Block Majorization Minimization with Extrapolation and Application to β-NMF

Le Thi Khanh Hien [1] **Valentin Leplat** [2] **Nicolas Gillis** [1]

Abstract

We propose a Block Majorization Minimization method with Extrapolation (BMMe) for solving a class of multi-convex optimization problems. The extrapolation parameters of BMMe are updated using a novel adaptive update rule. By showing that block majorization minimization can be reformulated as a block mirror descent method, with the Bregman divergence adaptively updated at each iteration, we establish subsequential convergence for BMMe. We use this method to design efficient algorithms to tackle nonnegative matrix factorization problems with the β-divergences (β-NMF) for $\beta \in [1,2]$. These algorithms, which are multiplicative updates with extrapolation, benefit from our novel results that offer convergence guarantees. We also empirically illustrate the significant acceleration of BMMe for β-NMF through extensive experiments.

1. Introduction

In this paper, we consider the following class of multi-convex optimization problems:

$$\min_{x_i \in \mathcal{X}_i} f(x_1, \dots, x_s), \tag{1}$$

where $x = (x_1, \dots, x_s)$ is decomposed into s blocks, $\mathcal{X}_i \subseteq \mathbb{E}_i$ is a closed convex set for $i = 1, \dots, s$, $\mathbb{E}_i$ is a finite dimensional real linear space equipped with the norm $\|\cdot\|_{(i)}$ and the inner product $\langle \cdot, \cdot \rangle_{(i)}$ (we will omit the lower-script (i) when it is clear in the context), $\mathcal{X} = \mathcal{X}_1 \times \dots \mathcal{X}_s \subseteq \operatorname{int}\operatorname{dom}(f)$, $f : \mathbb{E} = \mathbb{E}_1 \times \dots \times \mathbb{E}_s \to \mathbb{R} \cup \{+\infty\}$ is a differentiable function over the interior of its domain. Throughout the paper, we assume f is lower bounded and multi-convex, that is, $x_i \mapsto f(x)$ is convex.

1.1. Application to β-NMF, $\beta \in [1,2]$

Nonnegative matrix factorization (NMF) is a standard linear dimensionality reduction method tailored for data sets with nonnegative values (Lee & Seung, 1999). Given a nonnegative data matrix, $X \geq 0$, and a factorization rank, r, NMF aims to find two nonnegative matrices, W with r columns and H with r rows, such that $X \approx WH$. The β divergence is a widely used objective function in NMF to measure the difference between the input matrix, X, and its low-rank approximation, WH (Févotte & Idier, 2011). This problem is referred to as β-NMF and can be formulated in the form of (1) with two blocks of variables, W and H, as follows: given $X \in \mathbb{R}^{m \times n}_+$ and r, solve

$$\min_{\substack{W \in \mathbb{R}^{m \times r}, W \geq \varepsilon, \\ H \in \mathbb{R}^{r \times n}, H \geq \varepsilon}} D_\beta(X, WH), \tag{2}$$

where $D_\beta(X, WH) = \sum_{i=1}^m \sum_{j=1}^n d_\beta(X_{ij}, (WH)_{ij})$, with $d_\beta(x,y) = \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right)$ for $1 < \beta \leq 2$, and $d_\beta(x,y) = x \log \frac{x}{y} - x + y$ for $\beta = 1$. When $\beta = 2$, d_β is the Euclidean distance, and when $\beta = 1$ the Kullback-Leibler (KL) divergence; see Section 4.3 for a discussion on the KL divergence. Note that we consider a small positive lower bound, $\varepsilon > 0$, for W and H to allow the convergence analysis. In practice, we use the machine epsilon for ε, which does not influence the objective function much (Hien & Gillis, 2021).

[1]Department of Mathematics and Operational Research, University of Mons, Mons, Belgium. Emails: khanhhiennt@gmail.com, nicolas.gillis@umons.ac.be. We acknowledge the support by the European Union (ERC consolidator, eLinoR, no 101085607), and by the Francqui Foundation. [2]Center for Artificial Intelligence Technology, Skoltech, Moscow, Russia. Email: v.leplat@skoltech.ru.

1.2. Previous works

Block coordinate descent (BCD) methods serve as conventional techniques for addressing the multi-block Problem (1). These approaches update one block of variables at a time, while keeping the values of the other blocks fixed. There are three main types of BCD methods: classical BCD (Grippo & Sciandrone, 2000; Tseng, 2001), proximal BCD (Grippo & Sciandrone, 2000), and proximal gradient BCD (Beck & Tetruashvili, 2013; Bolte et al., 2014; Tseng & Yun, 2009). These methods fall under the broader framework known as the block successive upper-bound minimization algorithm (BSUM), as introduced by Razaviyayn et al. (2013). In BSUM, a block x_i of x is updated by minimizing a majorizer (also known as an upper-bound approximation function, or a surrogate function; see Defintion 2.1) of the corresponding block objective function.

To accelerate the convergence of BCD methods for nonconvex problems, a well-established technique involves the use of extrapolation points in each block update, as seen in (Xu & Yin, 2013; Ochs et al., 2014; Pock & Sabach, 2016; Ochs, 2019; Hien et al., 2020). Recently, Hien et al. (2023) proposed TITAN, an inertial block majorization-minimization framework for solving a more general class of multi-block composite optimization problems than (1), in which f is not required to be multi-convex. TITAN updates one block of x at a time by selecting a majorizer function for the corresponding block objective function, incorporating inertial force into this majorizer, and then minimizing the resulting inertial majorizer. Through suitable choices of majorizers and extrapolation operators, TITAN recovers several known inertial methods and introduces new ones, as detailed in (Hien et al., 2023, Section 4). TITAN has proven highly effective in addressing low-rank factorization problems using the Frobenius norm, as demonstrated in (Hien et al., 2022; 2023; 2020; Vu Thanh et al., 2021). However, to ensure convergence, TITAN requires the so-called nearly sufficiently decreasing property (NSDP) of the objective function between iterations. The NSDP is satisfied in particular when the majorizer is strongly convex or when the error function, that is, the difference between the majorizer and the objective, is lower bounded by a quadratic function (Hien et al., 2023, Section 2.2). Such requirements pose issues in some situations; for example the Jensen surrogate used to design the multiplicative updates (MU) for standard β-NMF (see Section 2 for the details) lacks strong convexity, nor the corresponding error function is lower bounded by a quadratic function. In other words, although TITAN does not require f to be multi-convex, utilizing TITAN for accelerating the MU in the context of β-NMF is very challenging. This scenario corresponds to a specific instance of Problem (1).

Consider β-NMF (2). For $\beta = 2$, NMF admits very efficient BCD algorithms with theoretically grounded extrapolation (Hien et al., 2020) and heuristic-based extrapolation mechanisms (Ang & Gillis, 2019). Otherwise, the most widely used algorithm to tackle β-NMF are the multiplicative updates (MU):

$$H \leftarrow \text{MU}(X, W, H) = \max\left(\varepsilon, H \circ \frac{\left[W^\top \frac{[X]}{[WH]^{\cdot(2-\beta)}}\right]}{\left[W^\top [WH]^{\cdot(\beta-1)}\right]}\right), \tag{3}$$

and $W^\top \leftarrow \text{MU}\big(X^\top, H^\top, W^\top\big)$, where $\circ$ and $\frac{[\cdot]}{[\cdot]}$ are the component-wise product and division between two matrices, respectively, and $(.)^{\cdot x}$ denotes the component-wise exponent. The MU are guaranteed to decrease the objective function (Févotte & Idier, 2011); see Section 2.1 for more details. Note that, by symmetry of the problem, since $X = WH \iff X^\top = H^\top W^\top$, the MU for H and W are the same, up to transposition.

As far as we know, there is currently no existing algorithm in the literature that accelerates the MU while providing convergence guarantees. On the other hand, it is worth noting that an algorithm with a guaranteed convergence in theory may not always translate to practical success. For instance, the block mirror descent method, while being the sole algorithm to ensure global convergence in KL-NMF, does not yield effective performance in real applications, as reported by Hien & Gillis (2021).

1.3. Contribution and outline of the paper

Drawing inspiration from the versatility of the BSUM framework (Razaviyayn et al., 2013) and the acceleration effect observed in TITAN (Hien et al., 2023), we introduce BMMe, which stands for Block Majorization Minimization with Extrapolation, to address Problem (1). Leveraging the multi-convex structure in Problem (1), BMMe does not need the NSDP condition to ensure convergence; instead, block majorization minimization for the multi-convex Problem (1) is reformulated as a block mirror descent method, wherein the Bregman divergence is adaptively updated at each iteration, and the extrapolation parameters in BMMe are dynamically updated using a novel adaptive rule. We establish subsequential convergence for BMMe, apply BMMe to tackle β-NMF problems with $\beta \in [1, 2]$, and showcase the obtained acceleration effects through extensive numerical experiments.

To give an idea of the simplicity and acceleration of BMMe, let us show how it works for β-NMF. Let (W, H) and (W^p, H^p) be the current and previous iterates, respectively. BMMe will provide the following MU with extrapolation (MUe):

$$\hat{H} = H + \alpha_H [H - H^p]_+, \quad H \leftarrow \mathrm{MU}(X, W, \hat{H}), \tag{4}$$

and similarly for W. We will show that MUe not only allows us to empirically accelerate the convergence of the MU significantly for a negligible additional cost per iteration and a slight modification of the original MU, but has convergence guarantees (Theorem 3.1). We will discuss in details how to choose the extrapolation parameters α_W and α_H in Section 3. It is important to note that no restarting step is required to ensure convergence. As a result, there is no need to compute objective function values during the iterative process, which would otherwise incur significant computational expenses. We will also show how to extend the MUe to regularized and constrained β-NMF problems in Section 4.3. Figure 1 illustrates the behavior of MU vs. MUe on the widely used CBCL facial image data set with $r = 49$, as in the seminal paper of Lee & Seung (1999) who introduced NMF, and with $\beta = 3/2$. MUe is more than twice faster than MU: over 10 random initializations, it takes MUe between 88 and 95 iterations with a median of 93 to obtain an objective smaller than the MU with 200 iterations. We will provide more experiments in Section 5 that confirm the significant acceleration effect of MUe.

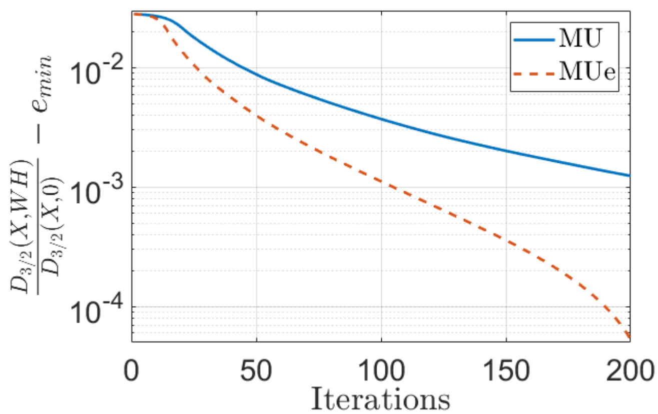

Figure 1. MU vs. MUe with Nesterov extrapolation sequence (15) on the CBCL data set with $r = 49$. Evolution of the median relative objective function values, over 10 random initial initializations, of β-NMF for $\beta = 3/2$ minus the smallest relative objective function found among all runs (denoted e_{min}).

The paper is organized as follows. In the next section, we provide preliminaries on majorizer functions, the majorization-minimization method, the multiplicative updates for β-NMF, and nonconvex optimizations. In Section 3, we describe our new proposed method, BMMe, and prove its convergence properties. In Section 4, we apply BMMe to solve standard β-NMF, as well as an important regularized and constrained KL-NMF model, namely the minimum-volume KL-NMF. We report numerical results in Section 5, and conclude the paper in Section 6.

Notation We denote $[s] = \{1, \ldots, s\}$. We use $I_{\mathcal{X}}$ to denote the indicator function associated to the set $\mathcal{X}$. For a given matrix X, we denote by $X_{:j}$ and $X_{i:}$ the j-th column and the i-th row of X, respectively. We denote the nonnegative part of X as $[X]_+ = \max(0, X)$ where the max is taken component wise. Given a multiblock differentiable function $f : x = (x_1, \ldots, x_s) \in \mathbb{E} \mapsto f(x)$, we use $\nabla_i f(x)$ to denote its partial derivative $\frac{\partial f(x)}{\partial x_i}$. We denote by e the vector of all ones of appropriate dimension.

2. Preliminaries

2.1. Majorizer, majorization-minimization method, and application to β-NMF

We adopt the following definition for a majorizer.

Definition 2.1 (Majorizer)**.** A continuous function $g : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is called a majorizer (or a surrogate function) of a differentiable function f over $\mathcal{Y}$ if the following conditions are satisfied:

(a) $g(x, x) = f(x)$ for all $x \in \mathcal{Y}$,

(b) $g(x, y) \geq f(x)$ for all $x, y \in \mathcal{Y}$, and

(c) $\nabla_1 g(x, x) = \nabla f(x)$ for all $x \in \mathcal{Y}$.

It is important noting that Condition (c) can be replaced by the condition on directional derivatives as in (Razaviyayn et al., 2013, Assumption 1 (A3)), and the upcoming analysis still holds. For simplicity, we use Condition (c) in this paper. Let us give some examples of majorizers. The second and third one will play a pivotal role in this paper. More examples of majorizers can be found in (Mairal, 2013; Sun et al., 2017; Hien et al., 2023).

1. Lipschitz gradient majorizer (see, e.g., Xu & Yin, 2013). If ∇f is L-Lipschitz continuous over $\mathcal{Y}$, then

$$g(x, y) = f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2$$

is called the Lipschitz gradient majorizer of f.

2. Bregman majorizer (see, e.g., Melo & Monteiro, 2017). Suppose there exists a differentiable convex function κ and $L > 0$ such that $x \mapsto L\kappa(x) - f(x)$ is convex. Then

$$\begin{aligned} g(x, y) = f(y) &+ \langle \nabla f(y), x - y\rangle \\ &+ L\big(\kappa(x) - \kappa(y) - \langle \nabla\kappa(y), x - y\rangle\big) \end{aligned}$$

is called a Bregman majorizer of f with kernel function κ. When $\kappa = \frac{1}{2}\|.\|^2$, the Breman majorizer coincides with the Lipschitz gradient majorizer.

3. Jensen majorizer (see, e.g., Dempster et al., 1977; Neal & Hinton, 1998; Lange et al., 2000). Suppose $\tilde{f} : \mathbb{R} \to \mathbb{R}$ is a convex function and $\omega \in \mathbb{R}^r$ is a given vector. Define $f : x \in \mathbb{R}^r \mapsto \tilde{f}(\omega^\top x)$. Then

$$g(x, y) = \sum_{i=1}^{r} \alpha_i \tilde{f}\left(\frac{\omega_i}{\alpha_i}(x_i - y_i) + \omega^\top y\right),$$

where $\alpha_i \geq 0$, $\sum_{i=1}^r \alpha_i = 1$, and $\alpha_i \neq 0$ whenever $\omega_i \neq 0$, is called a Jensen majorizer of f. The term "Jensen" in the name of the majorizer comes from the fact that the Jensen inequality for convex functions is used to form the majorizer. Indeed, by the Jensen inequality,

$$\sum_{i=1}^{r} \alpha_i \tilde{f}\left(\frac{\omega_i}{\alpha_i}(x_i - y_i) + \omega^\top y\right) \geq \tilde{f}\left(\sum_{i=1}^{r} \alpha_i [\frac{\omega_i}{\alpha_i}(x_i - y_i) + \omega^\top y]\right) = \tilde{f}(\omega^\top x).$$

Choosing $\alpha_i = \frac{\omega_i y_i}{\omega^\top y}$, $g(x, y) = \sum_{i=1}^r \frac{\omega_i y_i}{\omega^\top y} \tilde{f}\big(\frac{\omega^\top y}{y_i} x_i\big)$ is an example of a Jensen surrogate of f, if g is well-defined.

Given a majorizer of f, the minimization of f over $\mathcal{X}$ can be achieved by iteratively minimizing its majorizer, using

$$x^{t+1} \in \operatorname*{argmin}_{x \in \mathcal{X}} g(x, x^t),$$

where x^t denotes the t-th iterate. This is the majorization-minimization (MM) method which guarantees, by properties of the majorizer, that $f(x^{t+1}) \leq f(x^t)$ for all t; see Hunter & Lange (2004); Sun et al. (2017) for tutorials.

Example with the Multiplicative Updates for β-NMF The standard MU for β-NMF, given in (3), can be derived using the MM method. By symmetry of the problem, let us focus on the update of H. Moreover, we have that $D_\beta(X, WH) = \sum_i D_\beta(X_{:i}, WH_{:i})$, that is, the objective function is separable w.r.t. each column of H, and hence one can focus w.l.o.g. on the update of a single column of H. Let us thefore provide a majorizer for $D_\beta(v, Wh)$, and show how its closed-form solution leads to the MU (3).

The following proposition, which is a corollary of (Févotte & Idier, 2011, Theorem 1), provides a Jensen majorizer for $h \in \mathbb{R}^r \mapsto D_\beta(v, Wh) := \sum_{i=1}^m d_\beta(v_i, (Wh)_i)$ with $\beta \in [1, 2]$, where $v \in \mathbb{R}^m_+$ and $W \in \mathbb{R}^{m\times r}_+$ are given.

Proposition 2.2. *Let us denote $\tilde{v} = W\tilde{h}$, and let $\tilde{h}$ be such that $\tilde{v}_i > 0$ and $\tilde{h}_i > 0$ for all i. Then the following function is a majorizer for $h \mapsto D_\beta(v, Wh)$ with $\beta \in [1, 2]$:*

$$g(h, \tilde{h}) = \sum_{i=1}^{m} \sum_{k=1}^{r} \frac{W_{ik}\tilde{h}_k}{\tilde{v}_i} d_\beta\left(v_i, \tilde{v}_i \frac{h_k}{\tilde{h}_k}\right). \tag{5}$$

With this surrogate, we obtain the MU via the MM method (Févotte & Idier, 2011, Eq. 4.1) as follows:

$$\operatorname{argmin}_{x \geq \varepsilon} g(x, \tilde{h}) \;=\; \max\left(\varepsilon, \tilde{h} \circ \frac{\left[W^\top \frac{[v]}{[W\tilde{h}]^{\cdot(2-\beta)}}\right]}{\left[W^\top [W\tilde{h}]^{\cdot(\beta-1)}\right]}\right),$$

which leads to the MU in the matrix form (3). The term *multiplicative* in the name of the algorithm is because the new iterate is obtained by an element-wise multiplication between the current iterate, $\tilde{h}$, and a correction factor.

2.2. Critical point and coordinate-wise minimizer

Let us define three key notions for our purpose: subdifferential, critical point and coordinate-wise minimizer.

Definition 2.3 (Subdifferentials)**.** Let $g : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

(i) For each $x \in \operatorname{dom} g$, we denote $\hat{\partial} g(x)$ as the Fréchet subdifferential of g at x which contains vectors $v \in \mathbb{E}$ satisfying

$$\liminf_{y \neq x, y \to x} \frac{1}{\|y - x\|} \left(g(y) - g(x) - \langle v, y - x\rangle\right) \geq 0.$$

If $x \notin \operatorname{dom} g$, then we set $\hat{\partial} g(x) = \emptyset$.

(ii) The limiting-subdifferential $\partial g(x)$ of g at $x \in \operatorname{dom} g$ is defined as follows:

$$\partial g(x) := \left\{v \in \mathbb{E} :\ \exists x^k \to x,\ g\big(x^k\big) \to g(x),\ v^k \in \hat{\partial} g\big(x^k\big),\ v^k \to v\right\}.$$

Partial subdifferentials with respect to a subset of the variables are defined analogously by considering the other variables as parameters.

Definition 2.4 (Critical point)**.** We call $x^* \in \operatorname{dom} \mathrm{g}$ a critical point of g if $0 \in \partial g\left(x^*\right)$.

If x^* is a local minimizer of g then it is a critical point of g.

Definition 2.5 (Coordinate-wise minimizer)**.** We call $x^* \in \operatorname{dom} f$ a coordinate-wise minimizer of Problem (1) if

$$f(x_1^*, \ldots, x_{i-1}^*, x_i^*, x_{i+1}^*, \ldots, x_s^*) \;\leq f(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_s^*), \forall\, x_i \in \mathcal{X}_i,$$

or, equivalently, $\langle \nabla_i f(x^*), x_i - x_i^*\rangle \geq 0, \forall x_i \in \mathcal{X}_i$.

For Problem (1), a critical point of $f(x) + \sum_{i=1}^s I_{\mathcal{X}_i}(x_i)$, must be a coordinate-wise minimizer.

3. Block Majorization Minimization with Extrapolation (BMMe)

In this section, we introduce BMMe, see Algorithm 1 (top of the next page), and then prove its convergence, see Theorem 3.1.

3.1. Description of BMMe

BMMe, see Algorithm 1, updates one block of variables at a time, say x_i, by minimizing a majorizer G_i^t of

$$f_i^t(x_i) := f\big(x_1^{t+1}, \ldots, x_{i-1}^{t+1}, x_i, x_{i+1}^t, \ldots, x_s^t\big), \tag{8}$$

where the other blocks of variables $\{x_j\}_{j \neq i}$ are fixed, and t is the iteration index. The main difference between BMMe and standard block MM (BMM) is that the majorizer in BMMe is evaluated at the extrapolated block, $\hat{x}_i^t$ given in (6), while it is evaluated at the previous iterate x_i^t in BMM. The MUe (4) described in Section 1.1 follow exactly this scheme; we elaborate more on this specific case in Section 4.2.

In the following, we explain the notation that will be used in the sequel, and then state the convergence of BMMe.

• Denote $\bar{f}_i(x_i) := f(\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_s)$, where $\bar{x}$ is fixed. As assumed, the function $\bar{f}_i(\cdot)$, for $i \in [s]$, is convex and admits a majorizer $G_i^{(\bar{x})}(\cdot, \cdot)$ over its domain.

Algorithm 1 BMMe for solving Problem (1)

1: Choose initial points $x^{-1}, x^0 \in \mathrm{dom}(f)$. (Typically, $x^{-1} = x^0$.)
2: **for** $t = 0, \ldots$ **do**
3: **for** $i = 1, \ldots, s$ **do**
4: Extrapolate block i:

$$\hat{x}_i^t \; = \; x_i^t \, + \, \alpha_i^t \, \mathcal{P}_i \left(x_i^t - x_i^{t-1} \right), \quad \text{where} \tag{6}$$

• $\mathcal{P}_i$ is a mapping such that $\hat{x}_i^t \in \mathrm{dom}(f_i^t)$, where f_i^t is defined in (8) (for example, in β-NMF, we use $\mathcal{P}(a) = [a]_+$, see Section 4),
• α_i^t are the extrapolation parameters; see Section 3.2 for the conditions they need to satisfy.

5: Update block i:

$$x_i^{t+1} = \mathrm{argmin}_{x_i \in \mathcal{X}_i} \, G_i^t(x_i, \hat{x}_i^t), \tag{7}$$

where G_i^t is a majorizer of f_i^t over its domain.

6: **end for**
7: **end for**

• Given $\bar{x}$ and $\tilde{x}_i$, we denote $\xi_i^{(\bar{x}, \tilde{x}_i)}(x_i) = G_i^{(\bar{x})}(x_i, \tilde{x}_i)$ (i.e., we fix $\bar{x}$ and $\tilde{x}_i$) and

$$\begin{aligned} \mathcal{D}_{\bar{x}, \tilde{x}_i}(x_i, x_i') &= \mathcal{B}_{\xi_i}(x_i, x_i') \\ &= \xi_i(x_i) - \xi_i(x_i') - \langle \nabla \xi_i(x_i'), x_i - x_i' \rangle, \end{aligned}$$

where we omit the upperscript of $\xi_i^{(\bar{x}, \tilde{x}_i)}$ for notation succinctness. Using this notation, we can write

$$G_i^{(\bar{x})}(x_i, \tilde{x}_i) = \bar{f}_i(\tilde{x}_i) + \langle \nabla \bar{f}_i(\tilde{x}_i), x_i - \tilde{x}_i \rangle + \mathcal{D}_{\bar{x}, \tilde{x}_i}(x_i, \tilde{x}_i), \tag{9}$$

where we use the facts that $\bar{f}_i(\tilde{x}_i) = G_i^{(\bar{x})}(\tilde{x}_i, \tilde{x}_i)$ and $\nabla \bar{f}_i(\tilde{x}_i) = \nabla_1 G_i^{(\bar{x})}(\tilde{x}_i, \tilde{x}_i)$.

• Denote $x^{(t,i)} = (x_1^{t+1}, \ldots, x_{i-1}^{t+1}, x_i^t, x_{i+1}^t, \ldots, x_s^t)$, $G_i^t = G_i^{(x^{(t,i)})}$ be the majorizer of f_i^t, which is the notation used in Algorithm 1, and $\mathcal{D}_{\hat{x}_i}^t = \mathcal{D}_{x^{(t,i)}, \hat{x}_i}$.

Key observation for BMMe. Using the notation in (9), the MM update in (7) can be rewritten as

$$x_i^{t+1} \in \underset{x_i \in \mathcal{X}_i}{\mathrm{argmin}} f_i^t(\hat{x}_i) + \langle \nabla f_i^t(\hat{x}_i), x_i - \hat{x}_i \rangle + \mathcal{D}_{\hat{x}_i}^t(x_i, \hat{x}_i), \tag{10}$$

which has the form of a mirror descent step with the Bregman divergence $\mathcal{D}_{\hat{x}_i}^t(x_i, \hat{x}_i)$ being adaptively updated at each iteration. This observation will be instrumental in proving the convergence of BMMe.

3.2. Convergence of BMMe

We present the convergence of BMMe in Theorem 3.1.

Theorem 3.1. *Consider BMMe described in Algorithm 1 for solving Problem (1). We assume that the function $x_i \mapsto G_i^{(\bar{x})}(x_i, \tilde{x}_i)$ is convex for any given $\bar{x}$ and $\tilde{x}_i$. Furthermore, we assume the following conditions are satisfied.*

(C1) ***Continuity.*** *For $i \in [s]$, if $x^{(t,i)} \to \bar{x}$ when $t \to \infty$ then $G_i^t(\grave{x}_i^t, \acute{x}_i^t) \to G_i^{(\bar{x})}(\grave{x}_i, \acute{x}_i)$ for any $\grave{x}_i^t \to \grave{x}_i$ and $\acute{x}_i^t \to \acute{x}_i$, and $G_i^{(\bar{x})}(\cdot, \cdot)$ is a majorizer of $\bar{f}_i(\cdot)$.*

(C2) ***Implicit Lipschitz gradient majorizer.*** *At iteration t of Algorithm 1, for $i \in [s]$, there exists a constant $C_i > 0$ such that*

$$\begin{aligned} \mathcal{D}_{\hat{x}_i^t}^t(x_i^t, \hat{x}_i^t) \quad &\leq C_i \| x_i^t - \hat{x}_i^t \|^2 \\ &= C_i (\alpha_i^t)^2 \| \mathcal{P}_i(x_i^t - x_i^{t-1}) \|^2. \end{aligned} \tag{11}$$

(C3) The sequence of extrapolation parameters satisfies

$$\sum_{t=1}^{\infty} (\alpha_i^t)^2 \left\| \mathcal{P}_i(x_i^t - x_i^{t-1}) \right\|^2 < +\infty, \text{for } i \in [s]. \tag{12}$$

Then we have

$$\sum_{t=1}^{\infty}\sum_{i=1}^{s}\mathcal{D}^t_{\hat{x}^t_i}(x^t_i, x^{t+1}_i) < +\infty, \tag{13}$$

and the sequence generated by BMMe (Algorithm 1), $\{x^t\}_{t\geq 0}$, is bounded if f has bounded level sets.

Furthermore, under the following condition,

(C4) $\lim_{t\to\infty}\|x^t_i - x^{t+1}_i\| = 0$ *when* $\lim_{t\to\infty}\mathcal{D}^t_{\hat{x}^t_i}(x^t_i, x^{t+1}_i) = 0,$

any limit point of $\{x^t\}_{t\geq 0}$ is a critical point of $f(x) + \sum_{i=1}^{s} I_{\mathcal{X}_i}(x_i)$.

Let us discuss the conditions in Theorem 3.1:

• If $\mathcal{D}_{\bar{x},\tilde{x}_i}(x_i, \tilde{x}_i) \leq C_i\|x_i - \tilde{x}_i\|^2$ then (11) is satisfied. This condition means $\bar{f}_i(\cdot)$ is actually upper bounded by a Lipschitz gradient majorizer. However, it is crucial to realize that the introduction of C_i is primarily for the purpose of the convergence proof. Employing a Lipschitz gradient majorizer for updating x_i is discouraged due to the potential issue of C_i being excessively large (this situation can result in an overly diminishing/small step size, rendering the approach inefficient in practical applications). For example, $C_i = O(1/\varepsilon^2)$ in the case of KL-NMF, see the proof of Theorem 4.3.

• If $G^t_i(\cdot, \hat{x}^t_i)$ is θ_i-strongly convex then Condition (C4) is satisfied (here θ_i is a constant independent of $\{x^{(t,i)}\}$ and $\{\hat{x}^t_i\}$), since θ_i-strongly convexity implies

$$\mathcal{D}^t_{\hat{x}^t_i}(x^t_i, x^{t+1}_i) \geq \frac{\theta_i}{2}\|x^t_i - x^{t+1}_i\|^2.$$

• We see that the update in (10) has the form of an accelerated mirror descent (AMD) method (Tseng, 2008) for one-block convex problem. Hence, it makes sense to involve the extrapolation sequences that are used in AMD. This strategy has been used in (Xu & Yin, 2013; Hien et al., 2020; 2023). An example of choosing the extrapolation parameters satisfying (12) is

$$\alpha^t_i = \min\left\{\alpha^t_{Nes}, \frac{c}{t^{q/2}}\frac{1}{\|\mathcal{P}_i(x^t_i - x^{t-1}_i)\|}\right\}, \tag{14}$$

where $q > 1$, c is any large constant and α^t_{Nes} is the extrapolation sequence defined by $\eta_0 = 1$,

$$\eta_t = \frac{1}{2}\Big(1 + \sqrt{1 + 4\eta^2_{t-1}}\Big),\ \alpha^t_{Nes} = \frac{\eta_{t-1} - 1}{\eta_t}. \tag{15}$$

Another choice is replacing α^t_{Nes} in (14) by $\alpha^t_c = \frac{t-1}{t}$ (Tseng, 2008). In our experiments, we observe that when c is large enough, α^t_i coincides with α^t_{Nes} (or α^t_c).

4. Application of BMMe to β-NMF

Before presenting the application of BMMe to the standard β-NMF problem with $\beta \in [1, 2]$ (Section 4.2), and a constrained and regularized KL-NMF problem (Section 4.3), we briefly discuss the majorizers for the β divergences.

4.1. Majorizer of the β divergence, $\beta \in [1, 2]$

Recall that the function defined in (5) is a majorizer for $h \mapsto D_\beta(v, Wh)$, where $\tilde{v} = W\tilde{h}$, see Proposition 2.2. The function $g(\cdot,\cdot)$ is twice continuously differentiable over $\{(h, \tilde{h}) : h \geq \varepsilon, \tilde{h} \geq \varepsilon\}$, and

$$\nabla^2_{h_k} g(h, \tilde{h}) = \sum_{i=1}^{m}\frac{W_{ik}\tilde{v}_i}{\tilde{h}_k} d''_\beta\left(v_i, \tilde{v}_i\frac{h_k}{\tilde{h}_k}\right), \tag{16}$$

where $d''_\beta(x, y)$ denotes the second derivative with respect to y of $(x, y) \mapsto d_\beta(x, y)$. On the other hand, as already noted in Section 2.1 , $D_\beta(X, WH) = \sum_{j=1}^{n} D_\beta(X_{:j}, WH_{:j}) = \sum_{i=1}^{m} D_\beta(X^\top_{i:}, H^\top W^\top_{i:})$. Hence a majorizer of $H \mapsto D_\beta(X, WH)$ while fixing W is given by

$$G^{(W)}_2(H, \tilde{H}) = \sum_{j=1}^{n} g^{(W)}_j(H_{:j}, \tilde{H}_{:j}), \tag{17}$$

where $g_j^{(W)}(H_{:j}, \tilde{H}_{:j})$ is the majorizer of $H_{:j} \mapsto D_\beta(X_{:j}, WH_{:j})$, which is defined as in (5) with $v = X_{:j}$. Similarly, the following function is a majorizer of $W \mapsto D_\beta(X, WH)$ while fixing H

$$G_1^{(H)}(W, \tilde{W}) = \sum_{i=1}^{m} \mathbf{g}_i^{(H)}(W_{i:}^\top, \tilde{W}_{i:}^\top), \tag{18}$$

where $\mathbf{g}_i^{(H)}(W_{i:}^\top, \tilde{W}_{i:}^\top)$ is the majorizer of $W_{i:}^\top \mapsto D_\beta(X_{i:}^\top, H^\top W_{i:}^\top)$ defined as in (5) with v being replaced by $X_{i:}^\top$ and W being replaced by $H^\top$.

Note that there exists other majorizers for β-NMF, for example majorizers for both variables simultaneously (Marmin et al., 2023), and quadratic majorizers for KL divergence (Pham et al., 2023).

4.2. MU with extrapolation for β-NMF, $\beta \in [1, 2]$

We consider the standard β-NMF problem in (2) with $\beta \in [1, 2]$. Applying Algorithm 1 to solve Problem (2), we get MUe, a multiplicative update method with extrapolation described in Algorithm 2.

Algorithm 2 MUe for solving β-NMF (2)

1: Choose initial points $(W^{-1}, W^0, H^{-1}, H^0) \geq \varepsilon > 0$.
2: **for** $t = 0, \ldots$ **do**
3: Compute extrapolation points:

$$\hat{W}^t = W^t + \alpha_W^t [W^t - W^{t-1}]_+,$$
$$\hat{H}^t = H^t + \alpha_H^t [H^t - H^{t-1}]_+,$$

where α_W^t and α_H^t satisfy (19).
4: Update the two blocks of variables:

$$W^{t+1} = \underset{W \geq \varepsilon}{\operatorname{argmin}} G_1^t(W, \hat{W}^t) = \text{MU}\big(X^\top, (H^t)^\top, (\hat{W}^t)^\top\big)^\top \text{ [see (3)]},$$
$$H^{t+1} = \underset{H \geq \varepsilon}{\operatorname{argmin}} G_2^t(H, \hat{H}^t) = \text{MU}\big(X, W^{t+1}, \hat{H}^t\big),$$

where $G_1^t = G_1^{(H^t)}$ and $G_2^t = G_2^{(W^{t+1})}$ be the majorizers defined in (18) with $H = H^t$ and (17) with $W = W^{t+1}$, respectively.
5: **end for**

Theorem 4.1. *Suppose the extrapolation parameters in Algorithm 2 are chosen such that they are bounded and*

$$\begin{aligned} &\sum_{t=1}^{\infty} (\alpha_H^t)^2 \big\| [H^t - H^{t-1}]_+ \big\|^2 < +\infty, \text{ and} \\ &\sum_{t=1}^{\infty} (\alpha_W^t)^2 \big\| [W^t - W^{t-1}]_+ \big\|^2 < +\infty. \end{aligned} \tag{19}$$

Then MUe (Algorithm 2) generates a bounded sequence and any of its limit point, (W^*, H^*), is a KKT point of Problem (2), that is,

$$W^* \geq \varepsilon, \nabla_W D_\beta(X, W^* H^*) \geq 0,$$
$$\langle \nabla_W D_\beta(X, W^* H^*), W^* - \varepsilon e e^\top \rangle = 0,$$

and similarly for H^*.

4.3. MUe for constrained and regularized KL-NMF

The KL divergence is especially relevant when the statistical characteristics of the observed data samples conform to a Poisson distribution, turning KL-NMF into a meaningful choice for count data sets such as images (Richardson, 1972), documents (Lee & Seung, 1999), and single-cell sequencing (Carbonetto et al., 2023). In numerous scenarios, there are specific additional constraints and regularizers to add to KL-NMF. For example, the minimum-volume (min-vol) KL-NMF, which incorporates a regularizer encouraging the columns of matrix W to have a small volume, along with a normalization constraints (such as $H^\top e = e$ or $W^\top e = e$), enhances identifiability/uniqueness (Lin et al., 2015; Fu et al., 2015), a crucial aspect in various applications such as hyperspectral imaging (Miao & Qi, 2007) and audio source separation (Leplat et al., 2020). In this section, we consider the following general regularized KL-NMF problem

$$\min_{W\in\bar{\Omega}_W, H\in\bar{\Omega}_H} \Big\{ f(W,H) := D_{KL}(X,WH) + \lambda_1\phi_1(W) + \lambda_2\phi_2(H) \Big\}, \tag{20}$$

where $\bar{\Omega}_W := \big\{W : W \geq \varepsilon, W \in \Omega_W\big\}$ and $\bar{\Omega}_H := \big\{H : H \geq \varepsilon, H \in \Omega_H\big\}$. We assume that there exist continuous functions $L_{\phi_1}(\tilde{W}) \geq 0$ and $L_{\phi_2}(\tilde{H}) \geq 0$ such that for all $W, \tilde{W} \in \{W : W \geq 0, W \in \Omega_W\}$ and $H, \tilde{H} \in \{H : H \geq 0, H \in \Omega_H\}$, we have

$$\begin{aligned} \phi_1(W) &\leq \bar{\phi}_1(W,\tilde{W}) := \phi_1(\tilde{W}) + \langle \nabla\phi_1(\tilde{W}), W - \tilde{W}\rangle + \tfrac{L_{\phi_1}(\tilde{W})}{2}\|W - \tilde{W}\|^2, \\ \phi_2(H) &\leq \bar{\phi}_2(H,\tilde{H}) := \phi_2(\tilde{H}) + \langle \nabla\phi_2(\tilde{H}), H - \tilde{H}\rangle + \tfrac{L_{\phi_2}(\tilde{H})}{2}\|H - \tilde{H}\|^2. \end{aligned} \tag{21}$$

Furthermore, $L_{\phi_1}(\tilde{W})$ and $L_{\phi_2}(\tilde{H})$ in (21) are upper bounded by $\bar{L}_{\phi_1}$ and $\bar{L}_{\phi_2}$, respectively. We focus on the min-vol regularizer, $\phi_1(W) = \det(W^\top W + \delta I)$, and $\Omega_W = \{W \mid W^\top e = e\}$ (Leplat et al., 2020). In that case, $\phi_1(W)$ satisfies this condition, as proved in Lemma 4.2.

Lemma 4.2. *The function $\phi_1(W) = \log\det(W^\top W + \delta I)$ with $\delta > 0$ satisfies the condition in (21) with $L_{\phi_1}(\tilde{W}) = 2\|(\tilde{W}^\top\tilde{W} + \delta I)^{-1}\|_2$, which is upper bounded by $2/\delta^2$.*

We will use the following majorizer for $H \mapsto f(W,H)$ while fixing W

$$G_2^{(W)}(H,\tilde{H}) = \sum_{j=1}^{n} g_j^{(W)}(H_{:j}, \tilde{H}_{:j}) + \lambda_1\phi_1(W) + \lambda_2\bar{\phi}_2(H,\tilde{H}), \tag{22}$$

where $\bar{\phi}_2(H,\tilde{H})$ is defined in (21) and $g_j^{(W)}$ is defined as in (17). Similarly, we use the following majorizer for $W \mapsto f(W,H)$ while fixing H

$$G_1^{(H)}(W,\tilde{W}) = \sum_{i=1}^{m} \mathbf{g}_i^{(H)}(W_{i:}^\top, \tilde{W}_{i:}^\top) + \lambda_1\bar{\phi}_1(W,\tilde{W}) + \lambda_2\phi_2(H), \tag{23}$$

where $\bar{\phi}_1(W,\tilde{W})$ is defined in (21), and $\mathbf{g}_i^{(H)}$ is defined as in (18). Applying Algorithm 1 to solve Problem (20), we get Algorithm 3, a BMMe algorithm for regularized and constrained KL-NMF see its detailed description in Appendix A. It works exactly as Algorithm 2 but the majorizers $G_1^t = G_1^{(H^t)}$ and $G_2^t = G_2^{(W^{t+1})}$ are defined in (23) with $H = H^t$ and (22) with $W = W^{t+1}$, respectively.

Theorem 4.3. *Suppose the extrapolation parameters α_W^t and α_H^t satisfy (19). Then BMMe applied to Problem (20) (Algorithm 3) generates a bounded sequence and any of its limit point is a coordinate-wise minimizer of Problem (20).*

BMMe for solving Problem (20) (Algorithm 3) is not necessarily straightforward to implement, because its updates might not have closed forms. In the following, we derive such updates in the special case of min-vol KL-NMF (Leplat et al., 2020):

$$\begin{aligned} \min_{W\geq\varepsilon, H\geq\varepsilon} \; & D_{KL}(X,WH) + \lambda_1 \log\det(W^\top W + \delta I) \\ \text{such that} \quad & e^\top W_{:j} = 1, j = 1,\ldots,r. \end{aligned} \tag{24}$$

The update of H is as in (3) taking $\beta = 1$. The following lemma provides the update of W.

Lemma 4.4. *For notation succinctness, let $\hat{W} = W^t + \alpha_W^t[W^t - W^{t-1}]_+$ and $H = H^{t+1}$. BMMe for solving (24) (Algorithm 3) updates W as follows:*

$$W \leftarrow \max\left(\varepsilon, \frac{1}{2}\left(-B_2 + \Big[[B_2]^{.2} + 4\lambda_1 L_{\phi_1}(\hat{W})B_1\Big]^{.1/2}\right)\right), \tag{25}$$

where

$$\begin{aligned} &L_{\phi_1}(\hat{W}) = 2\|(\hat{W}^\top\hat{W} + \delta I)^{-1}\|_2, B_1 = \frac{[X]}{[\hat{W}H]}H^\top \circ \hat{W},\\ &B_2 = ee^\top H^\top + \lambda_1\big(A - L_{\phi_1}(\hat{W})\hat{W} + e\mu^\top\big),\\ &A = 2\hat{W}(\hat{W}^\top\hat{W} + \delta I)^{-1}, \quad \mu = (\mu_1, \ldots, \mu_r)^\top, \end{aligned}$$

and μ_k*, for* $k = 1, \ldots, r$*, is the unique solution of* $\sum_{j=1}^n W_{jk}(\mu_k) = 1$*. We can determine* μ_k *by using bisection method over* $\mu_k \in [\underline{\mu}_k, \overline{\mu}_k]$*, where*

$$\begin{aligned} \underline{\mu}_k &= \min_{j=1,\ldots,m} \tilde{\mu}_{jk}, \quad \overline{\mu}_k = \max_{j=1,\ldots,m} \tilde{\mu}_{jk},\\ \tilde{\mu}_{jk} &= \frac{1}{\lambda_1}\Big(4\lambda_1 L_{\phi_1}(\hat{W})b_1 m - 1/m - \textstyle\sum_{i=1}^n (H^\top)_{ik}\Big) + L_{\phi_1}(\hat{W})\hat{W}_{jk} - A_{jk}, \end{aligned} \tag{26}$$

with $b_1 = \sum_{i=1}^n \frac{(H^\top)_{ik}X_{ji}}{\tilde{v}_i}\hat{W}_{jk}$ *and* $\tilde{v} = H^\top\hat{W}_{j:}^\top$.

5. Numerical experiments

In this section, we show the empirical acceleration effect of BMMe. We use the Nesterov extrapolation parameters (15).

5.1. β-NMF for hyperspectral imaging

We consider β-NMF (2) with $\beta = 3/2$ which is among the best NMF models for hyperspectral unmixing (Févotte & Dobigeon, 2014). For this problem, the MU are the workhorse approach, and we compare it to MUe: Figure 2 provides the median evolution of objective function values for the Cuprite data set ($m = 188$, $n = 47750$, $r = 20$); see Appendix B.1 for more details and experiments on 3 other data sets. There is a significant acceleration effect: on average, MUe requires only 41 iterations to obtain a smaller objective than MU with 100 iterations.

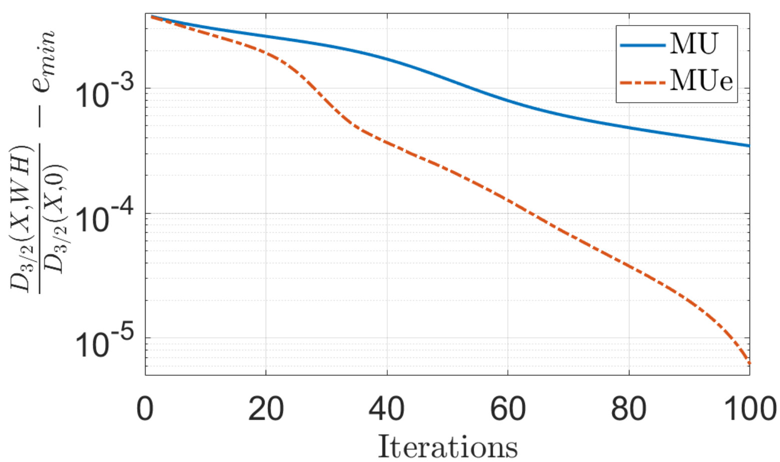

Figure 2. Median relative objective function of β-NMF for $\beta = 3/2$ minus the smallest objective function found among 10 random initializations (denoted e_{min}).

5.2. KL-NMF for topic modeling and imaging

We now consider KL-NMF which is the workhorse NMF model for topic modeling and also widely used in imaging; see Section 4.3. We compare MUe with MU and the cyclic coordinate descent (CCD) method of Hsieh & Dhillon (2011). As reported in (Hien & Gillis, 2021), MU and CCD are the state of the art for KL-NMF (sometimes one perform the best, sometimes the other). Figure 3 shows the median evolution of the relative objective function for two data sets: a dense image data set (ORL, $m = 10304$, $n = 400$), and a sparse document data set (hitech, $m = 2301$, $n = 10080$). For ORL, CCD and MU perform similarly while MUe performs the best. For hitech, MUe and CCD perform similarly, while they outperform MU. In all cases, MUe provides a significant acceleration effect to MU. Similar observations hold for 6 other data sets; see Appendix B.2.

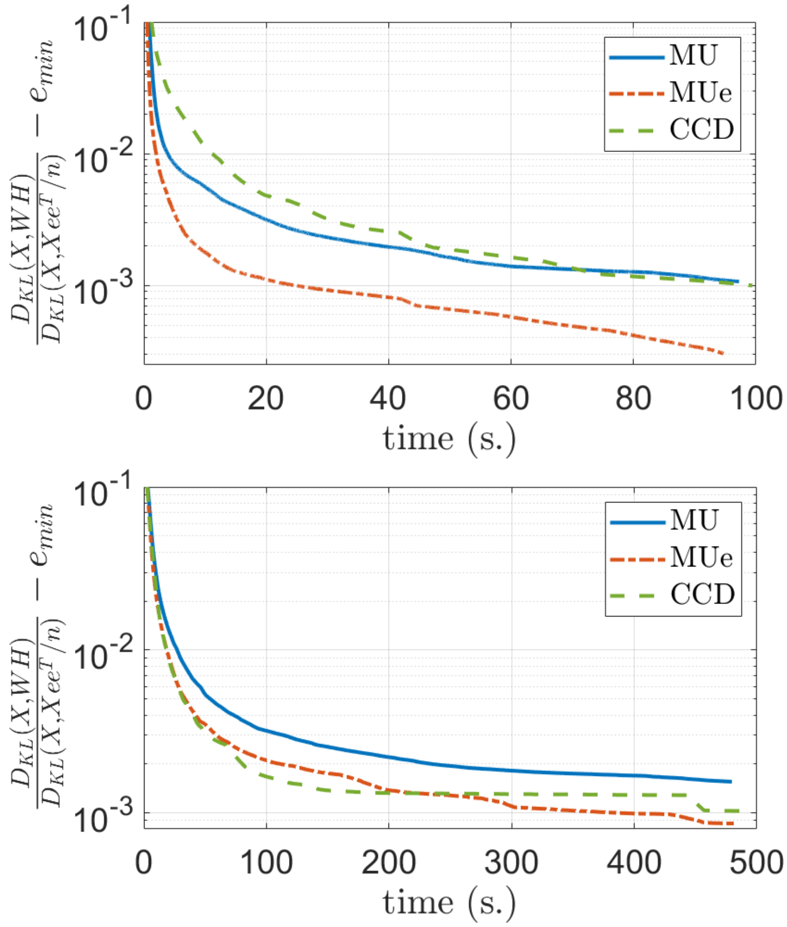

Figure 3. Median relative objective function of KL-NMF minus the smallest objective function found among 10 runs (denoted e_{min}): (top) ORL, (bottom) hitech, with $r = 10$ in both cases.

5.3. Min-vol KL-NMF for audio data sets

We address Problem (24) in the context of blind audio source separation (Leplat et al., 2020). We compare the following algorithms: MU with H update from (3) and W update from Lemma 4.4, its extrapolated variant, MUe, a recent MM algorithm by Leplat et al. (2021), denoted MM, and MMe incorporating the BMMe extrapolation step in MM.

Figure 4 displays the median relative objective function values,

$$e_{rel}(W, H) = \frac{D_{KL}(X, WH) + \lambda \log\det(WW + \delta I)}{D_{KL}(X, (Xe/n)e^\top)}, \tag{27}$$

minus the smallest relative objective found, for the prelude from J.S.-Bach ($m = 129$, $n = 2292$, $r = 16$); see Appendix B.3 for two other data sets, and more details. MUe exhibits accelerated convergence compared to MU. MM ranks second, while its new variant, MMe, integrating the proposed extrapolation, consistently achieves the best performance. This better performance stems from the nature of the majorizer employed for the logdet term which offers a more accurate approximation. It is worth noting that the majorizer used by MMe neither satisfies Condition (c) of Definition 2.1 nor Assumption 1 (A3) of Razaviyayn et al. (2013), and hence Theorem 3.1 does not apply to MMe.

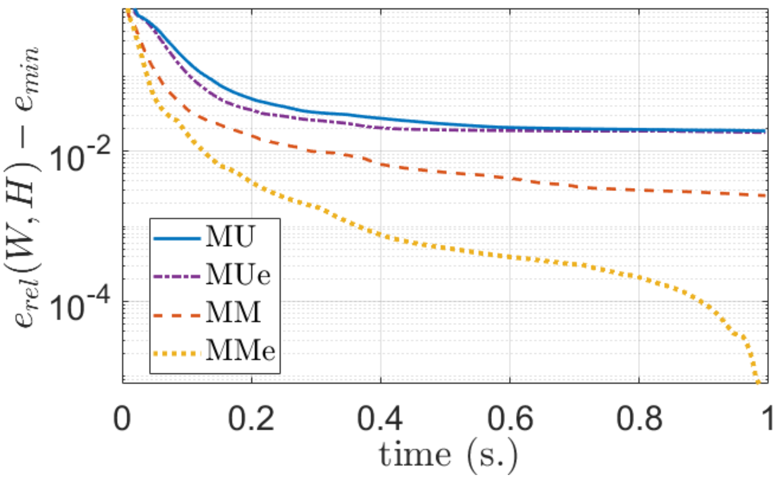

Figure 4. Evolution of the median relative errors (see (27)) minus the smallest relative error found among the 20 random initializations, of min-vol KL-NMF (24) of the four algorithms.

6. Conclusion and further work

In this paper, we considered multi-convex optimization (1). We proposed a new simple yet effective acceleration mechanism for the block majorization-minimization method, incorporating extrapolation (BMMe). We established subsequential convergence of BMMe, and leveraged it to accelerate multiplicative updates for various NMF problems. Through numerous numerical experiments conducted on diverse datasets, namely documents, images, and audio datasets, we showcased the remarkable acceleration impact achieved by BMMe. Further work include the use of BMMe for other applications, such as nonnegative tensor decompositions, and new theoretical developments, such as relaxing Condition (c) in Definition 2.1 (definition of a majorizer) or the condition on directional derivatives (Razaviyayn et al., 2013, Assumption 1 (A3)).

References

Ang, A. M. S. and Gillis, N. Accelerating nonnegative matrix factorization algorithms using extrapolation. *Neural Computation*, 31(2):417–439, 2019.

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

Beck, A. and Tetruashvili, L. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23:2037–2060, 2013.

Benetos, E., Dixon, S., Duan, Z., and Ewert, S. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, 2019.

Bertsekas, D. *Nonlinear Programming*. Athena Scientific, 2016.

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, April 2012.

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, Aug 2014.

Carbonetto, P., Luo, K., Sarkar, A., Hung, A., Tayeb, K., Pott, S., and Stephens, M. GoM DE: interpreting structure in sequence count data with differential expression analysis allowing for grades of membership. *Genome Biology*, 24(1): 236, 2023.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Févotte, C. and Dobigeon, N. Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization. *arXiv preprint arXiv:1401.5649*, 2014.

Févotte, C. and Idier, J. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9): 2421–2456, Sept 2011.

Févotte, C., Bertin, N., and Durrieu, J.-L. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. *Neural Computation*, 21(3):793–830, 03 2009.

Fu, X., Ma, W.-K., Huang, K., and Sidiropoulos, N. D. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing*, 63(9):2306–2320, 2015.

Grippo, L. and Sciandrone, M. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations Research Letters*, 26(3):127 – 136, 2000.

Hien, L., Phan, D., and Gillis, N. Inertial alternating direction method of multipliers for non-convex non-smooth optimization. *Computational Optimization and Applications*, 83:247–285, 2022.

Hien, L. T. K. and Gillis, N. Algorithms for nonnegative matrix factorization with the Kullback-Leibler divergence. *Journal of Scientific Computing*, (87):93, 2021.

Hien, L. T. K., Gillis, N., and Patrinos, P. Inertial block proximal method for non-convex non-smooth optimization. In *Thirty-seventh International Conference on Machine Learning (ICML)*, 2020.

Hien, L. T. K., Phan, D. N., and Gillis, N. An inertial block majorization minimization framework for nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 24(18):1–41, 2023.

Hsieh, C.-J. and Dhillon, I. S. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1064–1072, 2011.

Hunter, D. R. and Lange, K. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Lange, K., Hunter, D. R., and Yang, I. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

Lee, D. D. and Seung, H. S. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

Leplat, V., Gillis, N., and Ang, A. M. Blind audio source separation with minimum-volume beta-divergence NMF. *IEEE Transactions on Signal Processing*, 68:3400–3410, 2020.

Leplat, V., Gillis, N., and Idier, J. Multiplicative updates for NMF with beta-divergences under disjoint equality constraints. *SIAM Journal on Matrix Analysis and Applications*, 42(2):730–752, 2021.

Lin, C.-H., Ma, W.-K., Li, W.-C., Chi, C.-Y., and Ambikapathi, A. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5530–5546, 2015.

Ma, W., Bioucas-Dias, J. M., Chan, T., Gillis, N., Gader, P., Plaza, A. J., Ambikapathi, A., and Chi, C. A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. *IEEE Signal Processing Magazine*, 31(1):67–81, 2014.

Mairal, J. Optimization with first-order surrogate functions. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. 783–791, 2013.

Marmin, A., de Morais Goulart, J. H., and Févotte, C. Joint majorization-minimization for nonnegative matrix factorization with the β-divergence. *Signal Processing*, 209:109048, 2023.

Melo, J. G. and Monteiro, R. D. C. Iteration-complexity of a Jacobi-type non-Euclidean ADMM for multi-block linearly constrained nonconvex programs, 2017.

Miao, L. and Qi, H. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777, 2007.

Neal, R. M. and Hinton, G. E. *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*, pp. 355–368. Springer Netherlands, Dordrecht, 1998.

Nesterov, Y. *Lectures on Convex Optimization*. Springer, 2018.

Ochs, P. Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano. *SIAM Journal on Optimization*, 29(1):541–570, 2019.

Ochs, P., Chen, Y., Brox, T., and Pock, T. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

Pham, M.-Q., Cohen, J., and Chonavel, T. A fast multiplicative updates algorithm for non-negative matrix factorization. *arXiv preprint arXiv:2303.17992*, 2023.

Pock, T. and Sabach, S. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787, 2016.

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

Richardson, W. H. Bayesian-based iterative method of image restoration. *JoSA*, 62(1):55–59, 1972.

Smaragdis, P. and Brown, J. C. Non-negative matrix factorization for polyphonic music transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180. IEEE, 2003.

Sun, Y., Babu, P., and Palomar, D. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65:794–816, 02 2017.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, Jun 2001.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. Technical report, 2008.

Tseng, P. and Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, Mar 2009.

Vu Thanh, O., Ang, A., Gillis, N., and Hien, L. T. K. Inertial majorization-minimization algorithm for minimum-volume NMF. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1065–1069, 2021.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

Zhong, S. and Ghosh, J. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8:374–384, 2005.

Zhu, F. Hyperspectral unmixing: ground truth labeling, datasets, benchmark performances and survey. *arXiv preprint arXiv:1708.05125*, 2017.

A. BMMe for solving constrained and regularized KL-NMF (20)

Algorithm 3 is BMMe for the specific case of constrained and regularized KL-NMF.

Algorithm 3 BMMe for solving constrained and regularized KL-NMF (20)

1: Choose initial points $W^{-1} \geq \varepsilon, W^0 \geq \varepsilon, H^{-1} \geq \varepsilon, H^0 \geq \varepsilon$.
2: **for** $t = 1, \ldots$ **do**
3: Compute extrapolation points:

$$\begin{aligned} \hat{W}^t &= W^t + \alpha_W^t [W^t - W^{t-1}]_+, \\ \hat{H}^t &= H^t + \alpha_H^t [H^t - H^{t-1}]_+, \end{aligned}$$

where α_W^t and α_H^t satisfy the condition of Theorem 4.1.
4: Update the two blocks of variables:

$$\begin{aligned} W^{t+1} &\in \underset{W \in \bar{\Omega}_W}{\operatorname{argmin}} G_1^t(W, \hat{W}^t), \\ H^{t+1} &\in \underset{H \in \bar{\Omega}_H}{\operatorname{argmin}} G_2^t(H, \hat{H}^t), \end{aligned} \tag{28}$$

where $G_1^t = G_1^{(H^t)}$ and $G_2^t = G_2^{(W^{t+1})}$ are the majorizers defined in (23) with $H = H^t$ and (22) with $W = W^{t+1}$, respectively.
5: **end for**

B. Additional experiments

All experiments can be run using `https://github.com/vleplat/BMMe`. They have been performed on a laptop computer with Intel Core i7-11800H @ 2.30GHz and 16GB memory with MATLAB R2021b.

B.1. β-NMF with $\beta = 3/2$ for hyperspectral images

We consider β-NMF with $\beta = 3/2$ which has been shown to perform well for hyperspectral images; see Févotte & Dobigeon (2014). We use 4 widely used data sets summarized in Table 1; see, e.g., `http://lesun.weebly.com/hyperspectral-data-set.html` and Zhu (2017).

Data set	m	n	r
Urban	162	307×307	6
San Diego airport	158	400×400	8
Pines	142	145×145	16
Cuprite	188	250×191	20

Table 1. Summary of the hyperspectral image data sets.

A hyperspectral image (HSI) provides a spectral signature for each pixel of the image. The spectral signature measures the fraction of light reflected depending on the wavelength, and HSIs typically measure between 100 and 200 wavelengths. Given such an image, blind hyperspectral unmixing aims to decompose the image into pure materials and abundance maps (which indicates which pixel contains which material and in which proportion). One of the most successful model to perform this task is NMF applied on the wavelength-by-pixel matrix; see (Bioucas-Dias et al., 2012; Ma et al., 2014) and the references therein for more details.

Figure 5 displays the median relative objective function values minus the best solution found among 10 random initializations. Table 2 provides the minimum, median and maximum number of iterations for MUe to obtain an objective function value smaller than MU with 100 iterations, for the 10 random initializations. On average, MUe requires less than 50 iterations, meaning that MUe is more than twice faster than MU. MUe provides a significant acceleration over MU for all data sets and all initializations; in the worst case (for a total of 40 runs: 4 data sets with 10 random initializations each), it takes MUe 55 iterations to reach the error of MU with 100 iterations.

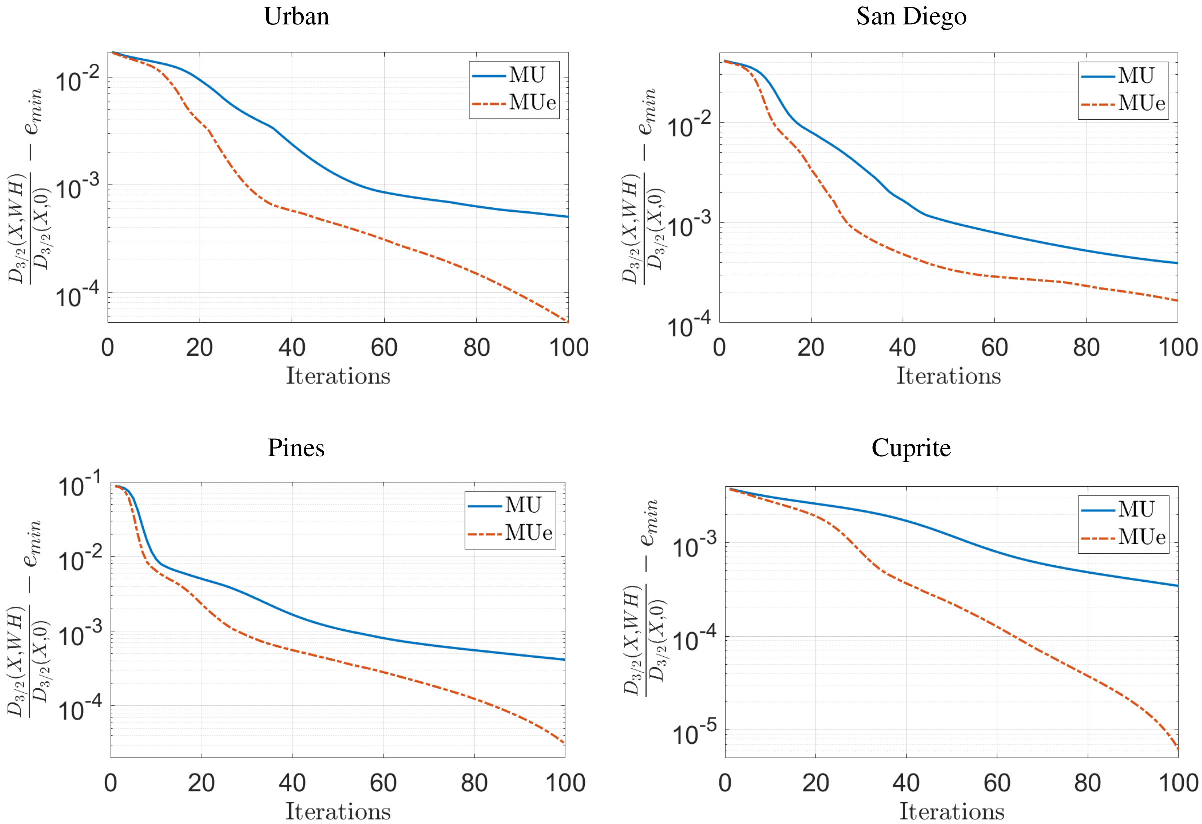

Figure 5. Evolution of the objective function of β-NMF for $\beta = 3/2$ minus the smallest objective function found among the 10 runs of the two algorithms (denoted e_{min}): MU vs. MUe on the four hyperspectral data sets ; see Table 1. We report the median over 10 randomly generated initial matrices for both algorithms.

Data set	min	meadian	max
San Diego	42	46	52
Urban	37	43	55
Cuprite	39	40.5	42
Pines	43	47	49

Table 2. Number of iterations needed for MUe to obtain an objective function value smaller than MU after 100 iterations. We report the minimum, median and maximum number over the 10 random initializations.

B.2. KL-NMF for topic modeling and imaging

As explained in Section 4.3, KL-NMF is widely used for imaging and topic modeling. In this section, we report extensive results for both applications, following the experimental setup of Hien & Gillis (2021) (in particular, we use $r = 10$ in all cases). We compare MU and MUe to the state-of-the-art algorithm CCD (Hsieh & Dhillon, 2011). In the extensive numerical experiments reported in Hien & Gillis (2021), MU and CCD were the best two algorithms for KL-NMF.

Dense facial images We use 4 popular facial image summarized in Table 3. Figure 6 displays the median relative objective function values minus the best solution found among 10 random initializations. For these dense data sets, MUe performs better than CCD, which was not the case of MU that performs on average worse than CCD on dense data sets (Hien & Gillis, 2021). This means that not only MUe provides a significant acceleration of MU, but also outperforms the state-of-the-art algorithm CCD for KL NMF on dense data sets.

Data set	m (# pixels)	n (# images)
CBCL	19×19	2429
Frey	28×20	1965
ORL	112×92	565
UMist	112×92	400

Table 3. Summary of the facial image data sets.

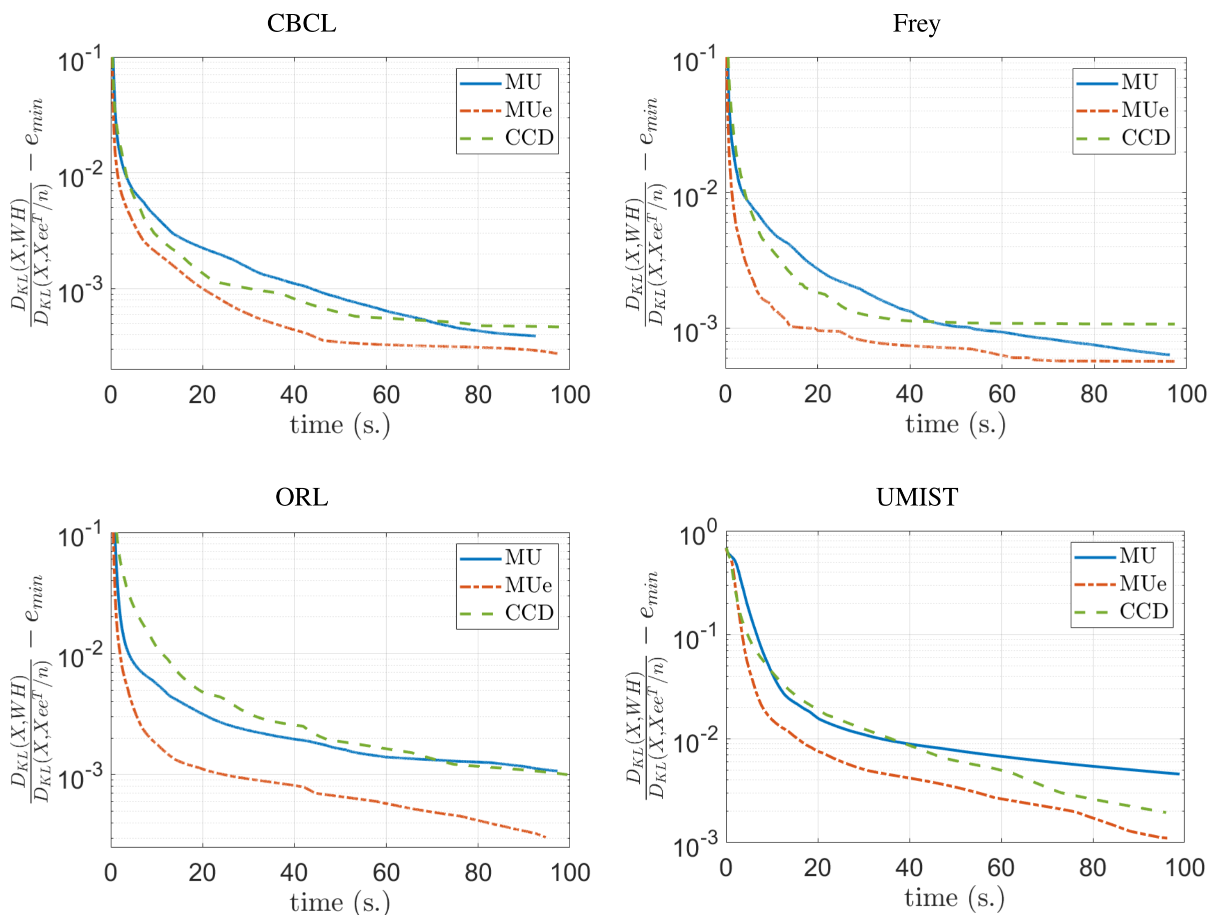

Figure 6. Evolution of the objective function of KL-NMF minus the smallest objective function found among the 10 runs of the three algorithms (denoted e_{min}) on the four facial image data sets. We report the median over 10 randomly generated initial matrices for the three algorithms.

Sparse document data sets We use 4 document data sets summarized in Table 3. Figure 7 displays the median relative

Data set	m (# documents)	n (# words)	sparsity (% zeros)
classic	7094	41681	99.92
hitech	2301	10080	98.57
la1	3204	31472	99.52
sports	8580	14870	99.14

Table 4. Summary of the document data sets; see (Zhong & Ghosh, 2005) for more details.

objective function values minus the best solution found among 10 random initializations. For sparse data sets, MUe outperforms MU, as in all our other experiments so far. However, MUe does not outperform CCD which performs better on two data sets (classic, la1). Hence, for sparse data sets, CCD is competitive with MUe, although no algorithm seem to

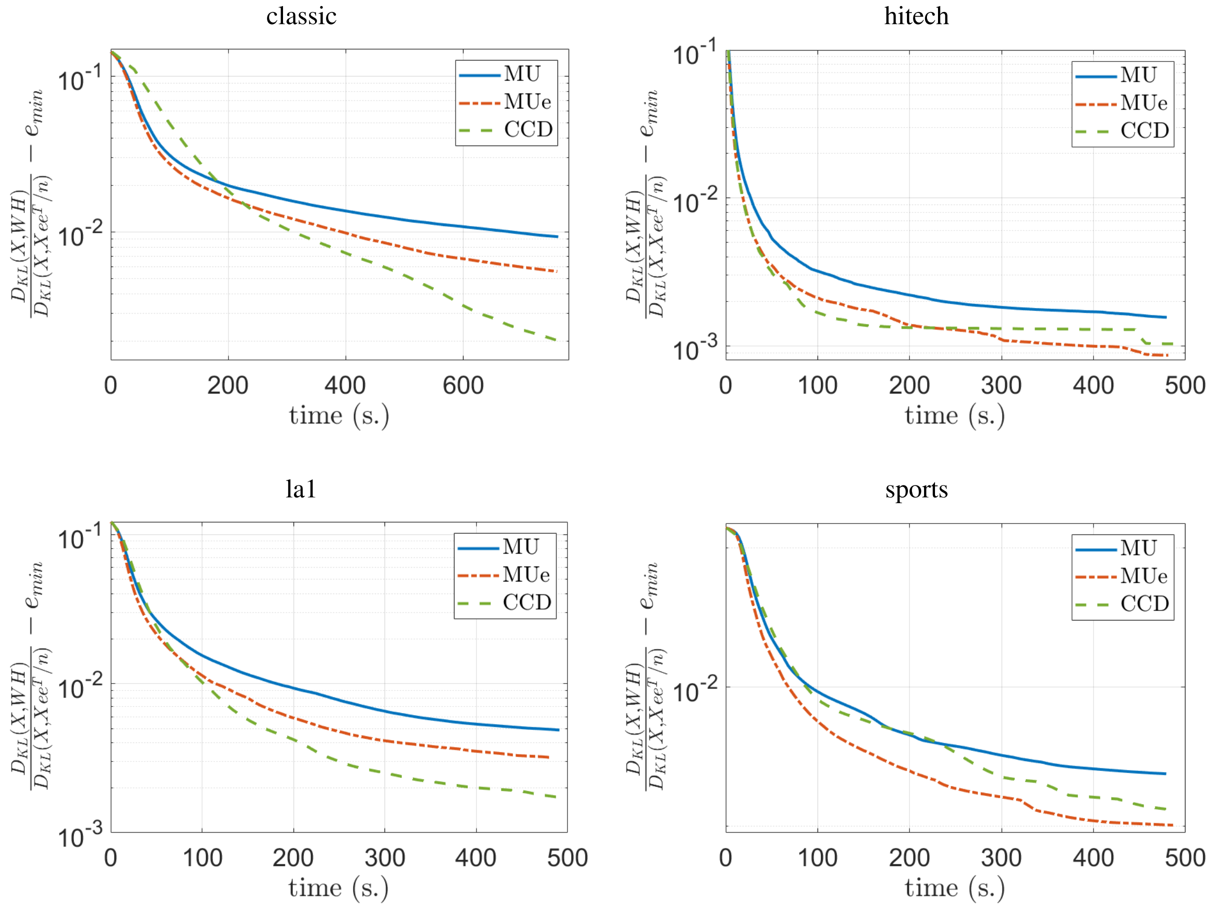

Figure 7. Evolution of the objective function of KL-NMF minus the smallest objective function found among the 10 runs of the three algorithms (denoted e_{min}) on the four facial image data sets. We report the median over 10 randomly generated initial matrices for the three algorithms.

outperform the other one.

B.3. Min-vol KL-NMF for audio data sets

NMF has been used successfully to perform blind audio source separation. The input matrix is a spectrogram that records the activations of the frequencies over time. Applying NMF on this spectrogram allows us to recover the frequency response of the sources as the columns of W, and the activations of the source over time as the rows of H (Smaragdis & Brown, 2003). For example, applying NMF on a piano recording will extract automatically the spectrum of the notes (their harmonics) and their activation over time, so that NMF can for example be used for automatic music transcription (Benetos et al., 2019).

In this section, we report the results obtained for the same algorithms as in Section 5.3 on the audio data sets ”Mary had a little lamb” and the audio sample from (Févotte et al., 2009), see Table 5.

Data set	m	n	r	$\tilde{\lambda}$
Mary had a little lamb	129	586	4	0.3
Prelude from J.S.-Bach	129	2582	16	0.04
Févotte et al. (2009)	513	676	7	0.015

Table 5. Audio data sets.

The parameter $\tilde{\lambda}$ is chosen as the initial ratio between $\lambda_1 \log\det(W^\top W + \delta I)$ and $D_{KL}(X, WH)$. In practice, given the initial point $(W^{(0)}, H^{(0)})$, we set $\tilde{\lambda}$, and the min-vol weight parameter λ_1 is determined using the formula:

$$\lambda_1 = \tilde{\lambda}\frac{D_{KL}(X, W^{(0)}H^{(0)})}{|\log\det(W^{(0),\top}W^{(0)} + \delta I)|}.$$

The value for $\tilde{\lambda}$ for each dataset are extracted from (Leplat et al., 2020), where successful audio source separation with min-vol β-NMF has been shown.

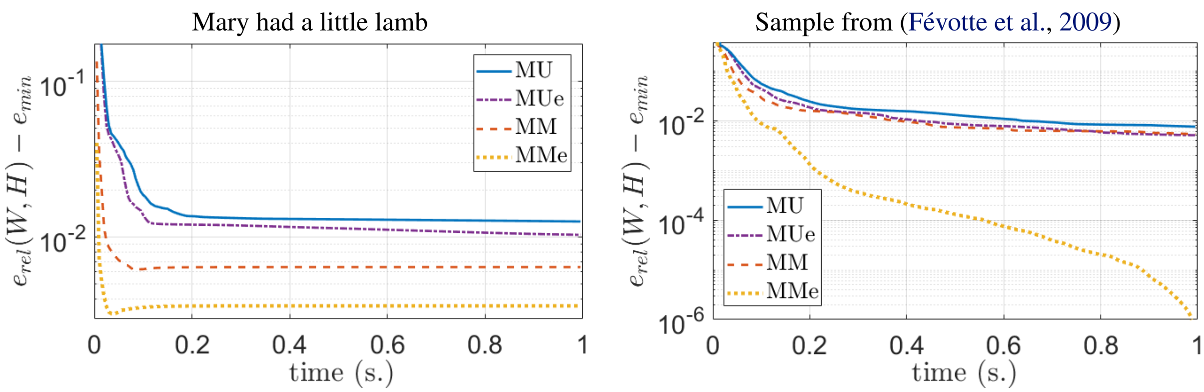

Figure 8. Benchmarked algorithms on first and third data set detailed in Table 5. Evolution of the median relative errors from Equation (27) for 20 random initial initializations minus the smallest relative error found (denoted e_{min}).

Figure 8 shows the median relative errors obtained by each algorithm for these two datasets across 20 runs, each result normalized by subtracting the smallest relative error achieved among all runs. The observations are similar as for the data set presented in the paper, namely MMe outperforms the other algorithms, showing once again that our BMMe framework provides a significant acceleration effect. Moreover, MUe allows us to accelerate MU.

C. Technical proofs

Before proving our convergence result for BMMe in Theorem 3.1, let us give the following useful proposition which is an extension of Property 1 of (Tseng, 2008).

Proposition C.1. *Let $z^+ = \arg\min_{u\in\mathcal{Y}} \varphi(u) + \mathcal{B}_\xi(u,z)$, where φ is a proper convex function, $\mathcal{Y}$ is a closed convex set, and $\mathcal{B}_\xi(u,z) = \xi(u) - \xi(z) - \langle\nabla\xi(z), u - z\rangle$, where ξ is a convex differentiable function in u while fixing z (note that ξ may also depend on z, and we should use $\xi^{(z)}$ for ξ but we omit the upperscript for notation succinctness). Then for all $u \in \mathcal{Y}$ we have*

$$\varphi(u) + \mathcal{B}_\xi(u,z) \geq \varphi(z^+) + \mathcal{B}_\xi(z^+,z) + \mathcal{B}_\xi(u,z^+).$$

Proof. Optimality condition gives us

$$\langle\varphi'(z^+) + \nabla_1\mathcal{B}_\xi(z^+,z), u - z^+\rangle \geq 0, \forall u \in \mathcal{Y},$$

where $\varphi'(z^+)$ is a subgradient of φ at z^+. Furthermore, as φ is convex, we have

$$\varphi(u) \geq \varphi(z^+) + \langle\varphi'(z^+), u - z^+\rangle.$$

Hence, for all $u \in \mathcal{Y}$,

$$\begin{aligned}
&\varphi(u) + \mathcal{B}_\xi(u,z)\\
&\geq \varphi(z^+) - \langle\nabla_u B_\xi(z^+,z), u - z^+\rangle + \mathcal{B}_\xi(u,z)\\
&= \varphi(z^+) - \langle\nabla\xi(z^+) - \nabla\xi(z), u - z^+\rangle + \xi(u) - \xi(z) - \langle\nabla\xi(z), u - z\rangle\\
&= \varphi(z^+) + \mathcal{B}_\xi(z^+,z) + \mathcal{B}_\xi(u,z^+).
\end{aligned}$$

$\square$

C.1. Proof of Theorem 3.1

Since $G_i^t(\cdot,\cdot)$ is a majorizer of $x_i \mapsto f_i^t(x_i)$,

$$f_i^t(x_i^{k+1}) \le G_i^t(x_i^{k+1}, \hat{x}_i^t) = f_i^t(\hat{x}_i^t) + \langle \nabla f_i^t(\hat{x}_i^t), x_i^{t+1} - \hat{x}_i^t \rangle + \mathcal{D}_{\hat{x}_i^t}^t(x_i^{t+1}, \hat{x}_i^t). \tag{29}$$

Applying Proposition C.1 for (10) with $\varphi(x_i) = f_i^t(\hat{x}_i^t) + \langle \nabla f_i^t(\hat{x}_i^t), x_i - \hat{x}_i^t \rangle$ and fixing $z = \hat{x}_i^t$,

$$\varphi(x_i) + \mathcal{D}_{\hat{x}_i^t}^t(x_i, \hat{x}_i^t) \ge \varphi(x_i^{t+1}) + \mathcal{D}_{\hat{x}_i^t}^t(x_i^{t+1}, \hat{x}_i^t) + \mathcal{D}_{\hat{x}_i^t}^t(x_i, x_i^{t+1}). \tag{30}$$

Hence, from (29) and (30), for all $x_i \in \mathcal{X}_i$, we have

$$f_i^t(x_i^{t+1}) \le \varphi(x_i) + \mathcal{D}_{\hat{x}_i^t}^t(x_i, \hat{x}_i^t) - \mathcal{D}_{\hat{x}_i^t}^t(x_i, x_i^{t+1}). \tag{31}$$

On the other hand, since $f_i^t(\cdot)$ is convex, we have $\varphi(x_i) \le f_i^t(x_i)$. Therefore, we obtain the following inequality for all $x_i \in \mathcal{X}_i$

$$f_i^t(x_i^{t+1}) + \mathcal{D}_{\hat{x}_i^t}^t(x_i, x_i^{t+1}) \le f_i^t(x_i) + \mathcal{D}_{\hat{x}_i^t}^t(x_i, \hat{x}_i^t). \tag{32}$$

Taking $x_i = x_i^t$ in (32), using the assumption in (11), and summing up the inequalities from $i = 1$ to s, we obtain

$$f(x^{t+1}) + \sum_{i=1}^s \mathcal{D}_{\hat{x}_i^t}^t(x_i^t, x_i^{t+1}) \le f(x^t) + \sum_{i=1}^s C_i(\alpha_i^t)^2 \|\mathcal{P}_i(x_i^t - x_i^{t-1})\|^2$$

which further implies the following inequality for all $T \ge 1$

$$f(x^{T+1}) + \sum_{t=1}^T \sum_{i=1}^s \mathcal{D}_{\hat{x}_i^t}^t(x_i^t, x_i^{t+1}) \le f(x^1) + \sum_{t=1}^T \sum_{i=1}^s C_i(\alpha_i^t)^2 \|\mathcal{P}_i(x_i^t - x_i^{t-1})\|^2. \tag{33}$$

From (33) and the condition in (12) we have (13) and $\{f(x^t)\}_{t\ge 0}$ is bounded. Hence $\{x^t\}_{t\ge 0}$ is also bounded if f is assumed to have bounded level sets.

Suppose x^* is a limit point of $\{x^k\}$, that is, there exist a subsequence $\{x^{t_k}\}$ converging to x^*. From (13), we have $\mathcal{D}_{\hat{x}_i^t}^t(x_i^t, x_i^{t+1}) \to 0$. Hence, $\{x^{t_k+1}\}$ also converges to x^*. On the other hand, as $\alpha_i^t \|\mathcal{P}_i(x_i^t - x_i^{t-1})\| \to 0$, we have $\hat{x}_i^t \to x_i^*$. From the update in (7),

$$G_i^{t_k}(x_i^{t_k+1}, \hat{x}_i^{t_k}) \le G_i^{t_k}(x_i, \hat{x}_i^{t_k}), \quad \forall x_i \in \mathcal{X}_i$$

Taking $t_k \to \infty$, from Condition (C1), we have

$$\bar{G}_i(x_i^*, x_i^*) \le \bar{G}_i(x_i, x_i^*), \quad \forall x_i \in \mathcal{X}_i,$$

where $\bar{G}_i(\cdot,\cdot)$ is a majorizer of $x_i \mapsto f_i^*(x_i) = (x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_s^*)$. Hence,

$$0 \in \partial(\bar{G}_i(x_i^*, x_i^*) + I_{\mathcal{X}_i}(x_i^*)), \quad \text{for } i = 1, \ldots, s.$$

Finally, using Proposition (Attouch et al., 2010, Proposition 2.1) and noting that $\nabla_1 \bar{G}_i(x_i^*, x_i^*) = \nabla_i f(x^*)$, this implies that x^* is a critical point of (1).

C.2. Proof of Theorem 4.1

Let us first prove that the generated sequence of Algorithm 2 is bounded. For simplicity, we denote $W = W^{t+1}$ and $\hat{H} = \hat{H}^t$ in the following. We have

$$\begin{aligned}
H_{kj}^{t+1} &= \hat{H}_{kj} \frac{\sum_{i=1}^m W_{ik} \frac{X_{ij}}{([W\hat{H}]_{ij})^{2-\beta}}}{\sum_{i=1}^m W_{ik}([W\hat{H}]_{ij})^{\beta-1}} = \sum_{i=1}^m \frac{\hat{H}_{kj} W_{ik} \frac{X_{ij}}{([W\hat{H}]_{ij})^{2-\beta}}}{\sum_{i=1}^m W_{ik}([W\hat{H}]_{ij})^{\beta-1}} \\
&\le \sum_{i=1}^m \frac{\hat{H}_{kj} W_{ik} \frac{X_{ij}}{([W\hat{H}]_{ij})^{2-\beta}}}{W_{ik}([W\hat{H}]_{ij})^{\beta-1}} = \sum_{i=1}^m \hat{H}_{kj} \frac{X_{ij}}{[W\hat{H}]_{ij}} = \sum_{i=1}^m \hat{H}_{kj} \frac{X_{ij}}{\sum_{l=1}^r W_{il}\hat{H}_{lj}} \\
&\le \sum_{i=1}^m \hat{H}_{kj} \frac{X_{ij}}{W_{ik}\hat{H}_{kj}} \le \sum_{i=1}^m \frac{X_{ij}}{\varepsilon}.
\end{aligned}$$

Hence, $\{H^t\}_{t\ge 0}$ is bounded. Similarly we can prove that $\{W^t\}_{t\ge 0}$ is bounded.

Now we verify the conditions of Theorem 3.1. Note that $W \mapsto G_1^{(H)}(W, \tilde{W})$ and $H \mapsto G_2^{(W)}(H, \tilde{H})$ are convex.

Condition (C1) of Theorem 3.1 We see that $(W, H, \tilde{H}) \mapsto G_2^{(W)}(H, \tilde{H})$ is continuously differentiable over $\{(W, H, \tilde{H}) : W \geq \varepsilon, H \geq \varepsilon, \tilde{H} \geq \varepsilon\}$. Furthermore, suppose $(W^{t+1}, H^t) \to (\bar{W}, \bar{H})$ then we have $\bar{W} \geq \varepsilon$ and $\bar{H} \geq \varepsilon$ as $W^{t+1} \geq \varepsilon$ and $H^t \geq \varepsilon$. Hence, it is not difficult to verify that G_2^t satisfies Condition (C1) of Theorem 3.1, and similarly for G_1^t.

Condition (C2) of Theorem 3.1 Considering Condition (C2) of Theorem 3.1, if we fix $\bar{x}$ and $\tilde{x}_i \in \mathcal{Y}_i$, where $\mathcal{Y}_i$ is a closed convex set, and $G_i^{(\bar{x})}(\cdot, \tilde{x}_i)$ is twice continuously differentiable over $\mathcal{Y}_i$ and the norm of its Hessian is upper bounded by $2C_i$ over $\mathcal{Y}_i$, then by the descent lemma (Nesterov, 2018), we have

$$G_i^{(\bar{x})}(x_i, \tilde{x}_i) \leq \bar{f}_i(\tilde{x}_i) + \langle \nabla \bar{f}_i(\tilde{x}_i), x_i - \tilde{x}_i \rangle + C_i \|x_i - \tilde{x}_i\|^2, \forall x_i \in \mathcal{Y}_i.$$

This implies that

$$\mathcal{D}_{\bar{x},\tilde{x}_i}(x_i, \tilde{x}_i) = G_i^{(\bar{x})}(x_i, \tilde{x}_i) - \big(\bar{f}_i(\tilde{x}_i) + \langle \nabla \bar{f}_i(\tilde{x}_i), x_i - \tilde{x}_i \rangle\big) \leq C_i \|x_i - \tilde{x}_i\|^2, \forall x_i \in \mathcal{Y}_i,$$

and hence that the Condition (C2) is satisfied.

Now consider Algorithm 2. Note that $W^t \geq \varepsilon$, $H^t \geq \varepsilon$, $\hat{H}^t = H^t + \alpha_H^t [H^t - H^{t-1}]_+ \geq \varepsilon$, $\hat{W}^t = W^t + \alpha_W^t [W^t - W^{t-1}]_+ \geq \varepsilon$, and we have proved that $\{(W^t, H^t)\}_{t \geq 0}$ generated by Algorithm 2 is bounded. This implies that $\{\hat{W}^t\}_{t \geq 0}$ and $\{\hat{H}^t\}_{t \geq 0}$ are also bounded.We verify (C2) for block H and it is similar for block W; recall that $G_2^{(W)}$ is defined in (17). Consider the compact set $\mathcal{C} = \{(H, \tilde{H}) : H \geq \varepsilon, \tilde{H} \geq \varepsilon, \|(H, \tilde{H})\| \leq C_H\}$, where C_H is a positive constant such that $\mathcal{C}$ contains $(H^t, \hat{H}^t)$. Since $G_2^{(W)}$, with $W \geq \varepsilon$, is twice continuously differentiable over the compact set $\mathcal{C}$, the Hessian $\nabla_H^2 G_2^t(\cdot, \hat{H}^t)$ is bounded by a constant that is independent of W^t and $\hat{H}^t$. As discussed above, this implies that the Condition (C2) of Theorem 3.1 is satisfied.

Condition (C4) of Theorem 3.1 Finally, from (16), we see that $\nabla_{h_k}^2 g_j^{(W)}(h, \hat{h})$ is lower bounded by a positive constant when $h \geq \varepsilon$, $\hat{h} \geq \varepsilon$, $W \geq \varepsilon$, and h, $\hat{h}$, and W are upper bounded. Hence the Condition (C4) of Theorem 3.1 is satisfied.

By Theorem 3.1, any limit point (W^*, H^*) of the generated sequence is a coordinate-wise minimizer of Problem (2). Hence,

$$\begin{aligned} W^* \geq \varepsilon, \quad \langle \nabla_W D_\beta(X, W^* H^*), W - W^* \rangle \geq 0 \quad \forall W \geq \varepsilon, \\ H^* \geq \varepsilon, \quad \langle \nabla_H D_\beta(X, W^* H^*), H - H^* \rangle \geq 0 \quad \forall H \geq \varepsilon. \end{aligned} \tag{34}$$

By choosing $H = H^* + \mathbf{E}_{(i,j)}$ in (34) for each (i, j), where $E_{(i,j)}$ is a matrix with a single component equal to 1 at position (i, j) and the other being 0, we get $\nabla_H D_\beta(X, W^* H^*) \geq 0$. Similarly, we have $\nabla_W D_\beta(X, W^* H^*) \geq 0$. By choosing $H = \varepsilon e e^\top$ and $H = 2H^* - \varepsilon e e^\top$ in (34), we have $\frac{\partial D_\beta(X, W^* H^*)}{\partial H_{ij}}(\varepsilon - H_{ij}^*) = 0$. Similarly, we also have $\frac{\partial D_\beta(X, W^* H^*)}{\partial W_{ij}}(\varepsilon - W_{ij}^*) = 0$. These coincide with the KKT conditions, and hence conclude the proof.

C.3. Proof of Theorem 4.3

We verify the conditions of Theorem 3.1. It is similar to the case of standard β-NMF with $\beta \in [1, 2]$, we have $W \mapsto G_1^{(H)}(W, \tilde{W})$ and $H \mapsto G_2^{(W)}(H, \tilde{H})$ are convex and Condition (C1) are satisfied.

Condition (C2) of Theorem 3.1 At iteration t, we verify (C2) for block H (recall that $G_2^{(W)}$ is defined in (22)), and it is similar for block W, by symmetry. For notation succinctness, in the following we denote $W = W^{t+1}$ and $\hat{H} = \hat{H}^t$. Note that $W^{t+1} \geq \varepsilon$ and $\hat{H}^t = H^t + \alpha_H^t [H^t - H^{t-1}]_+ \geq \varepsilon$. We observe that $G_2^{(W)}$ is separable with respect to the columns $H_{:j}$, $j = 1, \ldots, n$, of H. Specifically,

$$G_2^{(W)}(H, \tilde{H}) = \sum_{j=1}^{n} \Big(g_j^{(W)}(H_{:j}, \tilde{H}_{:j}) + \lambda_2 \bar{\phi}_2^j(H_{:j}, \tilde{H}) \Big) + \lambda_1 \phi_1(W),$$

where $\bar{\phi}_2^j(H_{:j}, \tilde{H}) = \phi_2(\tilde{H}) + [\nabla \phi_2(\tilde{H})]_{:j}^\top (H_{:j} - \tilde{H}_{:j}) + \frac{L_{\phi_2}(\tilde{H})}{2} \|H_{:j} - \tilde{H}_{:j}\|^2$. Hence, as discussed above in the proof of Theorem 4.1, it is sufficient to prove that the norm of the Hessian $\nabla_h^2 \big(g_j^{(W)}(h, \hat{H}_{:j}) + \lambda_2 \bar{\phi}_2^j(h, \hat{H})\big)$, for $j = 1, \ldots, n$, is upper bounded over $h \geq \varepsilon$ by a constant that is independent of W^{t+1} and $\hat{H}^t$. As $L_{\phi_2}(\tilde{H})$ is assumed to be upper bounded by $\bar{L}_{\phi_2}$, it is sufficient to prove that $\nabla_h^2 g_j^{(W)}(h, \hat{H}_{:j})$ is upper bounded over $h \geq \varepsilon$.

We have

$$\nabla_{h_k}^2 g_j^{(W)}(h, \hat{h}) = \sum_{i=1}^m \frac{W_{ik}\hat{v}_i}{\hat{h}_k} \frac{v_i(\hat{h}_k)^2}{(\hat{v}_i h_k)^2} = \sum_{i=1}^m \frac{W_{ik}\hat{h}_k}{\hat{v}_i} \frac{v_i}{(h_k)^2} \overset{(a)}{\leq} \sum_{i=1}^m \frac{W_{ik}\hat{h}_k}{\hat{v}_i} \frac{v_i}{\varepsilon^2} \overset{(b)}{\leq} \sum_{i=1}^m \frac{v_i}{\varepsilon^2},$$

where we used $h_k \geq \varepsilon$ in (a) and $W_{ik}\hat{h}_k \leq \hat{v}_i = \sum_{k=1}^r W_{ik}\hat{h}_k$ in (b). Hence Condition (C2) is satisfied. Together with (19), this implies that the generated sequence of Algorithm 3 is bounded as the objective of (20) has bounded level sets.

Finally, as the generated sequence is upper bounded and $W \geq \varepsilon$, $H \geq \varepsilon$, and $\hat{H} \geq \varepsilon$, we see that $\nabla_{h_k}^2 g_j^{(W)}(h, \hat{h})$ is lower bounded by a positive constant, which implies that the Condition (C4) is satisfied.

C.4. Proof of Lemma 4.2

We have

$$\begin{aligned}
\phi_1(W) &\leq \phi_1(\tilde{W}) + \left\langle (\tilde{W}^\top \tilde{W} + \delta I)^{-1}, W^\top W - \tilde{W}^\top \tilde{W} \right\rangle \\
&\leq \phi_1(\tilde{W}) + \left\langle 2\tilde{W}(\tilde{W}^\top \tilde{W} + \delta I)^{-1}, W - \tilde{W} \right\rangle + \|(\tilde{W}^\top \tilde{W} + \delta I)^{-1}\|_2 \|W - \tilde{W}\|_2^2 \\
&\leq \phi_1(\tilde{W}) + \left\langle \nabla\phi_1(\tilde{W}), W - \tilde{W} \right\rangle + \|(\tilde{W}^\top \tilde{W} + \delta I)^{-1}\|_2 \|W - \tilde{W}\|^2,
\end{aligned}$$

where we use the concavity of $\log\det(\cdot)$ for the first inequality, and the property that $W \mapsto \left\langle (\tilde{W}^\top \tilde{W} + \delta I)^{-1}, W^\top W \right\rangle$ is $2\|(\tilde{W}^\top \tilde{W} + \delta I)^{-1}\|_2$-smooth for the second inequality.

C.5. Proof of Lemma 4.4

The update of W is given by

$$\begin{aligned}
W^{t+1} \leftarrow \arg\min_{W \geq \varepsilon} G_1(W, \hat{W}) &= \sum_{i=1}^m g^i(W_{i:}^\top, \hat{W}_{i:}^\top) + \lambda_1 \bar{\phi}_1(W, \hat{W}) \\
\text{s.t} \quad e^\top W_{:k} &= 1, k = 1, \ldots, r.
\end{aligned} \tag{35}$$

Problem (35) is equivalent to

$$\min_{W \geq \varepsilon} \max_{\mu \in \mathbb{R}^r} \mathcal{L}(W, \mu) := G_1(W, \hat{W}) + \langle W^\top e - e, \mu \rangle.$$

Since $\mathcal{L}(\cdot, \mu)$ is convex, $\mathcal{L}(W, \mu) \to +\infty$ when $\|W\| \to +\infty$, and $\mathcal{L}(W, \cdot)$ is linear, we have strong duality (Bertsekas, 2016, Proposition 4.4.2), that is,

$$\min_{W \geq \varepsilon} \max_{\mu \in \mathbb{R}^r} \mathcal{L}(W, \mu) = \max_{\mu \in \mathbb{R}^r} \min_{W \geq \varepsilon} \mathcal{L}(W, \mu).$$

On the other hand, as $W \mapsto \mathcal{L}(W, \mu)$ is separable with respect to each W_{jk} of W, minimizing this function over $W \geq \varepsilon$ reduces to minimizing scalar strongly convex functions of W_{jk} over $W_{jk} \geq \varepsilon$, for $j = 1, \ldots, n$, $k = 1, \ldots, r$:

$$\min_{W_{jk} \geq \varepsilon} \Big\{ \sum_{i=1}^n \frac{(H^\top)_{ik}\hat{W}_{jk}}{\tilde{v}_i} X_{ji} \log\left(\frac{1}{W_{jk}}\right) + \sum_{i=1}^n (H^\top)_{ik} W_{jk} + \lambda_1 \big(A_{jk}W_{jk} + \frac{1}{2} L_{\phi_1}(\hat{W})(W_{jk} - \hat{W}_{jk})^2\big) + W_{jk}\mu_k \Big\},$$

where $\tilde{v} = H^\top \hat{W}_{j:}^\top$. This optimization problem can be rewritten as

$$\min_{W_{jk} \geq \varepsilon} -b_1 \log(W_{jk}) + b_2 W_{jk} + \frac{1}{2}\lambda_1 L_{\phi_1}(\hat{W}) W_{jk}^2,$$

which has the optimal solution

$$W_{jk}(\mu_k) = \max\left(\varepsilon, \frac{1}{2}\left(-b_2 + \left(b_2^2 + b_3 b_1\right)^{1/2}\right)\right),$$

where

$$b_1 = \sum_{i=1}^{n} \frac{(H^\top)_{ik} X_{ji}}{\tilde{v}_i} \hat{W}_{jk}, \quad b_3 = 4\lambda_1 L_{\phi_1}(\hat{W}),$$
$$b_2(\mu_k) = \sum_{i=1}^{n} (H^\top)_{ik} + \lambda_1 \big(A_{jk} - L_{\phi_1}(\hat{W})\hat{W}_{jk} + \mu_k\big).$$

In matrix form, we have (25). We need to find μ_k such that $\sum_{j=1}^{n} W_{jk}(\mu_k) = 1$. We have $W_{jk}(\mu_k) = \max(\varepsilon, \psi_{jk}(\mu_k))$, where $\psi_{jk}(\mu_k) = \frac{1}{2}(-b_2 + \sqrt{b_2^2 + 4b_3 b_1})$. Note that $\mu_k \mapsto W_{jk}(\mu_k)$ is a decreasing function since

$$\psi'_{jk}(\mu_k) = \frac{1}{2}\left(-\lambda_1 + \frac{\lambda_1 b_2}{\sqrt{b_2^2 + 4b_3 b_1}}\right) < 0.$$

We then apply the bisection method to find the solution of $\sum_{j=1}^{n} W_{jk}(\mu_k) = 1$. To determine the segment containing μ_k, we note that if $\varepsilon < 1/n$ then $W_{jk}(\tilde{\mu}_{jk}) = 1/n$, where $\tilde{\mu}_{jk}$ is defined in (26). Hence, $\mu_k \in [\underline{\mu}_k, \overline{\mu}_k]$, where $\underline{\mu}_k$ and $\overline{\mu}_k$ are defined in (26).