y}_i)+\dfrac{\alpha}{2}\left\|\boldsymbol{u}_i\boldsymbol{y}_i^\top-Q_i^k\right\|_F^2+\dfrac{\sigma_k}{2}\|\boldsymbol{y}_i-\boldsymbol{y}_i^k\|^2\right\},\end{cases}\tag{4.7}$$

where P_i^k and Q_i^k are defined by

$$\begin{aligned}P_i^k &:= Z^k-\textstyle\sum_{j=1}^{i-1}\boldsymbol{u}_j(\boldsymbol{y}_j^k)^\top-\sum_{j=i+1}^{r}\boldsymbol{x}_j^k(\boldsymbol{y}_j^k)^\top,\\ Q_i^k &:= Z^k-\textstyle\sum_{j=1}^{i-1}\boldsymbol{u}_j\boldsymbol{v}_j^\top-\sum_{j=i+1}^{r}\boldsymbol{u}_j(\boldsymbol{y}_j^k)^\top.\end{aligned}\tag{4.8}$$

Then, from Proposition 2.2, we can reformulate the subproblems in (4.7) and obtain the corresponding solutions by computing the proximal mappings of ψ_i and ϕ_i, which can be computed efficiently when ψ_i and ϕ_i are some common regularizers used in the literature. In particular, when $\psi_i(\cdot)$ and $\phi_i(\cdot)$ are $\|\cdot\|_1$, $\|\cdot\|_2^2$ or the indicator function of the box constraint, these subproblems have closed-form solutions. This updating strategy has also been used for NMF; see, for example, [8, 20, 21]. However, the methods used in [8, 20, 21] can only be applied for some specific problems with $\mathcal{A}=\mathcal{I}$, while NAUM can be applied for more general problems with $\mathcal{A}\mathcal{A}^*=\mathcal{I}_q$.

Our NAUM adapts a non-monotone line search criterion (see Step 2 in Algorithm 1) to improve the numerical performance. This is motivated by recent studies on

[2] This may happen when $0<\alpha<1$ so that $\beta=\alpha(\alpha-1)^{-1}<0$, or $0<\beta<1$ so that $\alpha=\beta(\beta-1)^{-1}<0$.

non-monotone algorithms with promising performances; see, for example, [7, 12, 39]. However, different from the non-monotone line search criteria used there, NAUM only includes (U, V) in the line search loop and checks the stopping criterion (4.6) after updating a pair of (U, V), rather than checking (4.6) immediately once U or V is updated. Thus, we do not need to compute the function value after updating each block of variable. This may reduce the cost of the line search and make NAUM more practical, especially when computing the function value is relatively expensive.

Before moving to the convergence analysis of NAUM, we would like to point out an interesting connection between NAUM and the low-rank matrix fitting algorithm, LMaFit [38], for solving the following matrix completion model without regularizers:

$$\min_{X,Y} \ \frac{1}{2}\left\|\mathcal{P}_\Omega(XY^\top - M)\right\|_F^2,$$

where Ω is the index set of the known entries of M, and $\mathcal{P}_\Omega(Z)$ keeps the entries of Z in Ω and sets the remaining ones to zero. If we apply our NAUM with (4.1a) and (4.2a), then at the k-th iteration, the iterates Z^k, X^{k+1} and Y^{k+1} are given by

$$\begin{aligned}
Z^k &= \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{P}_\Omega\right) X^k(Y^k)^\top + \tfrac{\beta}{\alpha+\beta}\mathcal{P}_\Omega(M),\\
X^{k+1} &= \left(\bar{\mu}_k X^k + \alpha Z^k Y^k\right)\left(\bar{\mu}_k I + \alpha (Y^k)^\top Y^k\right)^{-1},\\
Y^{k+1} &= \left(\bar{\sigma}_k Y^k + \alpha (Z^k)^\top X^{k+1}\right)\left(\bar{\sigma}_k I + \alpha (X^{k+1})^\top X^{k+1}\right)^{-1}.
\end{aligned}$$

One can verify that the sequence $\{(Z^k, X^{k+1}, Y^{k+1})\}$ above can be equivalently generated by the following scheme with $\widetilde{Z}^0 = \mathcal{P}_\Omega(M) + \mathcal{P}_{\Omega^c}\left(X^0(Y^0)^\top\right)$:

$$\begin{aligned}
Z^k &= \tfrac{\beta}{\alpha+\beta}\widetilde{Z}^k + \left(1 - \tfrac{\beta}{\alpha+\beta}\right) X^k(Y^k)^\top,\\
X^{k+1} &= \left(\bar{\mu}_k X^k + \alpha Z^k Y^k\right)\left(\bar{\mu}_k I + \alpha (Y^k)^\top Y^k\right)^{-1},\\
Y^{k+1} &= \left(\bar{\sigma}_k Y^k + \alpha (Z^k)^\top X^{k+1}\right)\left(\bar{\sigma}_k I + \alpha (X^{k+1})^\top X^{k+1}\right)^{-1},\\
\widetilde{Z}^{k+1} &= \mathcal{P}_\Omega(M) + \mathcal{P}_{\Omega^c}\left(X^{k+1}(Y^{k+1})^\top\right),
\end{aligned}$$

where Ω^c is the complement set of Ω. Surprisingly, when $\bar{\mu}_k = \bar{\sigma}_k = 0$, this scheme is exactly the SOR(successive over-relaxation)-like scheme used in LMaFit (see [38, Eq.(2.11)]) with $\omega := \frac{\beta}{\alpha+\beta}$ being an over-relaxation weight. With this connection, our NAUM, in some sense, can be viewed as an SOR-based algorithm. Moreover, just like the classical SOR for solving a system of linear equations, LMaFit with $\omega > 1$ also appears to be more efficient from the extensive numerical experiments reported in [38]. Then, it is natural to consider $\frac{\beta}{\alpha+\beta} > 1$ and hence $\frac{1}{\alpha} > 1$ (since $\frac{1}{\alpha} + \frac{1}{\beta} = 1$) in NAUM. This also gives some insights for the necessity of allowing more flexibilities in choosing α and β, and the promising performance of NAUM with a relatively small $\alpha \in (0, 1)$ as we shall see in Section 6.

5. Convergence analysis of NAUM.

In this section, we discuss the convergence properties of Algorithm 1. First, we present the first-order optimality conditions for the three different updating schemes in **(2a)** of Algorithm 1 as follows:

- **Proximal**

$$0 \in \partial\Psi(U) + \alpha\left(U(Y^k)^\top - Z^k\right)Y^k + \mu_k(U - X^k). \tag{5.1a}$$

- **Prox-linear**

$$0 \in \partial\Psi(U) + \alpha\left(X^k(Y^k)^\top - Z^k\right)Y^k + \mu_k(U - X^k). \tag{5.1b}$$

- **Hierarchical-prox** For $i = 1, 2, \cdots, r$,

$$0 \in \partial\psi_i(\boldsymbol{u}_i) + \alpha\left(\sum_{j=1}^{i}\boldsymbol{u}_j(\boldsymbol{y}_j^k)^\top + \sum_{j=i+1}^{r}\boldsymbol{x}_j^k(\boldsymbol{y}_j^k)^\top - Z^k\right)\boldsymbol{y}_i^k + \mu_k(\boldsymbol{u}_i - \boldsymbol{x}_i^k). \tag{5.1c}$$

Similarly, the first-order optimality conditions for the three different updating schemes in **(2b)** of Algorithm 1 are

- **Proximal**

$$0 \in \partial\Phi(V) + \alpha\left(UV^\top - Z^k\right)^\top U + \sigma_k(V - Y^k). \tag{5.2a}$$

- **Prox-linear**

$$0 \in \partial\Phi(V) + \alpha\left(U(Y^k)^\top - Z^k\right)^\top U + \sigma_k(V - Y^k). \tag{5.2b}$$

- **Hierarchical-prox** For $i = 1, 2, \cdots, r$,

$$0 \in \partial\phi_i(\boldsymbol{v}_i) + \alpha\left(\sum_{j=1}^{i}\boldsymbol{u}_j\boldsymbol{v}_j^\top + \sum_{j=i+1}^{r}\boldsymbol{u}_j(\boldsymbol{y}_j^k)^\top - Z^k\right)^\top \boldsymbol{u}_i + \sigma_k(\boldsymbol{v}_i - \boldsymbol{y}_i^k). \tag{5.2c}$$

We also need to make the following assumptions.

ASSUMPTION 5.1.

(a1) *Ψ, Φ are proper, closed, level-bounded functions and continuous on their domains respectively;*

(a2) $\mathcal{A}\mathcal{A}^* = \mathcal{I}_q$;

(a3) $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

REMARK 5.1. *(i) From* **(a1)***, one can see from [28, Theorem 1.9] that* $\inf\Psi$ *and* $\inf\Phi$ *are finite, i.e., Ψ and Φ are bounded from below. In particular, the iterates (4.1a), (4.1b), (4.1c), (4.2a), (4.2b) and (4.2c) are well defined; (ii) The continuity assumption in* **(a1)** *holds for many common regularizers, for example, ℓ_1-norm, nuclear norm and the indicator function of a nonempty closed set; (iii)* **(a2)** *is satisfied for some common linear maps, for example, the identity map and the sampling map.*

We start our convergence analysis by proving the following auxiliary lemma.

LEMMA 5.1 (**Sufficient descent of $\mathcal{F}$**). *Suppose that Assumption 5.1 holds. Let (X^k, Y^k) be generated by Algorithm 1 at the k-th iteration, and (U, V) be the candidate for (X^{k+1}, Y^{k+1}) generated by steps* **(2a)** *and* **(2b)***. Then, for any integer $k \geq 0$, we have*

$$\begin{aligned}&\mathcal{F}(U,V) - \mathcal{F}(X^k, Y^k)\\ &\leq -\frac{\mu_k - (\alpha + 2\gamma\rho)\|Y^k\|^2}{2}\|U - X^k\|_F^2 - \frac{\sigma_k - (\alpha+2\gamma\rho)\|U\|^2}{2}\|V - Y^k\|_F^2.\end{aligned} \tag{5.3}$$

Proof. First, from Lemma 3.1 and (4.5), we see that $\mathcal{F}(X^k, Y^k) = \Theta_{\alpha,\beta}(X^k, Y^k, Z^k)$. For any (U, V), let

$$W = \left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)\left(UV^\top\right) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b}). \tag{5.4}$$

Then, from Lemma 3.1, we have $\mathcal{F}(U, V) = \Theta_{\alpha,\beta}(U, V, W)$. Thus, to establish (5.3), we only need to consider the difference $\Theta_{\alpha,\beta}(U, V, W) - \Theta_{\alpha,\beta}(X^k, Y^k, Z^k)$.

We start by noting that

$$\begin{aligned}\mathcal{A}^*\mathcal{A}(W) &= \left(\mathcal{A}^*\mathcal{A} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\left(\mathcal{A}\mathcal{A}^*\right)\mathcal{A}\right)\left(UV^\top\right) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\left(\mathcal{A}\mathcal{A}^*\right)(\boldsymbol{b})\\ &= \tfrac{\alpha}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\left(UV^\top\right) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b}),\end{aligned} \tag{5.5}$$

where the last equality follows from **(a2)** in Assumption 5.1. Then, we obtain that

$$
\begin{aligned}
&\nabla_Z \Theta_{\alpha,\beta}(U,V,W) = \alpha(W - UV^\top) + \beta\mathcal{A}^*\mathcal{A}(W) - \beta\mathcal{A}^*(\boldsymbol{b}) \\
&= \alpha\Big[-\tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}(UV^\top) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b})\Big] + \beta\Big[\tfrac{\alpha}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\left(UV^\top\right) + \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*(\boldsymbol{b})\Big] - \beta\mathcal{A}^*(\boldsymbol{b}) = 0,
\end{aligned}
$$

where the second equality follows from (5.4) and (5.5). Moreover, since γ is chosen such that $(\alpha+\gamma)\mathcal{I} + \beta\mathcal{A}^*\mathcal{A} \succeq 0$ (see (4.4)), we see that, for any $k \geq 0$, the function $Z \longmapsto \Theta_{\alpha,\beta}(U,V,Z) + \frac{\gamma}{2}\|Z - Z^k\|_F^2$ is convex and hence

$$
\begin{aligned}
&\Theta_{\alpha,\beta}(U,V,Z^k) + \underbrace{\frac{\gamma}{2}\|Z^k - Z^k\|_F^2}_{=0} \\
&\geq \Theta_{\alpha,\beta}(U,V,W) + \frac{\gamma}{2}\|W - Z^k\|_F^2 + \langle \underbrace{\nabla_Z\Theta_{\alpha,\beta}(U,V,W)}_{=0} + \gamma(W - Z^k),\, Z^k - W\rangle,
\end{aligned}
$$

which implies that

$$
\Theta_{\alpha,\beta}(U,V,W) - \Theta_{\alpha,\beta}(U,V,Z^k) \leq \frac{\gamma}{2}\|W - Z^k\|_F^2. \tag{5.6}
$$

Then, substituting (4.5) and (5.4) into (5.6), we obtain

$$
\begin{aligned}
\Theta_{\alpha,\beta}(U,V,W) - \Theta_{\alpha,\beta}(U,V,Z^k) &\leq \tfrac{\gamma}{2}\left\|\left(\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right)\left(UV^\top - X^k(Y^k)^\top\right)\right\|_F^2 \\
&\leq \tfrac{\gamma}{2}\left\|\mathcal{I} - \tfrac{\beta}{\alpha+\beta}\mathcal{A}^*\mathcal{A}\right\|^2 \cdot \left\|UV^\top - X^k(Y^k)^\top\right\|_F^2 \\
&= \tfrac{\gamma\rho}{2}\left\|U(V - Y^k)^\top + (U - X^k)(Y^k)^\top\right\|_F^2 \\
&\leq \tfrac{\gamma\rho}{2}\Big(\left\|U(V - Y^k)^\top\right\|_F + \left\|(U - X^k)(Y^k)^\top\right\|_F\Big)^2 \\
&\overset{\text{(i)}}{\leq} \tfrac{\gamma\rho}{2}\Big(\|U\|\|V - Y^k\|_F + \|Y^k\|\|U - X^k\|_F\Big)^2 \\
&\overset{\text{(ii)}}{\leq} \gamma\rho\Big(\|U\|^2\|V - Y^k\|_F^2 + \|Y^k\|^2\|U - X^k\|_F^2\Big),
\end{aligned} \tag{5.7}
$$

where the equality follows from the definition of ρ in (4.3); (i) follows from the relation $\|AB\|_F \leq \|A\|\|B\|_F$; and (ii) follows from the relation $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

Next, we claim that

$$
\Theta_{\alpha,\beta}(U,V,Z^k) - \Theta_{\alpha,\beta}(U,Y^k,Z^k) \leq \frac{\alpha\|U\|^2 - \sigma_k}{2}\|V - Y^k\|_F^2, \tag{5.8}
$$

$$
\Theta_{\alpha,\beta}(U,Y^k,Z^k) - \Theta_{\alpha,\beta}(X^k,Y^k,Z^k) \leq \frac{\alpha\|Y^k\|^2 - \mu_k}{2}\|U - X^k\|_F^2. \tag{5.9}
$$

Below, we will only prove (5.8). The proof for (5.9) can be done in a similar way.

To prove (5.8), we consider the following three cases.

- Proximal: In this case, we have

$$
\begin{aligned}
&\Theta_{\alpha,\beta}(U,V,Z^k) - \Theta_{\alpha,\beta}(U,Y^k,Z^k) = \Phi(V) + \mathcal{H}_\alpha(U,V,Z^k) - \Phi(Y^k) - \mathcal{H}_\alpha(U,Y^k,Z^k) \\
&= \Big[\Phi(V) + \mathcal{H}_\alpha(U,V,Z^k) + \frac{\sigma_k}{2}\|V - Y^k\|_F^2\Big] - \big[\Phi(Y^k) + \mathcal{H}_\alpha(U,Y^k,Z^k)\big] - \frac{\sigma_k}{2}\|V - Y^k\|_F^2 \\
&\leq -\frac{\sigma_k}{2}\|V - Y^k\|_F^2,
\end{aligned}
$$

  where the inequality follows from the definition of V as a minimizer of (4.2a). This implies (5.8).

- Prox-linear: In this case, we have

$$\begin{aligned}
&\Theta_{\alpha,\beta}(U,V,Z^k)-\Theta_{\alpha,\beta}(U,Y^k,Z^k)=\Phi(V)+\mathcal{H}_\alpha(U,V,Z^k)-\Phi(Y^k)-\mathcal{H}_\alpha(U,Y^k,Z^k)\\
&\leq \Phi(V)+\mathcal{H}_\alpha(U,Y^k,Z^k)+\langle\nabla_Y\mathcal{H}_\alpha(U,Y^k,Z^k),\,V-Y^k\rangle+\frac{\alpha\|U\|^2}{2}\|V-Y^k\|_F^2\\
&\quad-\Phi(Y^k)-\mathcal{H}_\alpha(U,Y^k,Z^k)\\
&=\Phi(V)+\langle\nabla_Y\mathcal{H}_\alpha(U,Y^k,Z^k),V-Y^k\rangle+\frac{\sigma_k}{2}\|V-Y^k\|_F^2-\Phi(Y^k)+\frac{\alpha\|U\|^2-\sigma_k}{2}\|V-Y^k\|_F^2\\
&\leq\frac{\alpha\|U\|^2-\sigma_k}{2}\|V-Y^k\|_F^2,
\end{aligned}$$

  where the first inequality follows from the fact that $Y\mapsto\nabla_Y\mathcal{H}_\alpha(X,Y,Z)$ is Lipschitz with modulus $\alpha\|X\|^2$ and the last inequality follows from the definition of V as a minimizer of (4.2b).
- Hierarchical-prox: In this case, for any $1\leq i\leq r$, we have

$$\begin{aligned}
&\Theta_{\alpha,\beta}(U,\boldsymbol{v}_{j<i},\boldsymbol{v}_i,\boldsymbol{y}^k_{j>i},Z^k)-\Theta_{\alpha,\beta}(U,\boldsymbol{v}_{j<i},\boldsymbol{y}^k_i,\boldsymbol{y}^k_{j>i},Z^k)\\
&=\phi_i(\boldsymbol{v}_i)+\mathcal{H}_\alpha(U,\boldsymbol{v}_{j<i},\boldsymbol{v}_i,\boldsymbol{y}^k_{j>i},Z^k)-\phi_i(\boldsymbol{y}^k_i)-\mathcal{H}_\alpha(U,\boldsymbol{v}_{j<i},\boldsymbol{y}^k_i,\boldsymbol{y}^k_{j>i},Z^k)\\
&=\Big[\phi_i(\boldsymbol{v}_i)+\mathcal{H}_\alpha(U,\boldsymbol{v}_{j<i},\boldsymbol{v}_i,\boldsymbol{y}^k_{j>i},Z^k)+\frac{\sigma_k}{2}\|\boldsymbol{v}_i-\boldsymbol{y}^k_i\|^2\Big]-\frac{\sigma_k}{2}\|\boldsymbol{v}_i-\boldsymbol{y}^k_i\|^2\\
&\quad-\big[\phi_i(\boldsymbol{y}^k_i)+\mathcal{H}_\alpha(U,\boldsymbol{v}_{j<i},\boldsymbol{y}^k_i,\boldsymbol{y}^k_{j>i},Z^k)\big]\\
&\leq-\frac{\sigma_k}{2}\|\boldsymbol{v}_i-\boldsymbol{y}^k_i\|^2,
\end{aligned}$$

  where the inequality follows from the definition of $\boldsymbol{v}_i$ as a minimizer of (4.2c). Then, summing the above relation from $i=r$ to $i=1$ and simplifying the resulting inequality, we obtain (5.8).

The inequality (5.9) can be obtained via a similar argument.

Now, summing (5.7), (5.8) and (5.9), and using $\mathcal{F}(U,V)=\Theta_{\alpha,\beta}(U,V,W)$ and $\mathcal{F}(X^k,Y^k)=\Theta_{\alpha,\beta}(X^k,\,Y^k,Z^k)$, we obtain (5.3). This completes the proof. ∎

From Lemma 5.1, we see that the sufficient descent of $\mathcal{F}(X,Y)$ can be guaranteed as long as μ_k and σ_k are sufficiently large. Thus, based on this lemma, we can show in the following proposition that our non-monotone line search criterion in Algorithm 1 is well defined.

Proposition 5.2 (**Well-definedness of the non-monotone line search criterion**). *Suppose that Assumption 5.1 holds and Algorithm 1 is applied. Then, for each $k\geq 0$, the line search criterion (4.6) is satisfied after finitely many inner iterations.*

Proof. We prove this proposition by contradiction. Assume that there exists a $k\geq 0$ such that the line search criterion (4.6) cannot be satisfied after finitely many inner iterations. Note from **(2a)** and **(2d)** in Step 2 of Algorithm 1 that $\mu_k\leq\mu_k^{\max}=(\alpha+2\gamma\rho)\|Y^k\|^2+c$ and hence $\mu_k=\mu_k^{\max}$ must be satisfied after finitely many inner iterations. Let n_k denote the number of inner iterations when $\mu_k=\mu_k^{\max}$ is satisfied for the *first* time. If $\mu_k^0\geq\mu_k^{\max}$, then $n_k=1$; otherwise, we have

$$\mu^{\min}\tau^{n_k-2}\leq\mu_k^0\tau^{n_k-2}<\mu_k^{\max},$$

which implies that

$$n_k\leq\left\lfloor\frac{\log(\mu_k^{\max})-\log(\mu^{\min})}{\log\tau}+2\right\rfloor. \tag{5.10}$$

Then, from **(2d)** in Step 2 of Algorithm 1, we have $U \equiv U_{\mu_k^{\max}}$ and $\sigma_k^{\max} = (\alpha + 2\gamma\rho)\|U_{\mu_k^{\max}}\|^2 + c$ after at most $n_k + 1$ inner iterations, where $U_{\mu_k^{\max}}$ is computed by (4.1a), (4.1b) or (4.1c) with $\mu_k = \mu_k^{\max}$. Moreover, we see that $\sigma_k = \sigma_k^{\max}$ must be satisfied after finitely many inner iterations. Similarly, let $\hat{n}_k$ denote the number of inner iterations when $\sigma_k = \sigma_k^{\max}$ is satisfied for the *first* time. If $\sigma_k^0 > \sigma_k^{\max}$, then $\hat{n}_k = n_k$; if $\sigma_k^0 = \sigma_k^{\max}$, then $\hat{n}_k = 0$; otherwise, we have

$$\sigma^{\min}\tau^{\hat{n}_k - 1} \le \sigma_k^0 \tau^{\hat{n}_k - 1} < \sigma_k^{\max},$$

which implies that

$$\hat{n}_k \le \left\lfloor \frac{\log(\sigma_k^{\max}) - \log(\sigma^{\min})}{\log \tau} + 1 \right\rfloor.$$

Thus, after at most $\max\{n_k,\, \hat{n}_k\} + 1$ inner iterations, we must have $V \equiv V_{\sigma_k^{\max}}$, where $V_{\sigma_k^{\max}}$ is computed by (4.2a), (4.2b) or (4.2c) with $\sigma_k = \sigma_k^{\max}$. Therefore, after at most $\max\{n_k,\, \hat{n}_k\} + 1$ inner iterations, we have

$$\begin{aligned}
&\mathcal{F}(U_{\mu_k^{\max}}, V_{\sigma_k^{\max}}) - \mathcal{F}(X^k, Y^k) \\
&\le - \frac{\mu_k^{\max} - (\alpha + 2\gamma\rho)\|Y^k\|^2}{2} \|U_{\mu_k^{\max}} - X^k\|_F^2 - \frac{\sigma_k^{\max} - (\alpha + 2\gamma\rho)\|U_{\mu_k^{\max}}\|^2}{2} \|V_{\sigma_k^{\max}} - Y^k\|_F^2 \\
&= - \frac{c}{2} \left(\|U_{\mu_k^{\max}} - X^k\|_F^2 + \|V_{\sigma_k^{\max}} - Y^k\|_F^2 \right),
\end{aligned}$$

where the inequality follows from (5.3) and the equality follows from $\mu_k^{\max} = (\alpha + 2\gamma\rho)\|Y^k\|^2 + c$ and $\sigma_k^{\max} = (\alpha + 2\gamma\rho)\|U_{\mu_k^{\max}}\|^2 + c$. This together with

$$\mathcal{F}(X^k, Y^k) \le \max_{[k-N]_+ \le i \le k} \mathcal{F}(X^i, Y^i)$$

implies that (4.6) must be satisfied after at most $\max\{n_k,\, \hat{n}_k\} + 1$ inner iterations, which leads to a contradiction. ∎

Now, we are ready to prove our main convergence result, which characterizes a cluster point of the sequence generated by Algorithm 1. Our proof of statement (ii) in the following theorem is similar to that of [39, Lemma 4]. However, the arguments involved are more intricate since we have two blocks of variables in our line search loop.

THEOREM 5.3. *Suppose that Assumption 5.1 holds. Let $\{(X^k, Y^k)\}$ be the sequence generated by Algorithm 1. Then,*

(i) ***(boundedness of sequence)*** *$\{(X^k, Y^k)\}$, $\{\bar{\mu}_k\}$ and $\{\bar{\sigma}_k\}$ are bounded;*
(ii) ***(diminishing successive changes)*** $\lim\limits_{k\to\infty} \|X^{k+1} - X^k\|_F + \|Y^{k+1} - Y^k\|_F = 0$*;*
(iii) ***(global subsequential convergence)*** *any cluster point (X^*, Y^*) of $\{(X^k, Y^k)\}$ is a stationary point of $\mathcal{F}$.*

Proof. Statement (i). We first show that

$$\mathcal{F}(X^k, Y^k) \le \mathcal{F}(X^0, Y^0) \tag{5.11}$$

for all $k \ge 1$. We will prove it by induction. Indeed, for $k = 1$, it follows from Proposition 5.2 that

$$\mathcal{F}(X^1, Y^1) - \mathcal{F}(X^0, Y^0) \le -\frac{c}{2} \left(\|X^1 - X^0\|_F^2 + \|Y^1 - Y^0\|_F^2 \right) \le 0$$

is satisfied after finitely many inner iterations. Hence, (5.11) holds for $k = 1$. We now suppose that (5.11) holds for all $k \le K$ for some integer $K \ge 1$. Then, we only need to show that (5.11) also holds for $k = K + 1$. For $k = K + 1$, we have

$$\begin{aligned}\mathcal{F}(X^{K+1}, Y^{K+1}) - \mathcal{F}(X^0, Y^0) &\le \mathcal{F}(X^{K+1}, Y^{K+1}) - \max_{[K-N]_+ \le i \le K} \mathcal{F}(X^k, Y^k) \\ &\le -\frac{c}{2}\left(\|X^{K+1} - X^K\|_F^2 + \|Y^{K+1} - Y^K\|_F^2\right) \le 0,\end{aligned}$$

where the first inequality follows from the induction hypothesis and the second inequality follows from (4.6). Hence, (5.11) holds for $k = K + 1$. This completes the induction. Then, from (5.11), we have that for any $k \ge 0$,

$$\mathcal{F}(X^0, Y^0) \ge \mathcal{F}(X^k, Y^k) = \Psi(X^k) + \Phi(Y^k) + \frac{1}{2}\left\|\mathcal{A}(X^k(Y^k)^\top) - \boldsymbol{b}\right\|^2,$$

which, together with **(a1)** in Assumption 5.1, implies that the sequences $\{X^k\}$, $\{Y^k\}$ and $\{\|\mathcal{A}(X^k(Y^k)^\top) - \boldsymbol{b}\|\}$ are bounded. Moreover, from Step 2 and Step 3 in Algorithm 1, it is easy to see $\bar{\mu}_k \le \mu_k^{\max} = (\alpha + 2\gamma\rho)\|Y^k\|^2 + c$ for all k. Since $\{Y^k\}$ is bounded, the sequences $\{\mu_k^{\max}\}$ and $\{\bar{\mu}_k\}$ are bounded. Next, we prove the boundedness of $\{\bar{\sigma}_k\}$. Indeed, at the k-th iteration, there are three possibilities:

- $\bar{\mu}_k < \mu_k^{\max}$: In this case, we have $\bar{\sigma}_k \le \sigma_k^0 \tau^{\tilde{n}_k} \le \sigma^{\max}\tau^{\tilde{n}_k}$, where $\tilde{n}_k$ denotes the number of inner iterations for the line search at the k-th iteration and $\tilde{n}_k \le \max\left\{1, \left\lfloor \frac{\log(\mu_k^{\max}) - \log(\mu^{\min})}{\log \tau} + 2 \right\rfloor\right\}$ (see (5.10) and the discussions preceding it).
- $\bar{\mu}_k = \mu_k^{\max}$ and $\bar{\sigma}_k > \sigma_k^{\max}$: In this case, we have $\bar{\sigma}_k \le \sigma_k^0 \tau^{\tilde{n}_k} \le \sigma^{\max}\tau^{\tilde{n}_k}$, where $\tilde{n}_k \le \max\left\{1, \left\lfloor \frac{\log(\mu_k^{\max}) - \log(\mu^{\min})}{\log \tau} + 2 \right\rfloor\right\}$.
- Otherwise, we have $\bar{\sigma}_k \le \sigma_k^{\max} = (\alpha + 2\gamma\rho)\|X^{k+1}\|^2 + c$.

Note that $\{\tilde{n}_k\}$ is bounded as $\{\mu_k^{\max}\}$ is bounded. Thus, $\{\bar{\sigma}_k\}$ is bounded as the sequences $\{X^k\}$ and $\{\tilde{n}_k\}$ are bounded. This proves statement (i).

Statement (ii). We first claim that any cluster point of $\{(X^k, Y^k)\}$ is in $\mathrm{dom}\mathcal{F}$. Since $\{(X^k, Y^k)\}$ is bounded from statement (i), there exists at least one cluster point. Suppose that (X^*, Y^*) is a cluster point of $\{(X^k, Y^k)\}$ and let $\{(X^{k_i}, Y^{k_i})\}$ be a convergent subsequence such that $\lim_{i\to\infty}(X^{k_i}, Y^{k_i}) = (X^*, Y^*)$. Then, from the lower semicontinuity of $\mathcal{F}$ (since Ψ, Φ are closed by **(a1)** in Assumption 5.1) and (5.11), we have

$$\mathcal{F}(X^*, Y^*) \le \lim_{i\to\infty} \mathcal{F}(X^{k_i}, Y^{k_i}) \le \mathcal{F}(X^0, Y^0),$$

which implies that $\mathcal{F}(X^*, Y^*)$ is finite and hence $(X^*, Y^*) \in \mathrm{dom}\mathcal{F}$.

For notational simplicity, from now on, we let $\Delta_{X^k} := X^{k+1} - X^k$, $\Delta_{Y^k} := Y^{k+1} - Y^k$, $\Delta_{Z^k} := Z^{k+1} - Z^k$ and

$$\ell(k) = \arg\max_i \{\, \mathcal{F}(X^i, Y^i) \,:\, i = [k-N]_+, \cdots, k \,\}. \tag{5.12}$$

Then, the line search criterion (4.6) can be rewritten as

$$\mathcal{F}(X^{k+1}, Y^{k+1}) - \mathcal{F}(X^{\ell(k)}, Y^{\ell(k)}) \le -\frac{c}{2}\left(\|\Delta_{X^k}\|_F^2 + \|\Delta_{Y^k}\|_F^2\right) \le 0. \tag{5.13}$$

Observe that

$$
\begin{aligned}
&\mathcal{F}(X^{\ell(k+1)}, Y^{\ell(k+1)})\\
&= \max_{[k+1-N]_+ \le i \le k+1} \mathcal{F}(X^i, Y^i) = \max\left\{\mathcal{F}(X^{k+1}, Y^{k+1}), \max_{[k+1-N]_+ \le i \le k} \mathcal{F}(X^i, Y^i)\right\}\\
&\overset{\text{(i)}}{\le} \max\left\{\mathcal{F}(X^{\ell(k)}, Y^{\ell(k)}), \max_{[k+1-N]_+ \le i \le k} \mathcal{F}(X^i, Y^i)\right\}\\
&\le \max\left\{\mathcal{F}(X^{\ell(k)}, Y^{\ell(k)}), \max_{[k-N]_+ \le i \le k} \mathcal{F}(X^i, Y^i)\right\}\\
&\overset{\text{(ii)}}{=} \max\left\{\mathcal{F}(X^{\ell(k)}, Y^{\ell(k)}), \mathcal{F}(X^{\ell(k)}, Y^{\ell(k)})\right\} = \mathcal{F}(X^{\ell(k)}, Y^{\ell(k)}),
\end{aligned}
$$

where (i) follows from (5.13) and (ii) follows from (5.12). Therefore, the sequence $\{\mathcal{F}(X^{\ell(k)}, Y^{\ell(k)})\}$ is non-increasing. Since $\mathcal{F}(X^{\ell(k)},\, Y^{\ell(k)})$ is also bounded from below (due to **(a1)** in Assumption 5.1), we conclude that there exists a number $\widetilde{\mathcal{F}}$ such that

$$
\lim_{k\to\infty} \mathcal{F}(X^{\ell(k)},\, Y^{\ell(k)}) = \widetilde{\mathcal{F}}. \tag{5.14}
$$

We next prove by induction that for all $j \ge 1$,

$$
\begin{cases}
\lim\limits_{k\to\infty} \Delta_{X^{\ell(k)-j}} = \lim\limits_{k\to\infty} \Delta_{Y^{\ell(k)-j}} = 0, & \text{(5.15a)}\\
\lim\limits_{k\to\infty} \mathcal{F}(X^{\ell(k)-j}, Y^{\ell(k)-j}) = \widetilde{\mathcal{F}}. & \text{(5.15b)}
\end{cases}
$$

We first prove (5.15a) and (5.15b) for $j = 1$. Applying (5.13) with k replaced by $\ell(k) - 1$, we obtain

$$
\mathcal{F}(X^{\ell(k)}, Y^{\ell(k)}) - \mathcal{F}(X^{\ell(\ell(k)-1)}, Y^{\ell(\ell(k)-1)}) \le -\frac{c}{2}\left(\|\Delta_{X^{\ell(k)-1}}\|_F^2 + \|\Delta_{Y^{\ell(k)-1}}\|_F^2\right),
$$

which, together with (5.14), implies that

$$
\lim_{k\to\infty} \Delta_{X^{\ell(k)-1}} = \lim_{k\to\infty} \Delta_{Y^{\ell(k)-1}} = 0. \tag{5.16}
$$

Then, from (5.14) and (5.16), we have

$$
\begin{aligned}
\widetilde{\mathcal{F}} &= \lim_{k\to\infty} \mathcal{F}(X^{\ell(k)}, Y^{\ell(k)}) = \lim_{k\to\infty} \mathcal{F}(X^{\ell(k)-1} + \Delta_{X^{\ell(k)-1}}, Y^{\ell(k)-1} + \Delta_{Y^{\ell(k)-1}})\\
&= \lim_{k\to\infty} \mathcal{F}(X^{\ell(k)-1}, Y^{\ell(k)-1}),
\end{aligned}
$$

where the last equality follows because $\{(X^k, Y^k)\}$ is bounded, any cluster point of $\{(X^k, Y^k)\}$ is in $\mathrm{dom}\mathcal{F}$ and $\mathcal{F}$ is uniformly continuous on any compact subset of $\mathrm{dom}\mathcal{F}$ under **(a1)** in Assumption 5.1. Thus, (5.15a) and (5.15b) hold for $j = 1$.

We next suppose that (5.15a) and (5.15b) hold for $j = J$ for some $J \ge 1$. It remains to show that they also hold for $j = J + 1$. Indeed, from (5.13) with k replaced by $\ell(k) - J - 1$ (here, without loss of generality, we assume that k is large enough such that $\ell(k) - J - 1$ is nonnegative), we have

$$
\mathcal{F}(X^{\ell(k)-J}, Y^{\ell(k)-J}) - \mathcal{F}(X^{\ell(\ell(k)-J-1)}, Y^{\ell(\ell(k)-J-1)}) \le -\tfrac{c}{2}\left(\|\Delta_{X^{\ell(k)-J-1}}\|_F^2 + \|\Delta_{Y^{\ell(k)-J-1}}\|_F^2\right),
$$

which implies that

$$
\|\Delta_{X^{\ell(k)-J-1}}\|_F^2 + \|\Delta_{Y^{\ell(k)-J-1}}\|_F^2 \le \tfrac{2}{c}\left(\mathcal{F}(X^{\ell(\ell(k)-J-1)}, Y^{\ell(\ell(k)-J-1)}) - \mathcal{F}(X^{\ell(k)-J}, Y^{\ell(k)-J})\right).
$$

This together with (5.14) and the induction hypothesis implies that

$$\lim_{k\to\infty} \Delta_{X^{\ell(k)-(J+1)}} = \lim_{k\to\infty} \Delta_{Y^{\ell(k)-(J+1)}} = 0.$$

Thus, (5.15a) holds for $j = J+1$. From this, we further have

$$\begin{aligned}\lim_{k\to\infty} \mathcal{F}(X^{\ell(k)-(J+1)}, Y^{\ell(k)-(J+1)}) &= \lim_{k\to\infty} \mathcal{F}(X^{\ell(k)-J} - \Delta_{X^{\ell(k)-(J+1)}}, Y^{\ell(k)-J} - \Delta_{Y^{\ell(k)-(J+1)}}) \\ &= \lim_{k\to\infty} \mathcal{F}(X^{\ell(k)-J},\, Y^{\ell(k)-J}) = \widetilde{\mathcal{F}},\end{aligned}$$

where the second equality follows because $\{(X^k, Y^k)\}$ is bounded, any cluster point of $\{(X^k, Y^k)\}$ is in $\mathrm{dom}\mathcal{F}$ and $\mathcal{F}$ is uniformly continuous on any compact subset of $\mathrm{dom}\mathcal{F}$ under **(a1)** in Assumption 5.1. Hence, (5.15b) also holds for $j = J+1$. This completes the induction.

We are now ready to prove the main result in this statement. Indeed, from (5.12), we can see $k - N \leq \ell(k) \leq k$ (without loss of generality, we assume that k is large enough such that $k \geq N$). Thus, for any k, we must have $k - N - 1 = \ell(k) - j_k$ for $1 \leq j_k \leq N+1$. Then, we have

$$\begin{aligned}\|\Delta_{X^{k-N-1}}\|_F &= \|\Delta_{X^{\ell(k)-j_k}}\|_F \leq \max_{1\leq j\leq N+1} \|\Delta_{X^{\ell(k)-j}}\|_F, \\ \|\Delta_{Y^{k-N-1}}\|_F &= \|\Delta_{Y^{\ell(k)-j_k}}\|_F \leq \max_{1\leq j\leq N+1} \|\Delta_{Y^{\ell(k)-j}}\|_F.\end{aligned}$$

This together with (5.15a) implies that

$$\begin{aligned}\lim_{k\to\infty} \Delta_{X^k} &= \lim_{k\to\infty} \Delta_{X^{k-N-1}} = 0, \\ \lim_{k\to\infty} \Delta_{Y^k} &= \lim_{k\to\infty} \Delta_{Y^{k-N-1}} = 0.\end{aligned}$$

This proves the statement (ii).

Statement (iii). Again, let (X^*, Y^*) be a cluster point of $\{(X^k, Y^k)\}$ and let $\{(X^{k_i}, Y^{k_i})\}$ be a convergent subsequence such that $\lim\limits_{i\to\infty}(X^{k_i}, Y^{k_i}) = (X^*, Y^*)$. Recall that $(X^*, Y^*) \in \mathrm{dom}\mathcal{F}$. On the other hand, it is easy to see from (4.5) that $\lim\limits_{i\to\infty} Z^{k_i} = Z^*$, where Z^* is given by (3.4). Thus, it can be shown as in (3.9) that

$$\alpha(Z^* - X^*(Y^*)^\top) + \beta\mathcal{A}^*(\mathcal{A}(Z^*) - \boldsymbol{b}) = 0. \tag{5.17}$$

We next show that

$$\begin{cases} 0 \in \partial\Psi(X^*) + \alpha(X^*(Y^*)^\top - Z^*)Y^*, & (5.18\text{a}) \\ 0 \in \partial\Phi(Y^*) + \alpha(X^*(Y^*)^\top - Z^*)^\top X^*. & (5.18\text{b}) \end{cases}$$

We start by showing (5.18a) in the following cases:

- **Proximal**&**Prox-linear**: In these two cases, passing to the limit along $\{(X^{k_i}, Y^{k_i})\}$ in (5.1a) or (5.1b) with X^{k_i+1} in place of U and $\bar{\mu}_{k_i}$ in place of μ_k, and invoking **(a1)** in Assumption 5.1, statements (i), (ii), $(X^*, Y^*) \in \mathrm{dom}\mathcal{F}$ and (2.1), we obtain (5.18a).
- **Hierarchical-prox**: In this case, passing to the limit along $\{(X^{k_i}, Y^{k_i})\}$ in (5.1c) with X^{k_i+1} in place of U and $\bar{\mu}_{k_i}$ in place of μ_k, and invoking **(a1)** in Assumption 5.1, statements (i), (ii), $(X^*, Y^*) \in \mathrm{dom}\mathcal{F}$ and (2.1), we have

  $$0 \in \partial\psi_i(\boldsymbol{x}_i^*) + \alpha(X^*(Y^*)^\top - Z^*)\boldsymbol{y}_i^*$$

  for any $i = 1, 2, \cdots, r$. Then, stacking them up, we obtain (5.18a).

Similarly, we can obtain (5.18b). Thus, combining (5.17), (5.18a) and (5.18b), we see that (X^*, Y^*, Z^*) is a stationary point of $\Theta_{\alpha,\beta}$, which further implies (X^*, Y^*) is a stationary point of $\mathcal{F}$ from Theorem 3.3. This proves statement (iii). ∎

REMARK 5.2 (**Comment on (a3) in Assumption 5.1**). *If Φ and Ψ are the indicator functions of some nonempty closed sets, Theorem 5.3 can remain valid under the weaker condition on α and β that $\frac{1}{\alpha} + \frac{1}{\beta} > 0$ with a slight modification in (4.6) of Algorithm 1. Indeed, when Φ and Ψ are the indicator functions, one can see from Remark 3.1 and the proofs of Lemma 5.1 and Proposition 5.2 that if $\frac{1}{\alpha} + \frac{1}{\beta} > 0$, then*

$$\begin{aligned}&\mathcal{F}(U,V) - \mathcal{F}(X^k,Y^k) = \left(\tfrac{1}{\alpha} + \tfrac{1}{\beta}\right)\left(\Theta_{\alpha,\beta}(U,V,W) - \Theta_{\alpha,\beta}(X^k,Y^k,Z^k)\right)\\ &\le -\left(\tfrac{1}{\alpha} + \tfrac{1}{\beta}\right)\left(\tfrac{\mu_k - (\alpha+2\gamma\rho)\|Y^k\|^2}{2}\cdot\|U - X^k\|_F^2 + \tfrac{\sigma_k - (\alpha+2\gamma\rho)\|U\|^2}{2}\cdot\|V - Y^k\|_F^2\right),\end{aligned}$$

and the line search criterion is well defined with c replaced by $\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)c$. Moreover, recalling [28, Exercise 8.14], we see that $\partial\Psi$ and $\partial\Phi$ are normal cones. Thus, following Remark 3.2 and the similar augments in Theorem 5.3, we can obtain the same results when $\frac{1}{\alpha} + \frac{1}{\beta} > 0$ with c replaced by $\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)c$ in (4.6) of Algorithm 1.

REMARK 5.3 (**Comments on updating $\mu_k^{\max}$ and $\sigma_k^{\max}$**). *In Algorithm 1, we need to evaluate $\mu_k^{\max} = (\alpha + 2\gamma\rho)\|Y^k\|^2 + c$ and $\sigma_k^{\max} = (\alpha + 2\gamma\rho)\|U\|^2 + c$ in each iteration. However, computing the spectral norms of Y^k and U might be costly, especially when r is large. Hence, in our experiments, instead of computing $\|Y^k\|^2$ and $\|U\|^2$, we compute $\|Y^k\|_F^2$ and $\|U\|_F^2$, and update $\mu_k^{\max}$ and $\sigma_k^{\max}$ by $\mu_k^{\max} = (\alpha + 2\gamma\rho)\|Y^k\|_F^2 + c$ and $\sigma_k^{\max} = (\alpha + 2\gamma\rho)\|U\|_F^2 + c$ instead. Since $\|Y^k\| \le \|Y^k\|_F$ and $\|U\| \le \|U\|_F$, it follows from (5.3) that*

$$\mathcal{F}(U,V) - \mathcal{F}(X^k,Y^k) \le -\tfrac{\mu_k - (\alpha+2\gamma\rho)\|Y^k\|_F^2}{2}\,\|U - X^k\|_F^2 - \tfrac{\sigma_k - (\alpha+2\gamma\rho)\|U\|_F^2}{2}\,\|V - Y^k\|_F^2.$$

Then, one can show that Proposition 5.2 and Theorem 5.3 remain valid. In addition, we compute the quantities $\|U\|_F^2$ and $\|Y^k\|_F^2$ by $\mathrm{tr}(U^\top U)$ and $\mathrm{tr}((Y^k)^\top Y^k)$, respectively. For some cases, the matrices $U^\top U$ and $(Y^k)^\top Y^k$ can be used repeatedly in updating the variables and evaluating the objective value and successive changes to reduce the cost of line search; see a concrete example in Section 6.1.

6. Numerical experiments.

In this section, we conduct numerical experiments to test our algorithm for NMF and MC on real datasets. All experiments are run in MATLAB R2015b on a 64-bit PC with an Intel Core i7-4790 CPU (3.60 GHz) and 32 GB of RAM equipped with Windows 10 OS.

6.1. Non-negative matrix factorization.

We first consider NMF

$$\min_{X,Y} \ \frac{1}{2}\left\|XY^\top - M\right\|_F^2 \quad \text{s.t.} \quad X \ge 0, \ \ Y \ge 0, \tag{6.1}$$

where $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{n\times r}$ are decision variables. Note that the feasible set of (6.1) is unbounded. We hence focus on the following model:

$$\min_{X,Y} \ \frac{1}{2}\left\|XY^\top - M\right\|_F^2 \quad \text{s.t.} \quad 0 \le X \le X^{\max}, \ 0 \le Y \le Y^{\max}, \tag{6.2}$$

where $X^{\max} \ge 0$ and $Y^{\max} \ge 0$ are upper bound matrices. One can show that, when $X_{ij}^{\max}$ and $Y_{ij}^{\max}$ are sufficiently large[3], solving (6.2) gives a solution of (6.1). In our

[3]The estimations of $X_{ij}^{\max}$ and $Y_{ij}^{\max}$ have been discussed in [9, Page 67].

experiments, for simplicity, we set $X^{\max}_{ij} = 10^{16}$ and $Y^{\max}_{ij} = 10^{16}$ for all (i,j). Now, we see that (6.2) corresponds to (1.1) with $\Psi(X) = \delta_{\mathcal{X}}(X)$, $\Phi(Y) = \delta_{\mathcal{Y}}(Y)$ and $\mathcal{A} = \mathcal{I}$, where $\mathcal{X} = \{X \in \mathbb{R}^{m\times r} : 0 \le X \le X^{\max}\}$ and $\mathcal{Y} = \{Y \in \mathbb{R}^{n\times r} : 0 \le Y \le Y^{\max}\}$. We apply NAUM to solving (6.2), and use (4.1c) and (4.2c) to update U and V. The specific updates of Z^k, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are

$$
\begin{aligned}
Z^k &= \tfrac{\alpha}{\alpha+\beta} X^k (Y^k)^\top + \tfrac{\beta}{\alpha+\beta} M, \\
\boldsymbol{u}_i &= \max\left\{0,\ \min\left\{\boldsymbol{x}_i^{\max},\ \frac{\alpha P_i^k \boldsymbol{y}_i^k + \mu_k \boldsymbol{x}_i^k}{\alpha\|\boldsymbol{y}_i^k\|^2 + \mu_k}\right\}\right\}, \quad i = 1, 2\cdots, r, \\
\boldsymbol{v}_i &= \max\left\{0,\ \min\left\{\boldsymbol{y}_i^{\max},\ \frac{\alpha (Q_i^k)^\top \boldsymbol{u}_i + \sigma_k \boldsymbol{y}_i^k}{\alpha\|\boldsymbol{u}_i\|^2 + \sigma_k}\right\}\right\}, \quad i = 1, 2\cdots, r,
\end{aligned}
$$

where P_i^k and Q_i^k are defined in (4.8). Note that here it is not necessary to update Z^k explicitly. Indeed, we can directly compute $P_i^k \boldsymbol{y}_i^k$ and $(Q_i^k)^\top \boldsymbol{u}_i$ by substituting Z^k as below:

$$
\begin{aligned}
P_i^k \boldsymbol{y}_i^k &= \tfrac{\alpha}{\alpha+\beta} X^k (Y^k)^\top \boldsymbol{y}_i^k + \tfrac{\beta}{\alpha+\beta} M \boldsymbol{y}_i^k - \textstyle\sum_{j=1}^{i-1} \boldsymbol{u}_j (\boldsymbol{y}_j^k)^\top \boldsymbol{y}_i^k - \textstyle\sum_{j=i+1}^{r} \boldsymbol{x}_j^k (\boldsymbol{y}_j^k)^\top \boldsymbol{y}_i^k, \\
(Q_i^k)^\top \boldsymbol{u}_i &= \tfrac{\alpha}{\alpha+\beta} Y^k (X^k)^\top \boldsymbol{u}_i + \tfrac{\beta}{\alpha+\beta} M^\top \boldsymbol{u}_i - \textstyle\sum_{j=1}^{i-1} \boldsymbol{v}_j \boldsymbol{u}_j^\top \boldsymbol{u}_i - \textstyle\sum_{j=i+1}^{r} \boldsymbol{y}_j^k \boldsymbol{u}_j^\top \boldsymbol{u}_i.
\end{aligned} \tag{6.3}
$$

When computing $X^k(Y^k)^\top \boldsymbol{y}_i^k$ and $Y^k(X^k)^\top \boldsymbol{u}_i$ in the above, we first compute $(Y^k)^\top \boldsymbol{y}_i^k$ and $(X^k)^\top \boldsymbol{u}_i$ to avoid forming the huge $(m \times n)$ matrix $X^k(Y^k)^\top$. Moreover, the matrices $(X^k)^\top U$, $U^\top U$, $(Y^k)^\top Y^k$ and $M^\top U$ that have been computed in (6.3) can be used again to evaluate the successive changes and the objective value as follows:

$$
\begin{aligned}
\|U - X^k\|_F^2 &= \mathrm{tr}(U^\top U) - 2\mathrm{tr}((X^k)^\top U) + \mathrm{tr}((X^k)^\top X^k), \\
\|V - Y^k\|_F^2 &= \mathrm{tr}(V^\top V) - 2\mathrm{tr}((Y^k)^\top V) + \mathrm{tr}((Y^k)^\top Y^k), \\
\|UV^\top - M\|_F^2 &= \mathrm{tr}((U^\top U)(V^\top V)) - 2\mathrm{tr}((M^\top U)V^\top) + \|M\|_F^2.
\end{aligned}
$$

In the above relations, $(X^k)^\top X^k$ and $(Y^k)^\top Y^k$ can be obtained from $U^\top U$ and $V^\top V$ in the previous iteration, respectively, and $\|M\|_F^2$ can be computed in advance. Additionally, as we discussed in Remark 5.3, $\mathrm{tr}((Y^k)^\top Y^k)$ and $\mathrm{tr}(U^\top U)$ can also be used in computing $\mu_k^{\max}$ and $\sigma_k^{\max}$, respectively. These techniques were also used in many popular algorithms for NMF to reduce the computational cost (see, for example, [2, 9, 10, 18, 37]).

The experiments are conducted on the face datasets (dense matrices) and the text datasets (sparse matrices). For face datasets, we use CBCL[4], ORL[5] [29] and the extended Yale Face Database B (e-YaleB)[6] [19] for our test. CBCL contains 2429 images of faces with 19×19 pixels, ORL contains 400 images of faces with 112×92 pixels, and e-YaleB contains 2414 images of faces with 168×192 pixels. In our experiments, for each face dataset, each image is vectorized and stacked as a column of a data matrix M of size $m \times n$. For text datasets, we use three datasets from the CLUTO toolkit[7]. The specific values of m and n for each dataset and the values of r used for our tests are summarized in Table 1.

The parameters in NAUM are set as follows: $\mu^{\min} = \bar{\mu}_{-1} = 1$, $\sigma^{\min} = \bar{\sigma}_{-1} = 1$, $\sigma^{\max} = 10^6$, $\tau = 4$, $c = 10^{-4}$, $N = 3$, $\mu_k^0 = \max\left\{0.1\bar{\mu}_{k-1},\, \mu^{\min}\right\}$ and $\sigma_k^0 =$

[4]Available in http://cbcl.mit.edu/cbcl/software-datasets/FaceData2.html.
[5]Available in http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.
[6]Available in http://vision.ucsd.edu/~iskwak/ExtYaleDatabase/ExtYaleB.html.
[7]Available in http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download.

TABLE 1
Real data sets

Face Datasets (dense matrices)					Text Datasets (sparse matrices)				
Data	Pixels	m	n	r	Data	Sparsity	m	n	r
CBCL	19×19	361	2429	30, 60	classic	99.92%	7094	41681	10, 20
ORL	112×92	10304	400	30, 60	sports	99.14%	8580	14870	10, 20
e-YaleB	168×192	32256	2414	30, 60	ohscal	99.47%	11162	11465	10, 20

$\min\left\{\max\left\{0.1\bar{\sigma}_{k-1},\, \sigma^{\min}\right\},\, \sigma^{\max}\right\}$ for any $k \geq 0$. Moreover, we set $\beta = \frac{\alpha}{\alpha-1}$, $\gamma = \max\{0,\, -\alpha,\, -(\alpha+\beta)\}$ and $\rho = \max\left\{1,\, \alpha^2/(\alpha+\beta)^2\right\}$ for some given α.

We then compare the performances of NAUM with different α. In our comparisons, we initialize NAUM with different α at the same random initialization $(X^0,\, Y^0)$[8] and terminate them if one of the following stopping criteria is satisfied:

- $\frac{|\mathcal{F}^k_{\mathrm{nmf}} - \mathcal{F}^{k-1}_{\mathrm{nmf}}|}{\mathcal{F}^k_{\mathrm{nmf}}+1} \leq 10^{-4}$ holds for 3 consecutive iterations;
- $\frac{\|X^k - X^{k-1}\|_F + \|Y^k - Y^{k-1}\|_F}{\|X^k\|_F + \|Y^k\|_F + 1} \leq 10^{-4}$ holds,

where $\mathcal{F}^k_{\mathrm{nmf}} := \frac{1}{2}\left\|X^k(Y^k)^\top - M\right\|_F^2$ denotes the objective value at (X^k, Y^k). Table 2 presents the results of NAUM with different α for two face datasets (CBCL and ORL) and $r = 30, 60$. In the table, "iter" denotes the number of iterations; "relerr" denotes the relative error $\frac{\|X^*(Y^*)^\top - M\|_F}{\|M\|_F}$, where $(X^*,\, Y^*)$ is a terminating point obtained by each NUAM in a trial; "time" denotes the computational time (in seconds). All the results presented are the average of 10 independent trials. From Table 2, we can see that NAUM with a relatively small α (e.g., 0.6 and 0.8) has better numerical performance. However, α cannot be too small. Observe that NAUM with $\alpha = 0.5, 0.4, 0.2$ are not competitive and, surprisingly, $\alpha = 0.5$ leads to the worst performance. In view of this, we do not choose $\alpha < 0.6$ in our following experiments for NMF.

TABLE 2
Comparisons of NAUM with different α

α	iter	relerr	time	α	iter	relerr	time
CBCL, $r = 30$				CBCL, $r = 60$			
2.0	488	1.0519e-01	1.72	2.0	626	7.4388e-02	4.94
1.1	381	1.0448e-01	1.35	1.1	555	7.3477e-02	4.38
0.8	315	1.0426e-01	1.09	0.8	511	7.2986e-02	4.09
0.6	268	1.0406e-01	0.94	0.6	419	7.2998e-02	3.32
0.5	833	1.0593e-01	4.74	0.5	1372	7.5864e-02	19.49
0.4	440	1.0489e-01	3.05	0.4	599	7.4568e-02	10.02
0.2	556	1.0674e-01	4.18	0.2	782	7.7654e-02	14.30
ORL, $r = 30$				ORL, $r = 60$			
2.0	232	1.6673e-01	3.45	2.0	277	1.4078e-01	7.92
1.1	188	1.6619e-01	2.78	1.1	210	1.4042e-01	6.04
0.8	158	1.6603e-01	2.33	0.8	182	1.4017e-01	5.20
0.6	132	1.6578e-01	2.01	0.6	156	1.3996e-01	4.44
0.5	652	1.7216e-01	15.79	0.5	695	1.4583e-01	32.91
0.4	280	1.6615e-01	7.55	0.4	353	1.4061e-01	19.17
0.2	307	1.6753e-01	8.71	0.2	358	1.4272e-01	20.77

We next compare NAUM with two existing efficient algorithms[9] for NMF: the

[8]We use the Matlab commands: `X0 = max(0, randn(m, r)); Y0 = max(0, randn(n, r)); X0 = X0/norm(X0,'fro')*sqrt(norm(M, 'fro')); Y0 = Y0/norm(Y0,'fro')*sqrt(norm(M, 'fro'));`

[9]Most existing algorithms are directly developed for (6.1). However, they need the assumption that the sequence generated is bounded in their convergence analysis. Although this assumption is uncheckable and may fail, these algorithms always work well in practice. Thus, we directly use these algorithms in our comparisons, rather than modifying them for (6.2).

hierarchical alternating least squares (HALS) method[10] (see, for example, [8, 9, 10, 11, 20, 21]) and the block coordinate descent method for NMF (BCD-NMF[11]) (see Algorithm 2 in Section 3.2 in [40]).

To better evaluate the performances of different algorithms, we follow [11] to use an evolution of the objective function value. To define this evolution, we first define

$$e(k) := \frac{\mathcal{F}^k - \mathcal{F}_{\min}}{\mathcal{F}^0 - \mathcal{F}_{\min}},$$

where $\mathcal{F}^k$ denotes the objective function value obtained by an algorithm at (X^k, Y^k) and $\mathcal{F}_{\min}$ denotes the minimum of the objective function values obtained among *all* algorithms across *all* initializations. We also use $\mathcal{T}(k)$ to denote the total computational time after completing the k-th iteration of an algorithm. Thus, $\mathcal{T}(0) = 0$ and $\mathcal{T}(k)$ is non-decreasing with respect to k. Then, the evolution of the function value obtained from a particular algorithm with respect to time t is defined as

$$E(t) := \min\left\{e(k) : k \in \{i : \mathcal{T}(i) \leq t\}\right\}.$$

One can see that $0 \leq E(t) \leq 1$ (since $0 \leq e(k) \leq 1$ for all k) and $E(t)$ is non-increasing with respect to t. $E(t)$ can be considered as a normalized measure of the reduction of the function value with respect to time. For a given matrix M and a positive integer r, one can take the average of $E(t)$ over several independent trials with different initializations, and plot the average $E(t)$ within time t for a given algorithm.

In our experiments, we initialize all the algorithms at the same random initial point (X^0, Y^0) and terminate them *only* by the maximum running time $\mathrm{T}^{\max}$. The specific values of $\mathrm{T}^{\max}$ are given in Fig. 1 and Fig. 2. Additionally, we use the default settings for BCD-NMF. For NAUM, we choose $\alpha = 0.6, 0.8, 1.1, 2$. We then plot the average $E(t)$ for each algorithm within time $\mathrm{T}^{\max}$.

Fig. 1 and Fig. 2 show the average $E(t)$ of 30 independent trials for NMF on face datasets and text datasets, respectively. From the results, we can see that NAUM with $\alpha = 0.6$ performs best in most cases, and NAUM with $\alpha = 0.6$ or 0.8 always performs better than NAUM with $\alpha > 1$. This shows that choosing α and β under the weaker condition $\frac{1}{\alpha} + \frac{1}{\beta} = 1$ (hence α can be small than 1) can improve the numerical performance of NAUM.

6.2. Matrix completion. We next consider a recent model for MC:

$$\min_{X,Y} \;\; \frac{\eta}{2}\|X\|_* + \frac{\eta}{2}\|Y\|_* + \frac{1}{2}\left\|\mathcal{P}_\Omega(XY^\top - M)\right\|_F^2, \tag{6.4}$$

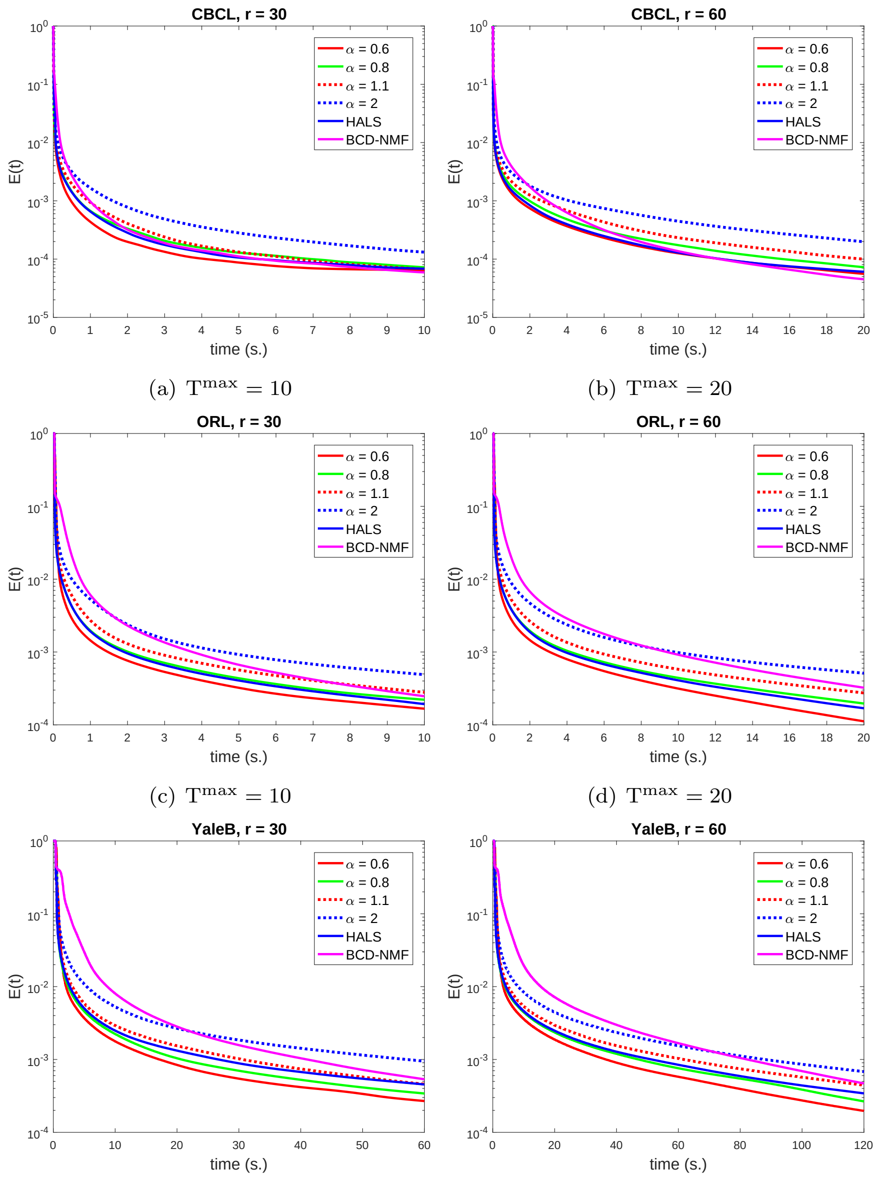
where $\eta > 0$ is a penalty parameter, Ω is the index set of the known entries of M, and $\mathcal{P}_\Omega(Z)$ keeps the entries of Z in Ω and sets the remaining ones to zero. This model was first considered in [30, 31] and was shown to be equivalent to Schatten-$\frac{1}{2}$ quasi-norm minimization. Encouraging numerical performance of this model has also

[10] HALS for (6.1) is given by

$$\boldsymbol{x}_i^{k+1} = \max\left\{0,\; \frac{M\boldsymbol{y}_i^k - \sum_{j=1}^{i-1}\boldsymbol{x}_j^{k+1}(\boldsymbol{y}_j^k)^\top\boldsymbol{y}_i^k - \sum_{j=i+1}^{r}\boldsymbol{x}_j^{k}(\boldsymbol{y}_j^k)^\top\boldsymbol{y}_i^k}{\|\boldsymbol{y}_i^k\|^2}\right\}, \quad i = 1, \cdots, r,$$

$$\boldsymbol{y}_i^{k+1} = \max\left\{0,\; \frac{M^\top\boldsymbol{x}_i^{k+1} - \sum_{j=1}^{i-1}\boldsymbol{y}_j^{k+1}(\boldsymbol{x}_j^{k+1})^\top\boldsymbol{x}_i^{k+1} - \sum_{j=i+1}^{r}\boldsymbol{y}_j^{k}(\boldsymbol{x}_j^{k+1})^\top\boldsymbol{x}_i^{k+1}}{\|\boldsymbol{x}_i^{k+1}\|^2}\right\}, \quad i = 1, \cdots, r.$$

[11] Available at http://www.math.ucla.edu/~wotaoyin/papers/bcu/nmf/index.html.



(a) $\mathrm{T}^{\max} = 10$ (b) $\mathrm{T}^{\max} = 20$

(c) $\mathrm{T}^{\max} = 10$ (d) $\mathrm{T}^{\max} = 20$

(e) $\mathrm{T}^{\max} = 60$ (f) $\mathrm{T}^{\max} = 120$

FIG. 1. *Average $E(t)$ of 30 independent trials for NMF on face datasets.*

been reported in [30, 31]. Note that (6.4) corresponds to (1.1) with $\Psi(X) = \frac{\eta}{2}\|X\|_*$, $\Phi(Y) = \frac{\eta}{2}\|Y\|_*$ and $\mathcal{A} = \mathcal{P}_\Omega$. Thus, we can apply NAUM with (4.1b) and (4.2b) to

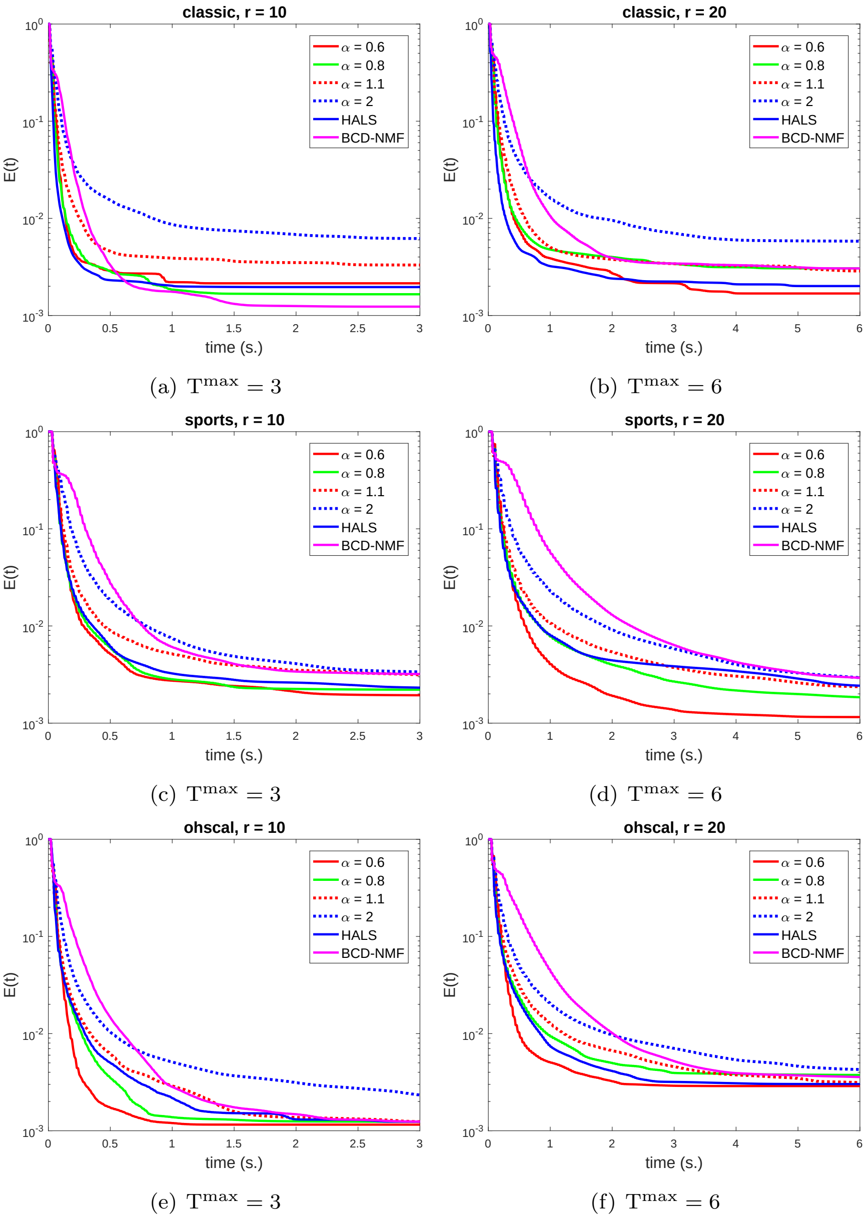


FIG. 2. *Average $E(t)$ of 30 independent trials for NMF on text datasets.*

solving (6.4). The updates of Z^k, U and V are

$$
\begin{aligned}
Z^k &= X^k(Y^k)^\top + \tfrac{\beta}{\alpha+\beta}\mathcal{P}_\Omega\left(M - X^k(Y^k)^\top\right),\\
U &= \mathcal{S}_{\eta/(2\mu_k)}\left(X^k - \tfrac{\alpha}{\mu_k}(X^k(Y^k)^\top - Z^k)Y^k\right),\\
V &= \mathcal{S}_{\eta/(2\sigma_k)}\left(Y^k - \tfrac{\alpha}{\sigma_k}(U(Y^k)^\top - Z^k)^\top U\right).
\end{aligned}
$$

Substituting Z^k into U and V and using $\frac{1}{\alpha}+\frac{1}{\beta}=1$ gives

$$
\begin{aligned}
U &= \mathcal{S}_{\eta/(2\mu_k)}\left(X^k - \tfrac{1}{\mu_k}\left[\mathcal{P}_\Omega(X^k(Y^k)^\top - M)\right]Y^k\right),\\
V &= \mathcal{S}_{\eta/(2\sigma_k)}\left(Y^k - \tfrac{\alpha}{\sigma_k}Y^k(U-X^k)^\top U - \tfrac{1}{\sigma_k}\left[\mathcal{P}_\Omega(X^k(Y^k)^\top - M)\right]^\top U\right).
\end{aligned}
\tag{6.5}
$$

Thus, similar to NAUM for NMF, we do not need to update Z^k explicitly for MC.

We compare NAUM with proximal alternating linearized minimization (PALM), which was proposed in [4] and was used to solve (6.4) in [30, 31]. For ease of future reference, we recall that the PALM for solving (6.4) is given by

$$
\begin{aligned}
X^{k+1} &= \mathcal{S}_{\frac{\eta}{2\|Y^k\|^2}}\left(X^k - \tfrac{1}{\|Y^k\|^2}\left[\mathcal{P}_\Omega(X^k(Y^k)^\top - M)\right]Y^k\right),\\
Y^{k+1} &= \mathcal{S}_{\frac{\eta}{2\|X^{k+1}\|^2}}\left(Y^k - \tfrac{1}{\|X^{k+1}\|^2}\left[\mathcal{P}_\Omega(X^{k+1}(Y^k)^\top - M)\right]^\top X^{k+1}\right).
\end{aligned}
$$

For NAUM, we use the same parameter settings as in Section 6.1, but choose $\alpha = 0.4, 0.6, 1.1$. All the algorithms are initialized at the same random initialization (X^0, Y^0)[12] and terminated if one of the following stopping criteria is satisfied:

- $\frac{|\mathcal{F}_{\mathrm{mc}}^k - \mathcal{F}_{\mathrm{mc}}^{k-1}|}{\mathcal{F}_{\mathrm{mc}}^k + 1} \leq 10^{-4}$ holds for 3 consecutive iterations;
- $\frac{\|X^k - X^{k-1}\|_F + \|Y^k - Y^{k-1}\|_F}{\|X^k\|_F + \|Y^k\|_F + 1} \leq 10^{-4}$ holds;
- the running time is more than 300 seconds,

where $\mathcal{F}_{\mathrm{mc}}^k := \frac{\eta}{2}\|X^k\|_* + \frac{\eta}{2}\|Y^k\|_* + \frac{1}{2}\left\|\mathcal{P}_\Omega(X^k(Y^k)^\top - M)\right\|_F^2$ denotes the objective function value obtained by each algorithm at (X^k, Y^k).

Table 3 presents the numerical results of different algorithms for different problems, where two face datasets (CBCL and ORL) are used as our test matrices M and a subset Ω of entries is sampled uniformly at random. In the table, sr denotes the sampling ratio, i.e., a subset Ω of (rounded) $mn * sr$ entries is sampled; r denotes the rank used for test; "iter" denotes the number of iterations; "Normalized fval" denotes the normalized function value $\frac{\mathcal{F}(X^*, Y^*) - \mathcal{F}_{\min}}{\mathcal{F}_{\max} - \mathcal{F}_{\min}}$, where (X^*, Y^*) is obtained by each algorithm, $\mathcal{F}(X^*, Y^*)$ is the function value at (X^*, Y^*) for each algorithm and $\mathcal{F}_{\max}$ (resp. $\mathcal{F}_{\min}$) denotes the maximum (resp. minimum) of the terminating function values obtained from *all* algorithms in *a* trial (one random initialization and Ω); "RecErr" denotes the recovery error $\frac{\|X^*(Y^*)^\top - M\|_F}{\|M\|_F}$. All the results presented are the average of 10 independent trials.

From Table 3, we can see that NAUM with $\alpha = 0.4$ gives the smallest function values and the smallest recovery error within least CPU time in most cases. Moreover, NAUM with $\alpha = 0.6$ also performs better than NAUM with $\alpha = 1.1$ and PALM with respect to the function value and the recovery error in most cases. This again shows that a flexible choice of α and β can lead to better numerical performances and the choice of $\alpha = 0.4$ performs best for MC from our experiments.

7. Concluding remarks. In this paper, we consider a class of matrix factorization problems involving two blocks of variables. To solve this kind of possibly nonconvex, nonsmooth and non-Lipschitz problems, we introduce a specially constructed potential function $\Theta_{\alpha,\beta}$ defined in (1.2) which contains one auxiliary block of variables. We then develop a non-monotone alternating updating method with a suitable line search criterion based on this potential function. Unlike other existing

[12] We use the Matlab commands: `X0 = randn(m, r); Y0 = randn(n, r);`

TABLE 3
Numerical results for MC on face datasets

η	data	sr	r	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 1.1$	PALM	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 1.1$	PALM
				iter				Normalized fval			
5	CBCL	0.5	30	780	1189	3320	3306	1.13e-01	**7.50e-02**	4.52e-01	1
		0.5	60	921	1218	3850	4654	**3.24e-02**	5.10e-02	3.85e-01	1
		0.2	30	1174	2366	4767	3573	**8.01e-03**	2.21e-01	6.87e-01	9.60e-01
		0.2	60	1577	1919	5360	5037	**1.03e-02**	8.95e-02	8.08e-01	8.86e-01
	ORL	0.5	30	1218	1243	1241	1468	**0**	2.94e-01	5.06e-01	1
		0.5	60	1049	1051	1051	1327	**0**	1	4.00e-01	7.73e-01
		0.2	30	2074	325	385	2691	**2.59e-03**	7.01e-01	1	1.31e-01
		0.2	60	1551	1551	356	2222	**0**	3.82e-01	1	2.12e-01
10	CBCL	0.5	30	457	654	1793	1935	**2.20e-02**	1.29e-01	3.60e-01	9.81e-01
		0.5	60	514	594	1950	2559	2.65e-01	**1.15e-01**	3.79e-01	8.71e-01
		0.2	30	627	1313	2513	2116	**1.91e-02**	3.75e-02	8.35e-01	7.79e-01
		0.2	60	866	1095	2713	2889	**2.07e-02**	2.89e-02	9.22e-01	4.86e-01
	ORL	0.5	30	1003	1186	1192	1402	**3.30e-02**	1.47e-01	4.30e-01	1
		0.5	60	975	1009	1012	1276	**0**	8.58e-01	6.11e-01	9.99e-01
		0.2	30	1409	364	411	2646	**0**	7.16e-01	1	8.10e-02
		0.2	60	1241	1504	376	2185	**4.05e-06**	3.97e-02	1	2.21e-01
				CPU time				RecErr			
5	CBCL	0.5	30	**35.56**	54.14	151.23	119.05	**1.05e-01**	**1.05e-01**	1.06e-01	1.08e-01
		0.5	60	**57.66**	76.09	240.19	206.47	**8.81e-02**	9.02e-02	9.04e-02	8.99e-02
		0.2	30	**34.04**	68.57	137.97	75.56	**1.37e-01**	**1.37e-01**	1.38e-01	1.43e-01
		0.2	60	**72.01**	87.82	245.21	147.08	**1.34e-01**	1.35e-01	1.35e-01	1.36e-01
	ORL	0.5	30	**294.20**	300	300	300	**1.72e-01**	1.84e-01	2.01e-01	2.12e-01
		0.5	60	300	300	300	300	**1.66e-01**	2.11e-01	2.05e-01	2.11e-01
		0.2	30	300	**47.35**	55.86	300	**2.08e-01**	3.04e-01	3.81e-01	2.24e-01
		0.2	60	300	300	**69.21**	300	**2.16e-01**	2.35e-01	3.49e-01	2.61e-01
10	CBCL	0.5	30	**21.01**	30.12	82.45	70.32	**1.16e-01**	1.19e-01	1.18e-01	1.17e-01
		0.5	60	**32.40**	37.38	122.51	113.80	**1.09e-01**	1.11e-01	1.14e-01	1.11e-01
		0.2	30	**18.15**	38.01	72.84	44.62	**1.60e-01**	1.61e-01	1.62e-01	**1.60e-01**
		0.2	60	**39.13**	49.37	123.74	83.52	1.57e-01	1.57e-01	1.58e-01	**1.56e-01**
	ORL	0.5	30	**252.15**	300	300	300	**1.71e-01**	1.77e-01	1.95e-01	2.08e-01
		0.5	60	**289.57**	300	300	300	**1.53e-01**	2.01e-01	2.03e-01	2.09e-01
		0.2	30	207.22	**53.08**	60.54	300	**1.95e-01**	3.06e-01	3.83e-01	2.14e-01
		0.2	60	243.45	295.60	**74.09**	300	**1.87e-01**	1.95e-01	3.60e-01	2.36e-01

methods such as those based on alternating minimization, our method essentially updates the two blocks of variables alternately by solving subproblems related to $\Theta_{\alpha,\beta}$ and then updates the auxiliary block of variables by an explicit formula (see (4.5)). Using the special structure of $\Theta_{\alpha,\beta}$, we demonstrate how some efficient computational strategies for NMF can be used to solve the associated subproblems in our method. Moreover, under some mild conditions, we establish that the sequence generated by our method is bounded and any cluster point of the sequence gives a stationary point of our problem. Finally, we conduct some numerical experiments for NMF and MC on real datasets to illustrate the efficiency of our method.

Note that the parameter α (and $\beta = \alpha/(\alpha-1)$) plays a significant role in our NAUM. Although it has been observed in our experiments that a relatively small α (e.g., 0.6, 0.8) can improve the numerical performance of NAUM, how to choose an optimal α is still unknown. In view of the recent work [24] on adaptively choosing the extrapolation parameter in FISTA for solving a class of possibly nonconvex problems, it may be possible to derive a strategy to adaptively update α in our NAUM. This is a possible future research topic.

Acknowledgments. The authors are grateful to Nicolas Gillis for his helpful comments. The authors are also grateful to the editor and the anonymous referees for their valuable suggestions and comments, which helped improve this paper.

REFERENCES

[1] H. Attouch, J. Bolte, and B. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods*, Mathematical Programming, 137 (2013), pp. 91–129.
[2] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons, *Algorithms and applications for approximate nonnegative matrix factorization*, Computational Statistics & Data Analysis, 52 (2007), pp. 155–173.
[3] P. Biswas, T. Liang, K. Toh, Y. Ye, and T. Wang, *Semidefinite programming approaches for sensor network localization with noisy distance measurements*, IEEE Transactions on Automation Science and Engineering, 3 (2006), pp. 360–371.
[4] J. Bolte, S. Sabach, and M. Teboublle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.
[5] E. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics, 9 (2009), pp. 717–772.
[6] E. Candès and T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Transactions on Information Theory, 56 (2010), pp. 2053–2080.
[7] X. Chen, Z. Lu, and T. Pong, *Penalty methods for a class of non-Lipschitz optimization problems*, SIAM Journal on Optimization, 26 (2016), pp. 1465–1492.
[8] A. Cichocki, R. Zdunek, and S. Amari, *Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization*, in International Conference on Independent Component Analysis and Signal Separation, Springer, 2007, pp. 169–176.
[9] N. Gillis, *Nonnegative Matrix Factorization: Complexity, Algorithms and Applications*, PhD thesis, Université catholique de Louvain, 2011.
[10] N. Gillis, *The why and how of nonnegative matrix factorization*, in Regularization, Optimization, Kernels, and Support Vector Machines, Chapman and Hall/CRC, 2014, pp. 257–291.
[11] N. Gillis and F. Glineur, *Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization*, Neural Computation, 24 (2012), pp. 1085–1105.
[12] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, *A general iterative shinkage and thresholding algorithm for non-convex regularized optimization problems*, in Proceedings of the International Conference on Machine Learning, vol. 28, 2013, pp. 37–45.
[13] M. Hong, *Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence and applications*, arXiv preprint arXiv:1604.00543, (2016).
[14] B. Jiang, T. Lin, S. Ma, and S. Zhang, *Structured nonconvex and nonsmooth optimization: Algorithms and iteration complexity analysis*, arXiv preprint arXiv: 1605.02408, (2016).
[15] R. Keshavan, A. Montanari, and S. Oh, *Matrix completion from a few entries*, IEEE Transactions on Information Theory, 56 (2010), pp. 2980–2998.
[16] M.-J. Lai, Y. Xu, and W. Yin, *Improved iteratively rewighted least squares for unconstrained smoothed ℓ_p minimization*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 927–957.
[17] D. Lee and H. Seung, *Learning the parts of objects by nonnegative matrix factorization*, Nature, 401 (1999), pp. 788–791.
[18] D. Lee and H. Seung, *Algorithms for non-negative matrix factorization*, in Proceedings of NIPS, 2001, pp. 556–562.
[19] K. Lee, J. Ho, and D. Kriegman, *Acquiring linear subspaces for face recognition under variable lighting*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (2005), pp. 684–698.
[20] L. Li and Y. Zhang, *FastNMF: Highly efficient monotonic fixed-point nonnegative matrix factorization algorithm with good applicability*, Journal of Electronic Imaging, 18 (2009), p. 033004.
[21] J. Liu, J. Liu, P. Wonka, and J. Ye, *Sparse non-negative tensor factorization using columnwise coordinate descent*, Pattern Recognition, 45 (2012), pp. 649–656.
[22] Z. Liu and L. Vandenberghe, *Interior-point method for nuclear norm approximation with application to system identification*, SIAM Journal on Matrix Analysis and Applications, 31 (2009), pp. 1235–1256.
[23] K. Mohan and M. Fazel, *Iterative reweighted algorithms for matrix rank minimization*, Journal of Machine Learning Researc, 13 (2012), pp. 3441–3473.
[24] P. Ochs and T. Pock, *Adaptive FISTA*, arXiv preprint arXiv:1711.04343, (2017).
[25] P. Paatero and U. Tapper, *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics, 5 (1994), pp. 111–126.
[26] B. Recht, M. Fazel, and P. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Review, 52 (2010), pp. 471–501.
[27] J. Rennie and N. Srebro, *Fast maximum margin matrix factorization for collaborative prediction*, in Proceedings of the 22nd international conference on Machine learning, ACM,

2005, pp. 713–719.
[28] R. Rockafellar and R.-B. Wets, *Variational Analysis*, Springer, 1998.
[29] F. Samaria and A. Harter, *Parameterisation of a stochastic model for human face identification*, in Proceedings of the Second IEEE Workshop on Applications of Computer Vision, IEEE, 1994, pp. 138–142.
[30] F. Shang, Y. Liu, and J. Cheng, *Scalable algorithms for tractable Schatten quasi-norm minimization*, in Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016, pp. 2016–2022.
[31] F. Shang, Y. Liu, and J. Cheng, *Tractable and scalable Schatten quasi-norm approximations for rank minimization*, in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, 2016, pp. 620–629.
[32] F. Shang, Y. Liu, and J. Cheng, *Unified scalable equivalent formulations for Schatten quasi-norms*, arXiv preprint arXiv: 1606.00668, (2016).
[33] N. Srebro, *Learning with Matrix Factorizations*, PhD thesis, Massachusetts Institute of Technology, 2004.
[34] R. Sun and Z.-Q. Luo, *Guaranteed matrix completion via non-convex factorization*, IEEE Transactions on Information Theory, 62 (2016), pp. 6535–6579.
[35] M. Udell, C. Horn, R. Zadeh, and S. Boyd, *Generalized low rank models*, Foundations and Trends in Machine Learning, 9 (2016), pp. 1–118.
[36] S. Vavasis, *On the complexity of nonnegative matrix factorization*, SIAM Journal on Optimization, 20 (2009), pp. 1364–1377.
[37] Y. Wang and Y. Zhang, *Nonnegative matrix factorization: A comprehensive review*, IEEE Transactions on Knowledge and Data Engineering, 25 (2013), pp. 1336–1353.
[38] Z. Wen, W. Yin, and Y. Zhang, *Solving a low-rank factorization model for matrix completion by a non-linear successive over-relaxation algorithm*, Mathematical Programming Computation, 4 (2012), pp. 333–361.
[39] S. Wright, R. Nowak, and M. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Transactions on Signal Processing, 57 (2009), pp. 2479–2493.
[40] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1758–1789.
[41] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, *An alternating direction algorithm for matrix completion with nonnegative factors*, Frontiers of Mathematics in China, 7 (2012), pp. 365–384.
[42] M. Zhang, Z.-H. Huang, and Y. Zhang, *Restricted p-isometry properties of nonconvex matrix recovery*, IEEE Transactions on Information Theory, 59 (2013), pp. 4316–4323.
[43] Y. Zhang, *An alternating direction algorithm for nonnegative matrix factorization*, Preprint, (2010).
[44] X. Zhou, C. Yang, H. Zhao, and W. Yu, *Low-rank modeling and its applications in image analysis*, ACM Computing Survey, 47 (2015), p. 36.