Alternating Block Linearized Bregman Iterations for Regularized Nonnegative Matrix Factorization

Beier Chen[1] and Hui Zhang[1]

[1]National University of Defense Technology

June 2, 2023

Abstract

In this paper, we propose an alternating block variant of the linearized Bregman iterations for a class of regularized nonnegative matrix factorization problems (NMF). The proposed method exploits the block structure of NMF, utilizes the smooth adaptable property of the loss function based on the Bregman distance, and at the same time follows the iterative regularization idea of the linearized Bregman iterations method. Theoretically, we show that the proposed method is a descent method by adjusting the involved parameters. Finally, we end with several illustrative numerical experiments.

Alternating Block Linearized Bregman Iterations for Regularized Nonnegative Matrix Factorization

Beier Chen * Hui Zhang †

June 1, 2023

Abstract

In this paper, we propose an alternating block variant of the linearized Bregman iterations for a class of regularized nonnegative matrix factorization problems (NMF). The proposed method exploits the block structure of NMF, utilizes the smooth adaptable property of the loss function based on the Bregman distance, and at the same time follows the iterative regularization idea of the linearized Bregman iterations method. Theoretically, we show that the proposed method is a descent method by adjusting the involved parameters. Finally, we end with several illustrative numerical experiments.

Keywords. nonnegative matrix factorization, sparse regularization, Bregman distance, linearized Bregman iterations, sufficient descent, alternating block.
AMS subject classifications. 90C30, 49M37, 65K10

1 Introduction

Nonnegative matrix factorization (NMF) is a dimensionality reduction technique widely used in machine learning, text mining, and image analysis. The objective of NMF is to factorize a nonnegative matrix $A \in \mathbb{R}^{m\times n}$ into two nonnegative matrices $X \in \mathbb{R}_+^{m\times r}$ and $Y \in \mathbb{R}_+^{n\times r}$ such that their product XY^T approximates A, where $r < \min\{m, n\}$. This problem is typically formulated as a non-convex optimization problem:

$$\min_{X\in\mathbb{R}_+^{m\times r}, Y\in\mathbb{R}_+^{n\times r}} \left\{ \Psi(X,Y) \equiv \frac{1}{2}\|A - XY^T\|_F^2 + \mathcal{R}_1(X) + \mathcal{R}_2(Y) \right\}, \tag{1.1}$$

where $\mathcal{R}_1$ and $\mathcal{R}_2$ are regularization terms that impose certain constraints or biases on the factor matrices X and Y, including sparsity and smoothness.

In the literature, there have developed many numerical methods to solve problem (1.1) with $\mathcal{R}_1 = \mathcal{R}_2 \equiv 0$, such as multiplicative updates (MU) [5, 6], projected gradient [7], block coordinate descent method [11] and hierarchical alternating least squares (HALS) [4]. All these methods can be categorized under alternating minimization, as at each iteration X is updated while Y is

*Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China. Email: chenbeier18@nudt.edu.cn

†Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China. Email: h.zhang1984@163.com

fixed, otherwise updating Y and fixing X. It is worth mentioning that any convergent subsequence generated by most of these approaches is guaranteed to converge to a critical point of NMF problem. However, these methods cannot be applied to solve the NMF problem with regularization (1.1).

In order to take the regularized terms into account, the author of [2] proposed the proximal alternating linearized minimization (PALM). Later on, an inertial variant, called iPALM was proposed in [9]. These two methods were built on the basic assumption that the loss function $\frac{1}{2}\|A - XY^T\|_F^2$ that satisfies the same gradient Lipschitz type continuity properties. In order to relax such assumptions, the authors of [10] proposed the Bregman proximal gradient method for NMF by following the concept of relative smoothness. Furthermore, the block Bregman proximal gradient method (BBPG) considers the blockwise variant of problem (1.1), which makes them easily amenable to parallel computation [10].

As a recently developed regularization technique, the linearized Bregman iterations method and its variants have been widely used in compressed sensing and image processing [8, 12, 13]. In this paper, we try to combine the LBreI and BBPG to develop a new method for NMF (1.1). To this end, we consider the following optimization problem which is more general than (1.1) and is similarly to that is focused on the reference [2]:

$$\inf\left\{\psi(x,y) \equiv E(x,y) + f(x) + g(y) : x \in \mathbb{R}^M, y \in \mathbb{R}^N\right\}, \tag{$\mathcal{P}$}$$

where f and g are convex extended real-valued functions, and E is a possibly non-convex function.

The remainder of the paper is organized as follows. Some preliminaries are given in Section 2. We propose an alternating block variant of the linearized Bregman iterations method for solving problem ($\mathcal{P}$) in Section 3. The convergence analysis is presented in Section 4. Moreover, the application to NMF with regularization is discussed in Section 5. Finally, we end with several illustrating numerical experiments in Section 6.

2 Preliminaries

Throughout the paper, we assume that $\langle\cdot,\cdot\rangle$ is the inner product and $\|\cdot\|$ is the induced norm. Let f be a real convex function, the domain (gradient of f and subgradient of f) is denoted by dom f. We begin with the definition of kernel generating distance, of which more details can be found in [3].

Definition 2.1 (Kernel generating distance)**.** *Let C be a nonempty, convex and open subset of $\mathbb{R}^n$. A function $h : C \to \mathbb{R}$ is called a kernel generating distance if it is continuously differentiable and satisfies the following conditions:*

(i) h is proper, lower semicontinuous and convex with dom $h \subset \bar{C}$ and dom $\partial h = C$;

(ii) h is differentiable on int dom h.

The proximity measure D_h is defined by setting a kernel generating distance $h : \mathbb{R}^n \to \mathbb{R}$. Given $u, v \in \mathbb{R}^n$, the Bregman distance between them can be expressed as

$$D_h(u,v) \equiv h(u) - [h(v) + \langle\nabla h(v), u - v\rangle].$$

This distance measure satisfies the property that $D_h(u,v) \geq 0$ if and only if h is convex. Furthermore, $D_h(u,v) = 0$ only when $u = v$ under strict convexity conditions for h. The literature [3] provides various examples of kernels h capable of generating Bregman distances.

This generalized Bregman distance of a proper lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R}$ between u and v is defined as:

$$D_f^p(u, v) \equiv f(u) - [f(v) + \langle p, u - v\rangle],$$

where p is a subgradient of $f(v)$, i.e., $p \in \partial f(v)$.

For simplicity, we define the symmetric generalized Bregman distance as follows.

Definition 2.2 (Symmetric generalized Bregman distance [14]). *$D_f^{symm}(u, v)$ is called the symmetric generalized Bregman distance of f between u and v, if*

$$D_f^{symm}(u, v) := D_f^q(u, v) + D_f^p(v, u) = \langle p - q, u - v\rangle,$$

for $u, v \in$ dom R with $p \in \partial f(u)$ and $q \in \partial f(v)$.

Let kernel generating distance h be defined as Definition 2.1, for all $w \in \text{dom } h$ and $u, v \in \text{int dom } h$, the wellknown three-points identity is expressed as

$$D_h(w, u) - D_h(w, v) - D_h(v, u) = \langle \nabla h(u) - h(v), v - w\rangle. \tag{2.1}$$

In addition, the next definition introduce a measure for the lack of symmetry in D_h.

Definition 2.3 (Symmetry coefficient [1]). *Given a kernel generating distance $h : \mathbb{R}^n \to \mathbb{R}$, its symmetry coefficient is defined by*

$$\eta(h) \equiv \inf\left\{\frac{D_h(u, v)}{D_h(v, u)} : u, v \in \mathbb{R}^n, u \neq v\right\} \in [0, 1]. \tag{2.2}$$

A commonly used method for analyzing NMF problems assumes that the gradient of function E is Lipschitz continuous. However, this paper introduces a simplified version of the L-smooth adaptable functions and refers to a more general definition outlined in [3]. This approach enhances the analysis of NMF problems and offers a general perspective into the optimization process.

Definition 2.4 (L-smooth adaptability). *Let $E, h : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable functions, and assume that h is convex. Then, we say that (E, h) is L-smooth adaptable if there exists some $L > 0$ such that $Lh - E$ is convex.*

The following Lemma 2.1 is also a simplified version correspond to Definition 2.4, which extends the L-smooth adaptability of the pair (E, h).

Lemma 2.1 (Descent lemma [3,10]). *Let $E, h : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable functions, and assume that h is convex. Then, (E, h) is L-smooth adaptable for $L > 0$ if and only if*

$$E(u) \leq E(v) + \langle \nabla E(v), u - v\rangle + LD_h(u, v), \quad \forall u, v \in \mathbb{R}^n.$$

3 The proposed method

3.1 Alternating linearized Bregman iterations method

In this section, we start by recalling the original Bregman iterations method and consider the bivariate version of the Bregman iterations method under the Gauss-Seidel scheme. Then, we propose the alternating linearized Bregman iterations method that updates the two variables x, y alternately during each iteration.

When we only consider x as a variable and $y = \bar{y}$ is fixed, the optimization problem $(\mathcal{P})$ is reduced to the following problem:

$$\inf \left\{ E(x, \bar{y}) + f(x) : x \in \mathbb{R}^M \right\}. \tag{$\mathcal{P}_1$}$$

The Bregman iterations method for the problem $(\mathcal{P}_1)$ is given by

$$x^{k+1} := \arg\min_x \left\{ E(x, \bar{y}) + D_f^{p_x^k}(x, x^k) \right\}, \tag{3.1}$$

where p_x^k is a subgradient of the function $f(x^k)$, i.e., $p_x^k \in \partial R(x^k)$, and $D_f^{p_x^k}(x, x^k) = f(x) - f(x^k) - \langle p_x^k, x - x^k \rangle$ is the Bregman distance of f between x and x^k. In fact, the Bregman iterations method replaces the term $f(x)$ in $(\mathcal{P}_1)$ with the generalized Bregman distance $D_f^{p_x^k}(x, x^k)$ to play the role of regularization. When $x = \bar{x}$ is fixed and consider y as a variable, we can obtain a similar approach with respect to y. To apply the Bregman iterations method to the problem $(\mathcal{P})$, we need to consider the bivariate version of the Bregman iterations method, which is given by

$$\begin{aligned} x^{k+1} &:= \arg\min_x \left\{ E(x, y^k) + D_f^{p_x^k}(x, x^k) \right\}, \\ y^{k+1} &:= \arg\min_y \left\{ E(x^{k+1}, y) + D_g^{p_y^k}(y, y^k) \right\}, \end{aligned} \tag{3.2}$$

where $p_x^k \in \partial f(x^k)$ and $p_y^k \in \partial g(y^k)$ are subgradients of the functions $f(x^k)$ and $g(y^k)$, respectively. The approach (3.2) is via the Gauss-Seidel iteration scheme, which updates the two variables x, y alternately during each iteration.

Suggested by the recent work of BPG and LBreI in [10,14], we can extend the Bregman method (3.2) beyond Lipschitz gradient continuity assumptions by replace the objective function E with the following approximation

$$\begin{aligned} E(x, y^k) &\approx E(x^k, y^k) + \langle \nabla_x E(x^k, y^k), x - x^k \rangle + \frac{1}{\delta_x^k} D_{h(x, y^k)}(x, x^k), \\ E(x^{k+1}, y) &\approx E(x^{k+1}, y^k) + \langle \nabla_y E(x^{k+1}, y^k), y - y^k \rangle + \frac{1}{\delta_y^k} D_{h(x^{k+1}, y)}(y, y^k), \end{aligned} \tag{3.3}$$

where h is a suitable kernel generating distance, and $\delta_x^k, \delta_y^k > 0$ are the step sizes. ∇_x and ∇_y are the subvectors of the the gradient of $E(x, \bar{y})$ and $E(\bar{x}, y)$, respectively.

So far, the iteration of the sequence $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ that is obtained from the linearized Bregman iterations method for the problem $(\mathcal{P})$ is given by

Algorithm 1 Alternating Linearized Bregman Iteration Method

Require: $\delta_x^k, \delta_y^k, x^0, y^0$ is given and $p_x^0 = \partial f(x^0), p_y^0 = \partial g(y^0)$.
Ensure: For $k = 1, 2, \ldots$, and compute:

$$
\begin{aligned}
x^{k+1} &\in \arg\min_{x\in\mathbb{R}^M} \left\{ \langle \nabla_x E(x^k, y^k), x - x^k\rangle + \frac{1}{\delta_x^k} D_{h(\cdot,y^k)}(x, x^k) + D_f^{p_x^k}(x, x^k) \right\}, \\
y^{k+1} &\in \arg\min_{y\in\mathbb{R}^N} \left\{ \langle \nabla_y E(x^{k+1}, y^k), y - y^k\rangle + \frac{1}{\delta_y^k} D_{h(x^{k+1},\cdot)}(y, y^k) + D_g^{p_y^k}(y, y^k) \right\}, \\
p_x^{k+1} &= p_x^k - \frac{1}{\delta_x^k}\left[\nabla_x h(x^{k+1}, y^k) - \nabla_x h(x^k, y^k) + \delta_x^k \nabla_x E(x^k, y^k) \right], \\
p_y^{k+1} &= p_y^k - \frac{1}{\delta_y^k}\left[\nabla_y h(x^{k+1}, y^{k+1}) - \nabla_y h(x^{k+1}, y^k) + \delta_y^k \nabla_y E(x^{k+1}, y^k) \right].
\end{aligned}
$$

Algorithm 1 extends the linearized Bregman iterations method in [14] to the bivariate version and update one of the two variants x and y when the other one is fixed.

3.2 The blockwise variant

To tackle the regularized NMF problem at hand, this paper introduces a blockwise variant that is denoted using the following notation. The i block of the vectors $x \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \cdots \times \mathbb{R}^{m_d}$ and $y \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_d}$ are represented as x_i and y_i, respectively. Moreover, x and y are vertically concatenated by their corresponding blocks x_i and y_i for $i = 1, 2, \ldots, d$, resulting in:

$$
\begin{aligned}
(x_1; x_2; \ldots; x_d) &= (x_1^T, x_2^T, \ldots, x_d^T)^T = x, \\
(y_1; y_2; \ldots; y_d) &= (y_1^T, y_2^T, \ldots, y_d^T)^T = y.
\end{aligned}
$$

Additionally, the convex functions f and g are consists of d blocks, i.e. $f(x) = \sum_{j=1}^d f_j(x_j)$ and $g(y) = \sum_{j=1}^d g_j(y_j)$. Let $M = \sum_{j=1}^d m_j, N = \sum_{j=1}^d n_j$, the non-convex optimization problem $(\mathcal{P})$ is then transformed into blockwise variant as follows:

$$
\inf \left\{ \psi(x, y) \equiv E(x_1, \ldots, x_d, y_1, \ldots, y_d) + \sum_{j=1}^d f_j(x_j) + \sum_{j=1}^d g_j(y_j) : x_j \in \mathbb{R}^{m_j}, y_j \in \mathbb{R}^{n_j} \right\}, \qquad (\mathcal{P}_0)
$$

where $E : \mathbb{R}^d \times \mathbb{R}^d \to (-\infty, +\infty]$ is a continuously differentiable function, and for all $j = 1, \ldots, d$, each $f_j : \mathbb{R}^{m_j} \to [0, +\infty]$ and $g_j : \mathbb{R}^{n_j} \to [0, +\infty]$ are both proper lower semicontinuous and convex functions. We assume that $\inf\{E(x, y) : x \in \mathbb{R}^M, y \in \mathbb{R}^N\} > -\infty$ and $\inf\left\{\psi(x, y) : x \in \mathbb{R}^M, y \in \mathbb{R}^N\right\} > -\infty$, which hold trivially for nonnegative regularized functions f_j, g_j and objective function E for all $j = 1, \ldots, d$.

For simplicity, let $\bar{x} \in \mathbb{R}^M$, $\bar{y} \in \mathbb{R}^N$ represent the given point, and for each $j = 1, 2, \ldots, d$, let $E^{(\bar{x},j,\bar{y})}(x_j)$ refer to the functions assigned by replacing the j-th block of $\bar{x}$ whereas, in $E^{(\bar{x},\bar{y},j)}(y_j)$, the j-th block of $\bar{y}$ is taken, i.e. $E^{(\bar{x},j,\bar{y})}(x_j) = E(\bar{x}_1, \ldots, x_j, \ldots, \bar{x}_d, \bar{y}_1, \ldots, \bar{y}_d)$, the other one similarly. This notation leads to the following optimization problem, which is a blockwise variant

of $(\mathcal{P}_0)$:

$$\begin{aligned} &\inf\left\{E^{(\bar{x},j,\bar{y})}(x_j)+f_j(x_j):x_j\in\mathbb{R}^{m_j}\right\},\\ &\inf\left\{E^{(\bar{x},\bar{y},j)}(y_j)+g_j(y_j):y_j\in\mathbb{R}^{n_j}\right\}. \end{aligned} \tag{$\mathcal{B}$}$$

In this case, the above model is throughout the paper. Moreover, the functions $h_x^{(j)}(x_j)$, $h_y^{(j)}(y_j)$ represent $h^{(j)}(x_j,\bar{y}_j)$ and $h^{(j)}(\bar{x}_j,y_j)$, respectively. In our work, the following standing assumption is proposed:

Assumption 3.1. *For any $(\bar{x},\bar{y})\in\mathbb{R}^M\times\mathbb{R}^N$ and $j=1,2,\ldots,d$,*

1) *there are convex differentiable functions $h_x^{(j)}:\mathbb{R}^{m_j}\to\mathbb{R}, h_y^{(j)}:\mathbb{R}^{n_j}\to\mathbb{R}$ such that $(h_x^{(j)},E^{(\bar{x},j,\bar{y})})$ is $L_x^{(j)}$-smooth adaptable, and $(h_y^{(j)},E^{(\bar{x},\bar{y},j)})$ is $L_x^{(j)}$-smooth adaptable;*

2) *the set of Lipschitz constant $\bigcup_{j=1}^d\{L_x^{(j)},L_y^{(j)}\}$ is a bounded set, the lower bound is $\underline{L}$ as well as the upper bound is $\overline{L}$;*

3) *the functions $h_x^{(j)}$ and $h_y^{(j)}$ are strongly convex with respect to x_j and y_j with parameter $\sigma_{j,1}$ and $\sigma_{j,2}$, respectively;*

4) *the functions $h_x^{(j)},h_y^{(j)}$ and E are locally gradient Lipschitz continuous on any bounded subset of $\mathbb{R}^{m_j},\mathbb{R}^{n_j}$ or $\mathbb{R}^M\times\mathbb{R}^N$, respectively.*

Our approach to solve the problem is via the Gauss-Seidel iteration scheme, which is also known as alternating minimization. We begin with a given initial point $(x^0,y^0)\in\mathbb{R}^M\times\mathbb{R}^N$ and generate a sequence $\left\{(x^k,y^k)\right\}_{k\in\mathbb{N}}$ using the following scheme: first, we set $T\geq d$. For $k=0,1,2,\cdots$, we choose $j_k\in\{1,2,\ldots,d\}$ and compute:

$$\text{For } j\neq j_k: \qquad \begin{cases} x_j^{k+1}=x_j^k,\\ y_j^{k+1}=y_j^k. \end{cases}$$

$$\text{For } j=j_k: \quad \begin{cases} x_j^{k+1}\in\arg\min\limits_x E(x,y^k),\\ y_j^{k+1}\in\arg\min\limits_y E(x^{k+1},y). \end{cases}$$

Remark 3.1. *The Gauss-Seidel scheme performs an iteration step on the j_k-th block of $(\mathcal{B})$. The rule for selecting the blocks is such that every d blocks are included consecutively in each T iterations of the scheme. Hence, for $k\geq T$, it holds that:*

$$\bigcup_{l=k-T+1}^{k}\{j_l\}=\{1,2,\ldots,d\}.$$

Remark 3.2. *The paper considers a general approach known as the essentially cyclic regime, where $j_k=(k \mod d)+1$. However, in the overlapping version, we can update multiple blocks simultaneously at each iteration. This means that for $k=0,1,2,\ldots,$ we choose a non-empty set*

$J_k \subset \{1, 2, \ldots, d\}$ *and perform the iteration on all blocks in* J_k *together [11]. It should be noted that a modification of Assumption 3.1 will be necessary in order to handle the subproblems:*

$$\inf \left\{ E^{(\bar{x}, J, \bar{y})}(x_J) + \sum_{j \in J} f_j(x_j) : x_J \in \mathbb{R}^{m_J} \right\},$$

$$\inf \left\{ E^{(\bar{x}, \bar{y}, J)}(y_J) + \sum_{j \in J} g_j(y_j) : y_J \in \mathbb{R}^{n_J} \right\},$$

where $m_J = \sum_{j \in J} m_j$ *and* $n_J = \sum_{j \in J} n_j$*. The subvectors* x_J, y_J *are composed from the blocks in* J*. Furthermore,* $E^{(\bar{x}, J, \bar{y})}(x_J)$ *is a function of the subvector* x_J *consisting of the blocks in* J*, and similarly,* $E^{(\bar{x}, \bar{y}, J)}(y_J)$ *is a function of the subvector* y_J *consisting of the blocks in* J*.*

Thus, we are now ready to propose the alternating block linearized Bregman iterations method:

Algorithm 2 Alternating Block Linearized Bregman Iteration Method

Require: δ_x^k, δ_y^k, (x^0, y^0) are given and $(p_{x_j}^0, p_{y_j}^0) \in (\partial f(x_j^0), \partial g(x_j^0)), \forall j = 1, \ldots, d$.
Ensure: For $k = 0, 1, \ldots$, choose $j_k \in \{1, \ldots, d\}$, and compute:
if $j \neq j_k$ **then**
$x_j^{k+1} = x_j^k$
$y_j^{k+1} = y_j^k$
else if $j = j_k$ **then**

$$x_j^{k+1} \in \arg\min_{x \in \mathbb{R}^{n_j}} \left\{ \left\langle \nabla E^{(\bar{x}^k, j, \bar{y}^k)}(x_j^k), x - x_j^k \right\rangle + \frac{1}{\delta_x^k} D_{h_x^{(j)}}(x, x_j^k) + D_{f_j}^{p_x^k}(x, x_j^k) \right\}, \tag{4.1a}$$

$$y_j^{k+1} \in \arg\min_{y \in \mathbb{R}^{m_j}} \left\{ \left\langle \nabla E^{(\bar{x}^{k+1}, \bar{y}^k, j)}(y_j^k), y - y_j^k \right\rangle + \frac{1}{\delta_y^k} D_{h_y^{(j)}}(y, y_j^k) + D_{g_j}^{p_y^k}(y, y_j^k) \right\}, \tag{4.1b}$$

$$p_{x_j}^{k+1} = p_{x_j}^k - \frac{1}{\delta_x^k} \left[\nabla h_x^{(j)}(x_j^{k+1}) - \nabla h_x^{(j)}(x_j^k) + \delta_x^k \nabla E^{(\bar{x}^k, j, \bar{y}^k)}(x_j^k) \right], \tag{4.1c}$$

$$p_{y_j}^{k+1} = p_{y_j}^k - \frac{1}{\delta_y^k} \left[\nabla h_y^{(j)}(y_j^{k+1}) - \nabla h_y^{(j)}(y_j^k) + \delta_y^k \nabla E^{(\bar{x}^{k+1}, \bar{y}^k, j)}(y_j^k) \right]. \tag{4.1d}$$

end if

Remark 3.3. *At the k-th iteration of the Gauss-Seidel scheme, we update* x_j *and* y_j *in turn using a linearized Bregman iterations framework. The minimization subproblems of the linearized Bregman iterations have unique solutions* (x_j^{k+1}, y_j^{k+1}) *under the assumption that the kernels* $h^{(j)}$ *are strongly convex, which satisfies Assumption 3.1.*

4 Convergence analysis

In this paper, we use an alternating update scheme in which we update the j_k-th block of x when the others of x and the blocks of y are fixed. This approach allows us to calculate the decline of the two variables separately, enabling us to obtain the overall decrease. To prove the descent property of the objective function with respect to the Bregman distance, we need the following lemma.

Lemma 4.1 (Descent inequalities)**.** *Under the settings of ($\mathcal{P}_0$) and Assumption 3.1, we have*

$$\begin{aligned} &E(x^{k+1},y^k)-E(x^k,y^k)\le L_x^{(j_k)}D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k)\\ &-\frac{1}{\delta_x^k}D_{h_x^{(j_k)}}(x_{j_k}^k,x_{j_k}^{k+1})-\frac{1}{\delta_x^k}D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k)-D_{f_{j_k}}(x_{j_k}^k,x_{j_k}^{k+1})-D_{f_{j_k}}(x_{j_k}^{k+1},x_{j_k}^k). \end{aligned} \tag{4.1}$$

Proof. To prove (4.1), we first rephrased (4.1a) as the following equality:

$$p_{x_j}^{k+1}-p_{x_j}^k+\nabla E^{(x^k,j_k,y^k)}(x_{j_k}^k)+\frac{1}{\delta_x^k}\left[\nabla h_x^{(j_k)}(x_{j_k}^{k+1})-\nabla h_x^{(j_k)}(x_{j_k}^k)\right]=0. \tag{4.2}$$

To obtain the desired result, the inner product is taken between the left-hand side of this expression and the term $x_{j_k}^k-x_{j_k}^{k+1}$. Using the well-known generalized three-points identity (2.1), we have:

$$\left\langle p_{x_j}^{k+1}-p_{x_j}^k,x_{j_k}^k-x_{j_k}^{k+1}\right\rangle=-D_{f_{j_k}}(x_{j_k}^k,x_{j_k}^{k+1})-D_{f_{j_k}}(x_{j_k}^{k+1},x_{j_k}^k). \tag{4.3}$$

In the same fashion, we can write:

$$\begin{aligned} &\frac{1}{\delta_x^k}\left\langle\nabla h_x^{(j_k)}(x_{j_k}^{k+1})-\nabla h_x^{(j_k)}(x_{j_k}^k),x_{j_k}^k-x_{j_k}^{k+1}\right\rangle\\ &=-\frac{1}{\delta_x^k}D_{h_x^{(j_k)}}(x_{j_k}^k,x_{j_k}^{k+1})-\frac{1}{\delta_x^k}D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k). \end{aligned} \tag{4.4}$$

Applying the inequality $E(u)\le E(v)+\langle\nabla E(v),u-v\rangle+LD_h(u,v)$ from Lemma 2.1 and setting $u=x_{j_k}^{k+1},v=x_{j_k}^k$, we obtain:

$$\begin{aligned} &\langle\nabla E^{(x^k,j_k,y^k)}(x_{j_k}^k),x_{j_k}^k-x_{j_k}^{k+1}\rangle\\ &\le E^{(x^k,j_k,y^k)}(x_{j_k}^k)-E^{(x^k,j_k,y^k)}(x_{j_k}^{k+1})+L_x^{(j_k)}D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k). \end{aligned} \tag{4.5}$$

Combining (4.3)-(4.5) and applying to (4.2), given the fact that $E^{(x^k,j_k,y^k)}(x_{j_k}^k)=E(x^k,y^k),\forall k\ge 1$, we can obtain the desired descent inequality (4.1). □

Applying Lemma 4.1 in a similar manner to fixing x and updating y, we obtain the following lemma, which guarantees sufficient descent property and summability of Bregman distance.

Lemma 4.2 (Sufficient descent and summability)**.** *Let $\{x^k,y^k\}$ be the sequence generated by the algorithm with the stepsizes δ_x^k and δ_y^k that satisfy*

$$\begin{aligned} &0<\delta_x^k<\frac{1+\eta(h_x^{(j_k)})-\omega_x^k}{L_x^{(j_k)}},\quad\exists\omega_x^k\in(0,1+\eta(h_x^{(j_k)})),\\ &0<\delta_y^k<\frac{1+\eta(h_y^{(j_k)})-\omega_y^k}{L_y^{(j_k)}},\quad\exists\omega_y^k\in(0,1+\eta(h_y^{(j_k)})). \end{aligned}$$

Denote

$$\rho_x^k:=\frac{L_x^{(j_k)}\omega_x^k}{1+\eta(h_x^{(j_k)})-\omega_x^k},\quad\rho_y^k:=\frac{L_y^{(j_k)}\omega_y^k}{1+\eta(h_y^{(j_k)})-\omega_y^k}.$$

The following inequality guarantees a sufficient decrease property in terms of the Bregman distance:

$$\begin{aligned} E(x^{k+1},y^{k+1})-E(x^k,y^k) \le & -\rho_x^k D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k)-D_{f_{j_k}}^{symm}(x_{j_k}^{k+1},x_{j_k}^k) \\ & -\rho_y^k D_{h_y^{(j_k)}}(y_{j_k}^{k+1},y_{j_k}^k)-D_{g_{j_k}}^{symm}(y_{j_k}^{k+1},y_{j_k}^k). \end{aligned}$$

In particular, we have

$$\begin{aligned} & \lim_{k\to\infty} D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k)=\lim_{k\to\infty} D_{h_y^{(j_k)}}(y_{j_k}^{k+1},y_{j_k}^k) \\ =&\lim_{k\to\infty} D_{f_{j_k}}^{symm}(x_{j_k}^{k+1},x_{j_k}^k)=\lim_{k\to\infty} D_{g_{j_k}}^{symm}(y_{j_k}^{k+1},y_{j_k}^k)=0. \end{aligned}$$

Proof. We begin by using (4.1) to derive an expression for $E(x^{k+1},y^{k+1})-E(x^k,y^k)$,

$$\begin{aligned} & E(x^{k+1},y^k)-E(x^k,y^k) \\ & \quad \le -D_{f_{j_k}}^{symm}(x_{j_k}^{k+1},x_{j_k}^k)+\left(L_x^{(j_k)}-\frac{1}{\delta_x^k}\right)D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k)-\frac{1}{\delta_x^k}D_{h_x^{(j_k)}}(x_{j_k}^k,x_{j_k}^{k+1}) \\ & \quad \le -D_{f_{j_k}}^{symm}(x_{j_k}^{k+1},x_{j_k}^k)+\frac{L_x^{(j_k)}\left(\eta(h_x^{(j_k)})-\omega_x^k\right)}{1+\eta(h_x^{(j_k)})-\omega_x^k}D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k) \\ & \qquad -\frac{L_x^{(j_k)}}{1+\eta(h_x^{(j_k)})-\omega_x^k}D_{h_x^{(j_k)}}(x_{j_k}^k,x_{j_k}^{k+1}) \\ & \quad \le -D_{f_{j_k}}^{symm}(x_{j_k}^{k+1},x_{j_k}^k)-\rho_x^k D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k), \end{aligned}$$

where the second inequality follows form the definition of $\eta(h_x^{(j_k)})$ as (2.2).

Similarly to derive the decrease inequalities updating y^k:

$$\begin{aligned} & E(x^{k+1},y^{k+1})-E(x^{k+1},y^k) \\ & \quad \le -D_{g_{j_k}}^{symm}(y_{j_k}^{k+1},y_{j_k}^k)+\left(L_y^{(j_k)}-\frac{1}{\delta_y^k}\right)D_{h_y^{(j_k)}}(y_{j_k}^{k+1},y_{j_k}^k)-\frac{1}{\delta_y^k}D_{h_y^{(j_k)}}(y_{j_k}^k,y_{j_k}^{k+1}) \\ & \quad \le -D_{g_{j_k}}^{symm}(y_{j_k}^{k+1},y_{j_k}^k)+\frac{L_y^{(j_k)}\left(\eta(h_y^{(j_k)})-\omega_y^k\right)}{1+\eta(h_y^{(j_k)})-\omega_y^k}D_{h_y^{(j_k)}}(y_{j_k}^{k+1},y_{j_k}^k) \\ & \qquad -\frac{L_x^{(j_k)}}{1+\eta(h_y^{(j_k)})-\omega_y^k}D_{h_y^{(j_k)}}(y_{j_k}^k,y_{j_k}^{k+1}) \\ & \quad \le -D_{g_{j_k}}^{symm}(y_{j_k}^{k+1},y_{j_k}^k)-\rho_y^k D_{h_y^{(j_k)}}(y_{j_k}^{k+1},y_{j_k}^k), \end{aligned}$$

Combining the above two inequalities yields the following

$$\begin{aligned} E(x^{k+1},y^{k+1})-E(x^k,y^k) \le & -\rho_x^k D_{h_x^{(j_k)}}(x_{j_k}^{k+1},x_{j_k}^k)-D_{f_{j_k}}^{symm}(x_{j_k}^{k+1},x_{j_k}^k) \\ & -\rho_y^k D_{h_y^{(j_k)}}(y_{j_k}^{k+1},y_{j_k}^k)-D_{g_{j_k}}^{symm}(y_{j_k}^{k+1},y_{j_k}^k). \end{aligned} \tag{4.6}$$

Summing these inequalities over $k = 0, \ldots, N$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{N} \left(\rho_x^k D_{h_x^{(j_k)}}(x_{j_k}^{k+1}, x_{j_k}^k) + D_{f_{j_k}}^{symm}(x_{j_k}^{k+1}, x_{j_k}^k) + \rho_y^k D_{h_y^{(j_k)}}(y_{j_k}^{k+1}, y_{j_k}^k) + D_{g_{j_k}}^{symm}(y_{j_k}^{k+1}, y_{j_k}^k) \right) \\
&\leq E(x^0, y^0) - E(x^{N+1}, y^{N+1}) \\
&\leq E(x^0, y^0) - \inf_{(x,y)\in \mathrm{dom}\ h} E(x, y).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\sum_{k=0}^{\infty} D_{h_x^{(j_k)}}(x_{j_k}^{k+1}, x_{j_k}^k) < \infty, \sum_{k=0}^{\infty} D_{f_{j_k}}^{symm}(x_{j_k}^{k+1}, x_{j_k}^k) < \infty, \\
\sum_{k=0}^{\infty} D_{h_y^{(j_k)}}(y_{j_k}^{k+1}, y_{j_k}^k) < \infty, \sum_{k=0}^{\infty} D_{g_{j_k}}^{symm}(y_{j_k}^{k+1}, y_{j_k}^k) < \infty.
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
&\lim_{k\to\infty} D_{h_x^{(j_k)}}(x_{j_k}^{k+1}, x_{j_k}^k) = \lim_{k\to\infty} D_{h_y^{(j_k)}}(y_{j_k}^{k+1}, y_{j_k}^k) \\
&= \lim_{k\to\infty} D_{f_{j_k}}^{symm}(x_{j_k}^{k+1}, x_{j_k}^k) = \lim_{k\to\infty} D_{g_{j_k}}^{symm}(y_{j_k}^{k+1}, y_{j_k}^k) = 0.
\end{aligned}$$

which completes the proof. □

The theorem below provides such a convergence property, which is a consequence of the sufficient decrease property summarized in the previous lemma.

Theorem 4.1 (Convergence)**.** *Assume that Assumptions 3.1 holds, and let* $\{z^k\}_{k\in\mathbb{N}} := \{(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by the ABLBreI method, with* $\omega = \min_{k\geq 1}\{\omega_x^k, \omega_y^k\}$*. Then, the following results hold:*

1) The sequence $\left\{E(z^k)\right\}_{k\in\mathbb{N}}$ *is non-increasing and in particular*

$$\frac{\rho}{2}\|z^{k+1} - z^k\|^2 \leq E(z^k) - E(z^{k+1}), \qquad \forall k \geq 0,$$

where

$$\rho = \min_{j=1,2,\cdots,d} \left\{ \frac{\underline{L}\omega_x^k \sigma_{j,1}}{1 + \eta(h_x^{(j_k)}) - \omega_x^k}, \frac{\underline{L}\omega_y^k \sigma_{j,2}}{1 + \eta(h_y^{(j_k)}) - \omega_y^k} \right\}.$$

2) We have

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2 + \|y^{k+1} - y^k\|^2 = \sum_{k=1}^{\infty} \|z^{k+1} - z^k\|^2 < \infty,$$

and hence $\lim_{k\to\infty} \|z^{k+1} - z^k\| = 0$.

Proof. (i) According to the conclusion in Lemma 4.1,

$$\begin{aligned}
&E(x^k, y^k) - E(x^{k+1}, y^{k+1}) \\
&\geq \rho_x^k D_{h_x^{(j_k)}}(x_{j_k}^{k+1}, x_{j_k}^k) + D_{f_{j_k}}^{symm}(x_{j_k}^{k+1}, x_{j_k}^k) + \rho_y^k D_{h_y^{(j_k)}}(y_{j_k}^{k+1}, y_{j_k}^k) + D_{g_{j_k}}^{symm}(y_{j_k}^{k+1}, y_{j_k}^k) \\
&\geq \rho_x^k D_{h_x^{(j_k)}}(x_{j_k}^{k+1}, x_{j_k}^k) + \rho_y^k D_{h_y^{(j_k)}}(y_{j_k}^{k+1}, y_{j_k}^k) \\
&\geq \frac{\rho_x^k \sigma_{j_k,1}}{2}\|x_{j_k}^{k+1} - x_{j_k}^k\|^2 + \frac{\rho_y^k \sigma_{j_k,2}}{2}\|y_{j_k}^{k+1} - y_{j_k}^k\|^2.
\end{aligned}$$

It follows that the sequence $\{E(z^k)\}_{k\in\mathbb{N}}$ is non-increasing and in particular

$$E(z^k) - E(z^{k+1}) \geq \frac{\rho}{2}\|z^{k+1} - z^k\|^2, \quad \forall k \geq 0, \tag{4.7}$$

where

$$\rho = \min_{j=1,2,\cdots,d} \left\{ \frac{\underline{L}\omega_x^k \sigma_{j,1}}{1+\eta(h_x^{(j_k)}) - \omega_x^k}, \frac{\underline{L}\omega_y^k \sigma_{j,2}}{1+\eta(h_y^{(j_k)}) - \omega_y^k} \right\} = \min_{k\geq 1} \left\{ \rho_x^k \sigma_{j_k,1}, \rho_y^k \sigma_{j_k,2} \right\}.$$

(ii) Let N be a positive integer. Summing the above equation from $k = 0$ to $N-1$ and since E is assumed to be bounded from below, it holds:

$$\begin{aligned}\sum_{k=1}^{N-1} \left\|x^{k+1} - x^k\right\|^2 + \left\|y^{k+1} - y^k\right\|^2 = \sum_{k=1}^{N-1} \left\|z^{k+1} - z^k\right\|^2 &\leq \frac{2}{\rho}\left(E(z^0) - E(z^N)\right) \\ &\leq \frac{2}{\rho}\left(E(z^0) - \inf E(z)\right).\end{aligned}$$

Taking the limit as $N \to \infty$, we obtain the desired assertion.

□

5 ABLBreI for RNMF

Therefore, we return to the optimization problem that needs to be solved, which is non-negative matrix factorization with regularization. We need to select the kernel that generates the distance that satisfies Assumption 3.1. Consequently, we propose a blockwise version for $j = 1, 2, \ldots, r$:

$$h^{(j)}(x_j, y_j) = \frac{1}{4}\left(\|x_j\|^2 + \|y_j\|^2\right)^2 + \frac{1}{2}\left(\|x_j\|^2 + \|y_j\|^2\right). \tag{5.1}$$

The kernel generating distance was used in [10] and verified to satisfy Assumption 3.1. After that, we recall two well-known operators that will be used to compute the explicit formula for the ABLBreI method.

Let $\mathcal{S}_\tau$ be the symbol of soft-thresholding(with parameter τ):; thus, for any $z \in \mathbb{R}^s$, it holds that

$$\mathcal{S}_\tau(z) = \arg\min_{w\in\mathbb{R}^d} \left\{ \tau\|w\|_1 + \frac{1}{2}\|w - z\|^2 \right\} = \max\{|z| - \tau, 0\}\,\mathrm{sgn}(z)$$

Let $[\cdot]_+$ be the projection onto $\mathbb{R}^s_+$, which is used to force the factorization matrices to be non-negative; thus, for any $z \in \mathbb{R}^s$, it holds that

$$[z]_+ = \arg\min_{w\in\mathbb{R}^d_+} \left\{ \|w - z\|^2 \right\} = \max\{z, 0\}.$$

And we choose ℓ_1 regularization to induce the sparsity, i.e.,

$$\psi(x, y) = \frac{1}{2}\|A - \sum_{i=1}^{r} x_i y_i^T\|_F^2 + \sum_{i=1}^{r} \mu_1\|x_i\|_1 + \sum_{i=1}^{r} \mu_2\|y_i\|_1.$$

The following proposition is the core to solve the optimization problem, which was used in [3].

Proposition 5.1 (Bregman proximal formula for the ℓ_1-norm regularization)**.** *The explicit formula of*

$$x^+ = \arg\min_{u\in\mathbb{R}^d} \left\{ \lambda\theta\|u\|_1 + \langle \mathcal{V}(x), u\rangle + \frac{1}{4}\|u\|^4 + \frac{\alpha}{2}\|u\|^2 \right\} \qquad (\mathcal{T})$$

is given by $x^+ = -t^*\mathcal{S}_{\lambda\theta}(\mathcal{V}(x))$*, where* x *is the current step and* x^+ *is the updated result,* t^* *is the unique positive real root of*

$$t^3\|\mathcal{S}_{\lambda\theta}(\mathcal{V}(x))\|^2 + \alpha t - 1 = 0.$$

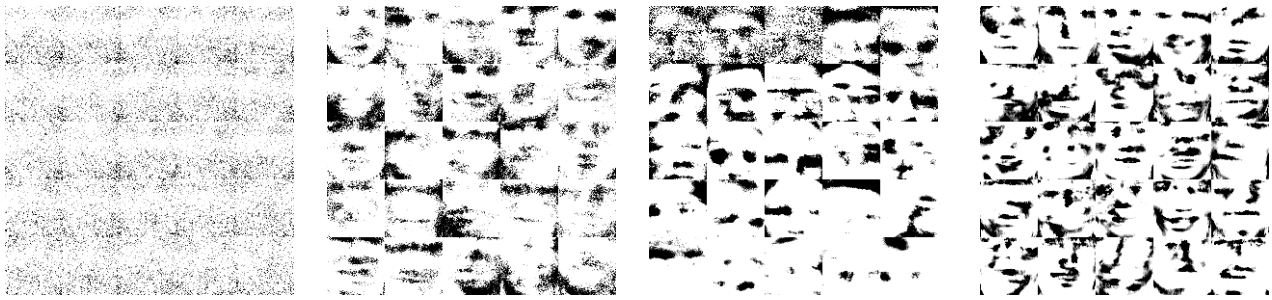
Let the step size parameters δ_x^k and δ_y^k both be equal to ρ. Thus, the closed-form solution of ABLBreI for NMF is in the form as Algorithm 3.

6 Numerical experiments

In our numerical experiments[1], we utilize standard methods under the cyclic block selection regime in the form of $j_k = (k \mod d) + 1$ to maintain the relevance and simplicity of the presentation. In order to test the performance of each method, we conduct experiments on both synthetic and real datasets. The synthetic datasets are generated using the Julia command A = sprand(Float64, 200, 10, 0.5)*sprand(Float64, 10, 200, 0.5), which produces a 200×200 matrix. For our real dataset, we chose the Olivetti Research Laboratory[2] (ORL) dataset. We consider constrained models with columnwise ℓ_1-regularization to induce sparsity of the blocks of the factorization matrices reflecting the form

$$\min_{X\geq 0, Y\geq 0} \left\{ \psi \equiv \frac{1}{2}\|A - xy^T\|_{\mathrm{F}}^2 + \mu_1 \sum_{i=1}^{r} \|x_i\|_1 + \mu_2 \sum_{i=1}^{r} \|y_i\|_1 \right\},$$

in which x_i and y_i are the column vectors of x and y, respectively, and μ_1 and μ_2 are the coefficient parameters of ℓ_1-regularization. Figure 1 visualizes the facial feature extraction through the recovery of weight matrices with the ABLBreI method.



(a) $t = 30$ (b) $t = 180$ (c) $t = 600$ (d) $t = 3000$

Figure 1: *The face extraction features of NMF visualized through the recovery of weight matrices, where t denotes the iteration time in seconds.*

We conduct experiments using the Block Bregman Proximal Gradient method as comparison, the alternating linearized Bregman iterations method (ALBreI) with $d = 1$ and the ABLBreI

[1]Code is available at `https://github.com/bellhello/ABLBreI-RNMF`

[2]AT&T Laboratories Cambridge.

Algorithm 3 Alternating Block Linearized Bregman Iteration for the NMF with ℓ_1-regularization

Require: $\rho \in (0,1)$, $(x^0, y^0) \in \mathrm{Lev}(E, 1/2)$, $(p_{x_j}^0, p_{y_j}^0) = (\partial\|x_j^0\|_1, \partial\|y_j^0\|_1), \forall j = 1, \cdots, r$.
Ensure: For $k = 1, 2, \ldots$ and $j_k \in \{1, 2 \ldots, r\}$, and compute:
if $j_k \neq j$ **then**

$$x_j^{k+1} = x_j^k, \quad y_j^{k+1} = y_j^k,$$

else if $j_k = j$ **then**

$$\begin{aligned}
\mathcal{V}_{x_j}^k &= \left(\sum_{i=1}^r x_i^k (y_i^k)^T - A\right) y_j^k - \rho^{-1}(\|x_j^k\|^2 + \|y_j^k\|^2 + 1)x_j^k - p_{x_j}^k, \\
\mathcal{V}_{y_j}^k &= \left(\sum_{i=1}^r x_i^{k+1} (y_i^k)^T - A\right)^T x_j^{k+1} - \rho^{-1}(\|x_j^{k+1}\|^2 + \|y_j^k\|^2 + 1)y_j^k - p_{y_j}^k, \\
x_j^{k+1} &= \left[-t_x^* \mathcal{S}_{\mu_1\rho}(\rho \mathcal{V}_{x_j}^k)\right]_+, \quad y_j^{k+1} = \left[-t_y^* \mathcal{S}_{\mu_2\rho}(\rho \mathcal{V}_{y_j}^k)\right]_+,
\end{aligned}$$

where t_x^* and t_y^* are the unique positive real root of

$$\begin{aligned}
t_x^3 \|\mathcal{S}_{\mu_1\rho}(\rho \mathcal{V}_{x_j}^k)\|^2 + (1 + \|y_j^k\|^2)t_x - 1 = 0, \\
t_y^3 \|\mathcal{S}_{\mu_2\rho}(\rho \mathcal{V}_{y_j}^k)\|^2 + (1 + \|x_j^{k+1}\|^2)t_y - 1 = 0,
\end{aligned}$$

respectively.

$$\begin{aligned}
p_{x_j}^{k+1} &= p_{x_j}^k - \frac{1}{\rho}\left[(\|x_j^{k+1}\|^2 + \|y_j^k\|^2 + 1)x_j^{k+1} - (\|x_j^k\|^2 + \|y_j^k\|^2 + 1)x_j^k\right] \\
&\quad + \left(\sum_{i=1}^r x_i^k (y_i^k)^T - A\right) y_i^k, \\
p_{y_j}^{k+1} &= p_{y_j}^k - \frac{1}{\rho}\left[(\|x_j^{k+1}\|^2 + \|y_j^{k+1}\|^2 + 1)y_j^{k+1} - (\|x_j^{k+1}\|^2 + \|y_j^k\|^2 + 1)y_j^k\right] \\
&\quad + \left(\sum_{i=1}^r x_i^{k+1} (y_i^k)^T - A\right)^T x_j^{k+1}.
\end{aligned}$$

end if

method that divides the factorization matrices into r blocks evenly, i.e. $d = r$. Furthermore, we conduct experiments using two different sparsity levels—0.2 and 0.05 for the BBPG method. These sparsity levels represent the percentage of non-zero elements in the column vector as 20% and 5%, respectively. We implement all the methods described in this paper using *Julia 1.8.3*, and the numerical experiments are run on a laptop equipped with an Intel i7-11800F CPU @2.30GHz and 32 GB RAM.

We generate the initial matrices X^0 and Y^0 by randomly assigning values to each entry from the interval $[0, 1]$. We ensure that the initial matrices respected the sparsity levels by creating two random mask matrices $M_1 \in \mathbb{R}^{m\times r}$ and $M_2 \in \mathbb{R}^{n\times r}$, so that

$$X := X^0 \odot M_1, \qquad Y := Y^0 \odot M_2.$$

Each entry of the mark matrices are distributed following a Bernoulli distribution with parameter $\gamma \in [0, 1]$, i.e.

$$M_{i,j} \sim \mathcal{B}(\gamma),$$

where the parameter γ is the expected percentage of non-zero elements.

We then apply a rescaling procedure to the initial matrices to ensure that (1) $\|X_i^0\| = \|Y_i^0\|$ for all $i = 1, 2, \ldots, r$, and (2) $\arg\min_t \|A - tX^0(Y^0)^T\| = 1$. We run 20 iterations of each method examined, initializing each with a different point, and reported the average results. For a fair comparison, we use the same initial points for all the methods.

6.1 Synthesis matrix

For the NMF of synthesis matrix, we choose the factorization rank as $r = 10$ and use the regularization parameter $\mu = \mu_1 = \mu_2 \in \{1.00, 1.01\}$. We present the average Frobenius norm of the residual matrix $A - XY^T$ for the NMF model with the sparsity of 80% in Figure 2 and of 95% in Figure 3.

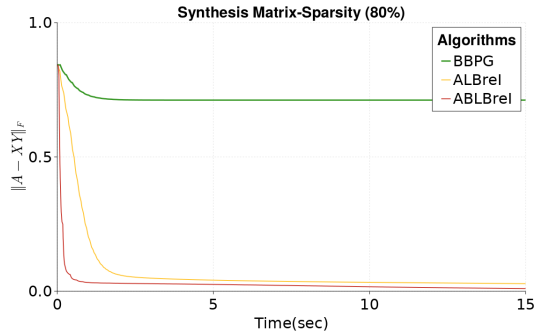


Figure 2: *The Frobenius norm of the residual matrix for the regularized NMF model for the synthesis matrix with sparsity 80%.*

Synthesis-Sparsity (80%), $\mu = 1.00$		
Algorithms	non-zero ratio	
	X	Y
BBPG	0.190	0.190
ALBreI	0.206	0.198
ABLBreI	0.219	0.195

Table 1: *The ratio of the non-zero elements of the factorization matrices respect to Figure 2.*

As we know, the ALBreI method is a special case of the ABLBreI method when the block size is $d = 1$, which is also the reason for removing the "block" from its name. However, the ALBreI method is still a kind of method that updates the variables in "blockwise"; in fact, there are only two blocks, x and y. From the figures above, we can see that the ABLBreI method is more efficient than the ALBreI method, which reflects the accelerating effect of having more blocks.

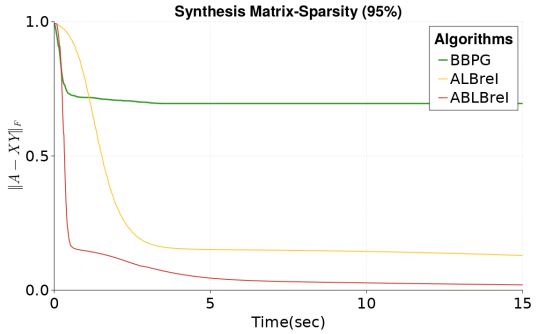


Figure 3: *The Frobenius norm of the residual matrix for the regularized NMF model for the synthesis matrix with sparsity 95%.*

Synthesis-Sparsity (95%), $\mu = 1.01$		
Algorithms	non-zero ratio	
	X	Y
BBPG	0.0500	0.0500
ALBreI	0.0535	0.0460
ABLBreI	0.0542	0.0575

Table 2: *The ratio of the non-zero elements of the factorization matrices with respect to Figure 3.*

6.2 ORL dataset

The ORL dataset consists of 10 face images in different angles for each of the 40 heads. Each face image has a size of 92×112 pixels, resulting in a data matrix of size 10304×400. The size of the data matrix is also the primary reason for the out-of-memory error for the non-blockwise methods. We stress that the factorization rank is an important parameter in NMF research, however, for simplicity, we use $r = 25$ in our experimental setting.

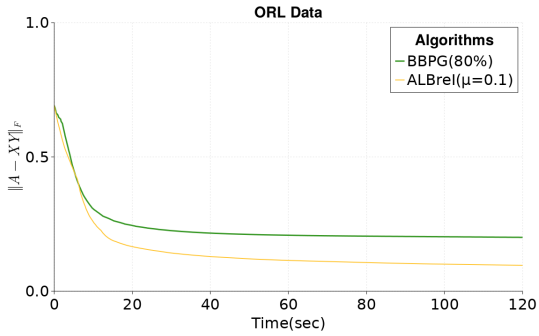


Figure 4: *The Frobenius norm of the residual matrix for the regularized NMF model for ORL dataset.*

ORL-Sparsity (80%), $\mu = 1.00$		
Algorithms	non-zero ratio	
	X	Y
BBPG	0.0500	0.0500
ABLBreI	0.0535	0.0460

Table 3: *The ratio of the non-zero elements of the factorization matrices with respect to Figure 4.*

7 Concluding remarks

Based on the purpose of applying LBreI to solve the NMF with regularization, this paper discusses the strategy of alternating iteration in the case of block variable, and proposes the alternating block linear Bregman iterations algorithm beyond the convexity and Lipschitz gradient continuity to the objective function under the assumption of the convexity of the regularization term. In this paper, we also analyze the descending property and the convergence of the ABLBreI algorithm. Finally,

we give simple demonstrations by numerical experiment. Accelerated variants of the ABLBreI algorithm will be the next step of research.

Acknowledgements

This work is supported by the National Science Foundation of China (Nos.11971480).

References

[1] Bauschke, H., Bolte, J., Teboulle, M.: A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. Mathematics of Operations Research. **42**, 330–348 (2017). DOI 10.1287/moor.2016.0817

[2] Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for non-convex and non-smooth problems. Mathematical Programming. **146**, 459–494 (2014). DOI 10.1007/s10107-013-0701-9

[3] Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First Order Methods Beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems. SIAM Journal on Optimization. **28**, 2131-2151 (2018). DOI 10.1137/17M1138558

[4] Cichocki, A., Zdunek, R., Amari, S.: Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization. Independent Component Analysis and Signal Separation. 169–176 (2007).

[5] Lee, D. D., Seung, H. S.:Learning the parts of objects by non-negative matrix factorization. Nature. **401**, 788-791 (1999). DOI 10.1038/44565

[6] Lee, D. D., Seung, H. S.:Algorithms for Non-Negative Matrix Factorization. Proceedings of the 13th International Conference on Neural Information Processing Systems. 535–541 (2001).

[7] Lin, C.: Projected Gradient Methods for Nonnegative Matrix Factorization. Neural Computation, **19**, 2756–2779, (2007). DOI 10.1162/neco.2007.19.10.2756.

[8] Osher, S., Burger, M., Goldfarb,D.: An Iterative Regularization Method for Total Variation-Based Image Restoration. Multiscale Modeling & Simulation. **4**, 490–509 (2005). DOI 10.1137/040605412

[9] Pock, T., Sabach, S.: Inertial Proximal Alternating Linearized Minimization (iPALM) for Nonconvex and Nonsmooth Problems. SIAM Journal on Imaging Sciences. **9**, 1756–1787 (2016). DOI 10.1137/16M1064064

[10] Teboulle, M., Vaisbourd, Y.: Novel Proximal Gradient Methods for Nonnegative Matrix Factorization with Sparsity Constraints. SIAM Journal on Imaging Sciences. **13**, 381–421 (2020). DOI 10.1137/19M1271750

[11] Tseng, P.: Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. Journal of Optimization Theory and Applications. **109**, 475–494 (2001). DOI 10.1023/A:1017501703105

[12] Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman Iterative Algorithms for ℓ_1-Minimization with Applications to Compressed Sensing. SIAM Journal on Imaging Sciences. **1**, 143–168 (2008). DOI 10.1137/070703983

[13] Yin, W.: Analysis and Generalizations of the Linearized Bregman Method. SIAM Journal on Imaging Sciences. **3**, 856-877 (2010). DOI 10.1137/090760350

[14] Zhang, H., Zhang, L., Yang, H.: Revisiting Linearized Bregman Iterations under Lipschitz-like Convexity Condition. Math. Comp. **92**, 779-803 (2023). DOI 10.1090/mcom/3792