Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

SIAM J. MATH. DATA SCI.
Vol. 7, No. 3, pp. 1292–1314
© 2025 Society for Industrial and Applied Mathematics

Block Majorization Minimization with Extrapolation and Application to β-NMF*

Le Thi Khanh Hien†, Valentin Leplat‡, and Nicolas Gillis§

Abstract. We propose a Block Majorization Minimization method with Extrapolation (BMMe) for solving a class of multiconvex 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 β-divergences (β-NMF) for $\beta \in [1,2]$. These algorithms, which are multiplicative updates with extrapolation, benefit from our novel results, which offer convergence guarantees. We also empirically illustrate the significant acceleration of BMMe for β-NMF through extensive experiments.

Key words. block majorization minimization, extrapolation, nonnegative matrix factorization, β-divergences, Kullback–Leibler divergence

MSC codes. 15A23, 90C26, 94A12

DOI. 10.1137/24M1660188

1. Introduction.

In this paper, we consider the following class of multiconvex optimization problems:

$$\min_{x_i \in \mathcal{X}_i} f(x_1, \ldots, x_s), \tag{1.1}$$

where $x = (x_1, \ldots, x_s)$ is decomposed into s blocks, $\mathcal{X}_i \subseteq \mathbb{E}_i$ is a closed convex set for $i = 1, \ldots, s$, $\mathbb{E}_i$ 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), $\mathcal{X} = \mathcal{X}_1 \times \ldots \mathcal{X}_s \subseteq \operatorname{int}\operatorname{dom}(f)$, and $f : \mathbb{E} = \mathbb{E}_1 \times \cdots \times \mathbb{E}_s \to \mathbb{R} \cup \{+\infty\}$ is a differentiable function over the interior of its domain. Throughout the paper, we assume f is lower bounded and multiconvex, 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 datasets with nonnegative values [20]. 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 $X \approx WH$.

*Received by the editors May 9, 2024; accepted for publication (in revised form) April 14, 2025; published electronically August 13, 2025.
https://doi.org/10.1137/24M1660188
Funding: The third and first authors acknowledge the support by the European Union (ERC consolidator, eLinoR, 101085607).
†Acapela Group, Rue du Crossage 2a, 7012 Mons, Belgium (khanhhiennt@gmail.com).
‡Institute of Data Sciences, Faculty of Engineering, Innopolis Univerity, Innopolis, Russia (V.Leplat@innopolis.ru).
§Department of Mathematics and Operational Research, University of Mons, 7000 Mons, Belgium (nicolas.gillis@umons.ac.be).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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 [10]. This problem is referred to as β-NMF and can be formulated in the form of (1.1) with two blocks of variables, W and H, as follows: given $X \in \mathbb{R}_+^{m\times n}$ and r, solve

$$\min_{\substack{W\in\mathbb{R}^{m\times r}, W\geq\varepsilon,\\ H\in\mathbb{R}^{r\times n}, H\geq\varepsilon}} D_\beta(X, WH), \tag{1.2}$$

where $D_\beta(X,WH) = \sum_{i=1}^m \sum_{j=1}^n d_\beta(X_{ij}, (WH)_{ij})$, with

$$d_\beta(x,y) = \begin{cases} x\log\dfrac{x}{y} - x + y & \text{for } \beta = 1,\\ \dfrac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right) & \text{for } 1 < \beta \leq 2. \end{cases}$$

When $\beta = 2$, d_β is the Euclidean distance, and when $\beta = 1$, it is 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 [14].

1.2. Previous works. Block coordinate descent (BCD) methods serve as conventional techniques for addressing the multiblock problem (1.1). 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 [12, 39], proximal BCD [12], and proximal gradient BCD [4, 6, 41]. These methods fall under the broader framework known as the block successive upper-bound minimization algorithm (BSUM), as introduced in [36]. 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 Definition 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 [43, 33, 35, 32, 15]. Recently, [16] proposed TITAN, an inertial block majorization-minimization framework for solving a more general class of multiblock composite optimization problems than (1.1), in which f is not required to be multiconvex. 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 [16, section 4]. TITAN has proven highly effective in addressing low-rank factorization problems using the Frobenius norm, as demonstrated in [13, 16, 15, 42]. 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 [16, 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 4.1 for the details) lacks strong convexity, and the corresponding error function is not lower bounded by a quadratic function. In other words, although TITAN

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

does not require f to be multiconvex, utilizing TITAN for accelerating the MU in the context of β-NMF is very challenging. This scenario corresponds to a specific instance of problem (1.1).

Consider β-NMF (1.2). For $\beta = 2$, NMF admits very efficient BCD algorithms with theoretically grounded extrapolation [15] and heuristic-based extrapolation mechanisms [1]. Otherwise, the most widely used algorithm to tackle β-NMF is the multiplicative updates (MU): for $1 \le \beta \le 2$,

$$H \leftarrow \text{MU}(X,W,H) = \max\left(\varepsilon, H \circ \frac{[W^\top \frac{[X]}{[WH]^{.(2-\beta)}}]}{[W^\top [WH]^{.(\beta-1)}]}\right), \tag{1.3}$$

and $W^\top \leftarrow \text{MU}(X^\top, H^\top, W^\top)$, where $\circ$ and $\frac{[\cdot]}{[\cdot]}$ are the componentwise product and division between two matrices, respectively, and $(.)^{.x}$ denotes the componentwise exponent. The MU are guaranteed to decrease the objective function [10]; 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 MUs 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 in [14].

1.3. Contribution and outline of the paper.

Drawing inspiration from the versatility of the BSUM framework [36] and the acceleration effect observed in TITAN [16], we introduce BMMe, which stands for Block Majorization Minimization with Extrapolation, to address problem (1.1). Leveraging the multiconvex structure in problem (1.1), BMMe does not need the NSDP condition to ensure convergence; instead, block majorization minimization for the multiconvex problem (1.1) 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):

$$\hat{H} = H + \alpha_H [H - H^p]_+, \quad H \leftarrow \text{MU}(X,W,\hat{H}), \tag{1.4}$$

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 (see Remark 1.1 below) and a slight modification of the original MU, but also has convergence guarantees (Theorem 3.2). We will discuss in detail 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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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 versus MUe on the widely used CBCL facial image dataset with $r = 49$, as in the seminal paper of [20], which 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.

Remark 1.1 (time versus iterations). The extra cost of MUe compared to MU is only the computation of the extrapolated point, $\hat{H} = H + \alpha_H [H - H^p]_+$. For the update of H, this costs $O(nr)$ operations and $O(nr)$ memory. The MU itself requires the computation of WH in $O(mnr)$ operations and $O(mn)$ memory, and multiplying $[X]./[[WH]^{.(2-\beta)}]$ and $[WH]^{.(\beta-1)}$ by $W^\top$ requires $O(mnr)$ operations. The same observation holds for W where the role of m and n are exchanged. For example, for the CBCL dataset experiment in Figure 1, with $m = 361$, $n = 2429$, $r = 49$, MUe requires less than 1% more time than MU: on 30 runs with 1000 iterations, the average time for MU is 11.63 s and for MUe it is 11.71 s, which is about 0.7% more than MU. Given this negligible difference, for simplicity we report the iteration number instead of the computational time when comparing MU with MUe. When comparing with other algorithms, we will use the computational time.

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.

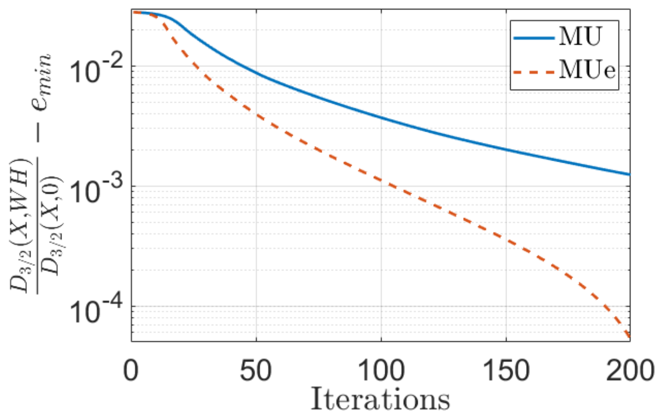

Figure 1. *MU versus MUe with Nesterov extrapolation sequence (3.10) on the CBCL dataset with $r = 49$. Evolution of the median relative objective function values, over 10 random initial initializations, of β-NMF for $\beta = 3/2$ minus the smallest relative objective function found among all runs (denoted e_{min}).*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Notation. We denote $[s]=\{1,\dots,s\}$. 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 jth column and the ith row of X, respectively. We denote the nonnegative part of X as $[X]_+=\max(0,X)$ where the max is taken componentwise. Given a multiblock differentiable function $f:x=(x_1,\dots,x_s)\in\mathbb{E}\mapsto f(x)$, 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) $g(x,y)\ge f(x)$ *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 to note that condition (c) can be replaced by the condition on directional derivatives as in [36, 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 ones will play a pivotal role in this paper. More examples of majorizers can be found in [25, 38, 16].

1. *Lipschitz gradient majorizer* (see, e.g., [43]). 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., [28]). Suppose there exist a differentiable convex function κ and $L>0$ such that $x\mapsto L\kappa(x)-f(x)$ is convex. Then

$$g(x,y)=f(y)+\langle\nabla f(y),x-y\rangle+L(\kappa(x)-\kappa(y)-\langle\nabla\kappa(y),x-y\rangle) \tag{2.1}$$

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

3. *Jensen majorizer* (see, e.g., [8, 30, 19]). 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 $\alpha_i\ge0$, $\sum_{i=1}^r\alpha_i=1$, and $\alpha_i\ne0$ whenever $\omega_i\ne0$, 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,

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Choosing $\alpha_i = \frac{\omega_i y_i}{\omega^\top y}$, $g(x,y) = \sum_{i=1}^r \frac{\omega_i y_i}{\omega^\top y} \tilde{f}(\frac{\omega^\top y}{y_i} x_i)$ 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 tth iterate. This is the majorization minimization (MM) method, which guarantees, by properties of the majorizer, that $f(x^{t+1}) \le f(x^t)$ for all t; see [18, 38] for tutorials.

Example with the multiplicative updates for β-NMF. The standard MU for β-NMF, given in (1.3), 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 therefore provide a majorizer for $D_\beta(v, Wh)$ and show how its closed-form solution leads to the MU (1.3).

The following proposition, which is a corollary of [10, 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 (see [10]). *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 $\beta \in [1,2]$:*

$$g(h, \tilde{h}) = \sum_{i=1}^m \sum_{k=1}^r \frac{W_{ik} \tilde{h}_k}{\tilde{v}_i} d_\beta \left(v_i, \tilde{v}_i \frac{h_k}{\tilde{h}_k} \right). \tag{2.2}$$

With this surrogate, the MU obtained via the MM method [10, eq. 4.1] are as follows:

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

which leads to the MU in the matrix form (1.3). The term *multiplicative* in the name of the algorithm is due to the fact that the new iterate is obtained by an elementwise multiplication between the current iterate, $\tilde{h}$, and a correction factor.

2.2. Critical point and coordinatewise minimizer. Let us define three key notions for our purpose: subdifferential, critical point, and coordinatewise 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*

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

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

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

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

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* $0 \in \partial g(x^*)$.

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

Definition 2.5 (coordinatewise minimizer). *We call* $x^* \in \operatorname{dom} f$ *a coordinatewise minimizer of problem* (1.1) *if*

$$f(x_1^*, \dots, x_{i-1}^*, x_i^*, x_{i+1}^*, \dots, x_s^*) \le f(x_1^*, \dots, x_{i-1}^*, x_i, x_{i+1}^*, \dots, x_s^*) \quad \forall x_i \in \mathcal{X}_i,$$

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

For problem (1.1), a critical point of $f(x) + \sum_{i=1}^s I_{\mathcal{X}_i}(x_i)$ must be a coordinatewise minimizer.

3. Block Majorization Minimization with Extrapolation (BMMe).

In this section, we introduce BMMe (see Algorithm 3.1) and then prove its convergence (see Theorem 3.2).

Remark 3.1 (mappings $\mathcal{P}_i$). BMMe requires the mappings $\mathcal{P}_i$, one for each block of variables. There are multiple choices possible for a given case. For example, if $\operatorname{dom}(f_i^t)$ is the full space, then both $[.]_+$ and the identity mapping would satisfy condition $\hat{x}_i^t \in \operatorname{int} \operatorname{dom}(f_i^t)$. If $\operatorname{dom}(f_i^t)$ is a convex cone, then $\mathcal{P}_i\left(x_i^t - x_i^{t-1}\right) = [x_i^t - x_i^{t-1}]_{\mathcal{C}}$ would satisfy the condition,

---

Algorithm 3.1. BMMe for solving problem (1.1).

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

$$\hat{x}_i^t = x_i^t + \alpha_i^t \mathcal{P}_i\left(x_i^t - x_i^{t-1}\right), \quad \text{where} \tag{3.1}$$

- $\mathcal{P}_i$ is a mapping such that $\hat{x}_i^t \in \operatorname{int} \operatorname{dom}(f_i^t)$, where f_i^t is defined in (3.3) (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:

$$x_i^{t+1} = \operatorname{argmin}_{x_i \in \mathcal{X}_i} G_i^t(x_i, \hat{x}_i^t), \tag{3.2}$$

    where G_i^t is a majorizer of f_i^t over its domain.
6:   **end for**
7: **end for**

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where $[\cdot]_{\mathcal{C}}$ denotes the projection onto any closed convex subset $\mathcal{C}$ of $\mathrm{dom}(f_i^t)$. This is what we use for β-NMF, with $\mathcal{P}_i(a) = [a]_+$ for all i.

3.1. Description of BMMe. BMMe (see Algorithm 3.1) updates one block of variables at a time, say, x_i, by minimizing a majorizer G_i^t of

$$f_i^t(x_i) := f(x_1^{t+1}, \dots, x_{i-1}^{t+1}, x_i, x_{i+1}^t, \dots, x_s^t), \tag{3.3}$$

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 (3.1), while it is evaluated at the previous iterate x_i^t in BMM. The MUe (1.4) described in section 1.1 follows 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 what follows and then state the convergence of BMMe.

- Denote $\bar{f}_i(x_i) := f(\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_s)$, 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.
- 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 superscript of $\xi_i^{(\bar{x},\tilde{x}_i)}$ for notation succinctness. Using this notation, we can write

$$G_i^{(\bar{x})}(x_i, \tilde{x}_i) = \bar{f}_i(\tilde{x}_i) + \langle \nabla \bar{f}_i(\tilde{x}_i), x_i - \tilde{x}_i \rangle + \mathcal{D}_{\bar{x},\tilde{x}_i}(x_i, \tilde{x}_i), \tag{3.4}$$

  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 $x^{(t,i)} = (x_1^{t+1}, \dots, x_{i-1}^{t+1}, x_i^t, x_{i+1}^t, \dots, x_s^t)$, and let $G_i^t = G_i^{(x^{(t,i)})}$ be the majorizer of f_i^t, which is the notation used in Algorithm 3.1, and $\mathcal{D}_{\hat{x}_i}^t = \mathcal{D}_{x^{(t,i)},\hat{x}_i}$.

Key observation for BMMe. Using the notation in (3.4), the MM update in (3.2) can be rewritten as

$$x_i^{t+1} \in \operatorname*{argmin}_{x_i \in \mathcal{X}_i} f_i^t(\hat{x}_i) + \langle \nabla f_i^t(\hat{x}_i), x_i - \hat{x}_i \rangle + \mathcal{D}_{\hat{x}_i}^t(x_i, \hat{x}_i), \tag{3.5}$$

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 3.2. All the technical proofs, except for our main Theorem 3.2, are relegated to Appendix A.

Theorem 3.2. *Consider BMMe described in Algorithm 3.1 for solving problem (1.1). 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 $\bar{f}_i(\cdot)$.*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

(C2) Implicit Lipschitz gradient majorizer. *At iteration t of Algorithm 3.1, for $i \in [s]$, there exists a constant $C_i > 0$ such that*

$$\mathcal{D}^t_{\hat{x}^t_i}(x^t_i, \hat{x}^t_i) \le C_i \|x^t_i - \hat{x}^t_i\|^2 = C_i (\alpha^t_i)^2 \|\mathcal{P}_i(x^t_i - x^{t-1}_i)\|^2. \tag{3.6}$$

(C3) *The sequence of extrapolation parameters satisfies*

$$\sum_{t=1}^{\infty} (\alpha^t_i)^2 \|\mathcal{P}_i(x^t_i - x^{t-1}_i)\|^2 < +\infty \quad \textit{for } i \in [s]. \tag{3.7}$$

Then we have

$$\sum_{t=1}^{\infty} \sum_{i=1}^{s} \mathcal{D}^t_{\hat{x}^t_i}(x^t_i, x^{t+1}_i) < +\infty, \tag{3.8}$$

and the sequence generated by BMMe (Algorithm 3.1), $\{x^t\}_{t\ge 0}$, is bounded if f has bounded level sets.

Furthermore, under the 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 $\{x^t\}_{t\ge 0}$ is a critical point of $f(x) + \sum_{i=1}^s I_{\mathcal{X}_i}(x_i)$.

Before proving Theorem 3.2, let us discuss its conditions:

- If $\mathcal{D}_{\bar{x},\tilde{x}_i}(x_i, \tilde{x}_i) \le C_i \|x_i - \tilde{x}_i\|^2$, then (3.6) 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 4.4.
- 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-strong convexity implies

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

- We see that the update in (3.5) has the form of an accelerated mirror descent (AMD) method [40] for the one-block convex problem. Hence, it makes sense to involve the extrapolation sequences that are used in AMD. This strategy has been used in [43, 15, 16]. An example of choosing the extrapolation parameters satisfying (3.7) is

$$\alpha^t_i = \min\left\{ \alpha^t_{Nes}, \frac{c}{t^{q/2}} \frac{1}{\|\mathcal{P}_i(x^t_i - x^{t-1}_i)\|} \right\}, \tag{3.9}$$

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

$$\eta_t = \frac{1}{2}\left(1 + \sqrt{1 + 4\eta^2_{t-1}}\right), \; \alpha^t_{Nes} = \frac{\eta_{t-1} - 1}{\eta_t}. \tag{3.10}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Another choice is replacing α^t_{Nes} in (3.9) by $\alpha^t_c = \frac{t-1}{t}$ [40]. In our experiments, we observe that when c is large enough, α^t_i coincides with α^t_{Nes} (or α^t_c). A motivation to choose α^t_{Nes} is that, for the one-block problem with an L-smooth convex objective, BMMe with the Lipschitz gradient majorizer recovers the famous Nesterov fast gradient method, which is an optimal first-order method. However, it is pertinent to note that the best choice may vary depending on the particular application and dataset characteristics.

Proof of Theorem 3.2. Since $G^t_i(\cdot,\cdot)$ is a majorizer of $x_i \mapsto f^t_i(x_i)$,

$$f^t_i(x^{k+1}_i) \le G^t_i(x^{k+1}_i, \hat{x}^t_i) = f^t_i(\hat{x}^t_i) + \langle \nabla f^t_i(\hat{x}^t_i), x^{t+1}_i - \hat{x}^t_i\rangle + \mathcal{D}^t_{\hat{x}^t_i}(x^{t+1}_i, \hat{x}^t_i). \tag{3.11}$$

Applying Proposition A.1 in Appendix A.1 for (3.5) with $\varphi(x_i) = f^t_i(\hat{x}^t_i) + \langle \nabla f^t_i(\hat{x}^t_i), x_i - \hat{x}^t_i\rangle$ and fixing $z = \hat{x}^t_i$, we get

$$\varphi(x_i) + \mathcal{D}^t_{\hat{x}^t_i}(x_i, \hat{x}^t_i) \ge \varphi(x^{t+1}_i) + \mathcal{D}^t_{\hat{x}^t_i}(x^{t+1}_i, \hat{x}^t_i) + \mathcal{D}^t_{\hat{x}^t_i}(x_i, x^{t+1}_i). \tag{3.12}$$

Hence, from (3.11) and (3.12), for all $x_i \in \mathcal{X}_i$, we have

$$f^t_i(x^{t+1}_i) \le \varphi(x_i) + \mathcal{D}^t_{\hat{x}^t_i}(x_i, \hat{x}^t_i) - \mathcal{D}^t_{\hat{x}^t_i}(x_i, x^{t+1}_i). \tag{3.13}$$

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

$$f^t_i(x^{t+1}_i) + \mathcal{D}^t_{\hat{x}^t_i}(x_i, x^{t+1}_i) \le f^t_i(x_i) + \mathcal{D}^t_{\hat{x}^t_i}(x_i, \hat{x}^t_i). \tag{3.14}$$

Taking $x_i = x^t_i$ in (3.14), using the assumption in (3.6), and summing up the inequalities from $i = 1$ to s, we obtain

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

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

$$f(x^{T+1}) + \sum_{t=1}^{T}\sum_{i=1}^{s} \mathcal{D}^t_{\hat{x}^t_i}(x^t_i, x^{t+1}_i) \le f(x^1) + \sum_{t=1}^{T}\sum_{i=1}^{s} C_i(\alpha^t_i)^2 \|\mathcal{P}_i(x^t_i - x^{t-1}_i)\|^2. \tag{3.15}$$

From (3.15) and the condition in (3.7) we have (3.8) 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 exists a subsequence $\{x^{t_k}\}$ converging to x^*. From (3.8), we have $\mathcal{D}^t_{\hat{x}^t_i}(x^t_i, x^{t+1}_i) \to 0$. Hence, $\{x^{t_k+1}\}$ also converges to x^*. On the other hand, as $\alpha^t_i \|\mathcal{P}_i(x^t_i - x^{t-1}_i)\| \to 0$, we have $\hat{x}^t_i \to x^*_i$. From the update in (3.2),

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

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 [3, 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 (1.1). ■

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Complexity, scalability, and practical implementation aspects. BMMe serves as an accelerated version of the block majorization-minimization method (BMM). In essence, while BMM updates each block x_i by $x_i^{t+1} = \operatorname{argmin}_{x_i \in \mathcal{X}_i} G_i^t(x_i, x_i^t)$, BMMe achieves the same by replacing x_i^t by an extrapolation point $\hat{x}_i^t$, which is computed with a marginal additional cost, namely $O(n)$ operations, where n is the number of variables. Consequently, BMMe inherits crucial properties regarding complexity, scalability, and practical implementation from BMM. BMM updates one block of variables at a time while keeping others fixed, thus scaling effectively with data size in terms of the number of block variables. However, the complexity of each block update grows with the dimension of the block variable. The efficiency of BMM heavily relies on selecting suitable majorizers, ensuring closed-form solutions for block updates, and thus circumventing the need for outer solvers in large-scale problems (this may help avoid substantial computational resources). The choice of appropriate majorizers is pivotal and application-specific; for instance, in applications utilizing the β-NMF model, Jensen majorizers are commonly and effectively employed; see section 4 below for the details. Moreover, it is worth noting that a properly designed majorizer allows for the computation of its closed-form minimizer in an elementwise manner. This characteristic is particularly beneficial for handling large-scale problems, as it can be efficiently executed on a parallel computation platform.

Iteration complexity. Iteration complexity of BMM-type methods is a challenging topic, especially for nonconvex problems. Considering the use of general majorizers together with inertial/extrapolated parameters in each block update, the work that is most related to our paper is [16]. As explained in [16, Remark 9], as long as a global convergence (that is, the whole generated sequence converges to a critical point) is guaranteed, a convergence rate for the generated sequence can be derived by using the same technique as in the proof of [2, Theorem 2]. In fact, the technique of [2] has been commonly used to establish the convergence rate in other block coordinate methods, for which specific majorizers are used in each block update (for example, [43] uses Lipschitz gradient majorizers). However, it is challenging to extend the result to BMMe. Along with the Kurdyka–Łojasiewicz assumption, the BMM-type methods with extrapolation, such as [16, 43], need to establish the NSDP, as discussed in section 1.2. And as such, extending the usual convergence rate result to BMMe is an open question. A recent work [24] establishes iteration complexity of a BMM method for solving constrained nonconvex nonsmooth problems; however, the majorizers are required to have a Lipschitz gradient, which is not satisfied by many Bregman majorizers.

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 (2.2) 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

$$\nabla^2_{h_k} g(h,\tilde{h}) = \sum_{i=1}^{m} \frac{W_{ik}\tilde{v}_i}{\tilde{h}_k} d''_\beta\left(v_i, \tilde{v}_i \frac{h_k}{\tilde{h}_k}\right), \tag{4.1}$$

where $d''_\beta(x,y)$ denotes the second derivative with respect to y of $(x,y) \mapsto d_\beta(x,y)$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

On the other hand, as already noted in section 2.1,

$$D_\beta(X,WH)=\sum_{j=1}^{n} D_\beta(X_{:j},WH_{:j})=\sum_{i=1}^{m} D_\beta(X_{i:}^\top,H^\top W_{i:}^\top).$$

Hence a majorizer of $H\mapsto D_\beta(X,WH)$ while fixing W is given by

$$G_2^{(W)}(H,\tilde{H})=\sum_{j=1}^{n} g_j^{(W)}(H_{:j},\tilde{H}_{:j}), \tag{4.2}$$

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 (2.2) with $v=X_{:j}$. Similarly, the following function is a majorizer of $W\mapsto D_\beta(X,WH)$ while fixing H:

$$G_1^{(H)}(W,\tilde{W})=\sum_{i=1}^{m} \mathbf{g}_i^{(H)}(W_{i:}^\top,\tilde{W}_{i:}^\top), \tag{4.3}$$

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 (2.2) with v being replaced by $X_{i:}^\top$ and W being replaced by $H^\top$.

Note that there exist other majorizers for β-NMF—for example, majorizers for both variables simultaneously [26], for ℓ_1-regularized β-NMF with sum-to-one constraints [27], and quadratic majorizers for the KL divergence [34].

Remark 4.1 (choice of majorizer). Incorporating a regularization term $\lambda|x_i-x_i^t|^2$ in the Jensen surrogate to have a strongly convexity majorizer would fulfill the conditions outlined in TITAN [16]. However, this is not recommended, as it would result in a regularized Jensen surrogate that lacks a closed-form solution for the subproblem, necessitating an outer solver and requiring adaptation of the convergence analysis of TITAN to accommodate inexact solutions (the current analysis of TITAN does not support inexact solutions). In contrast, the extrapolation strategy employed by BMMe will preserve the closed-form update for β-NMF in its iterative step by embedding the extrapolation point directly into the majorizer.

4.2. MU with extrapolation for β-NMF, $\beta\in[1,2]$. We consider the standard β-NMF problem in (1.2) with $\beta\in[1,2]$. Applying Algorithm 3.1 to solve problem (1.2), we get MUe, a multiplicative update method with extrapolation described in Algorithm 4.1. The convergence property of MUe is given in Theorem 4.2; see Appendix A.2 for the proof.

Theorem 4.2. *Suppose the extrapolation parameters in Algorithm 4.1 are chosen such that they are bounded and*

$$\sum_{t=1}^{\infty}(\alpha_H^t)^2\|[H^t-H^{t-1}]_+\|^2<+\infty, \quad \text{and} \quad \sum_{t=1}^{\infty}(\alpha_W^t)^2\|[W^t-W^{t-1}]_+\|^2<+\infty. \tag{4.4}$$

Then MUe (Algorithm 4.1) generates a bounded sequence and any one of its limit points, (W^*,H^*), is a KKT point of problem (1.2), that is,

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

and similarly for H^*.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 4.1. MUe for solving β-NMF (1.2).

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 (4.4).
4: Update the two blocks of variables:

$$W^{t+1} = \operatorname*{argmin}_{W \geq \varepsilon} G_1^t(W, \hat{W}^t) = \mathrm{MU}(X^\top, (H^t)^\top, (\hat{W}^t)^\top)^\top \text{ [see (1.3)]},$$
$$H^{t+1} = \operatorname*{argmin}_{H \geq \varepsilon} G_2^t(H, \hat{H}^t) = \mathrm{MU}(X, W^{t+1}, \hat{H}^t),$$

where $G_1^t = G_1^{(H^t)}$ and $G_2^t = G_2^{(W^{t+1})}$ be the majorizers defined in (4.3) with $H = H^t$ and (4.2) with $W = W^{t+1}$, respectively.
5: **end for**

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 datasets such as images [37], documents [20], and single-cell sequencing [7]. 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 [23, 11], a crucial aspect in various applications such as hyperspectral imaging [29] and audio source separation [21]. In this section, we consider the following general regularized KL-NMF problem:

$$\min_{W \in \bar{\Omega}_W, H \in \bar{\Omega}_H} \Big\{ f(W,H) := D_{KL}(X, WH) + \lambda_1 \phi_1(W) + \lambda_2 \phi_2(H) \Big\}, \tag{4.5}$$

where $\bar{\Omega}_W := \{W : W \geq \varepsilon, W \in \Omega_W\}$ and $\bar{\Omega}_H := \{H : H \geq \varepsilon, H \in \Omega_H\}$. 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

$$\begin{aligned} \phi_1(W) &\leq \bar{\phi}_1(W, \tilde{W}) := \phi_1(\tilde{W}) + \langle \nabla \phi_1(\tilde{W}), W - \tilde{W} \rangle + \frac{L_{\phi_1}(\tilde{W})}{2} \|W - \tilde{W}\|^2, \\ \phi_2(H) &\leq \bar{\phi}_2(H, \tilde{H}) := \phi_2(\tilde{H}) + \langle \nabla \phi_2(\tilde{H}), H - \tilde{H} \rangle + \frac{L_{\phi_2}(\tilde{H})}{2} \|H - \tilde{H}\|^2. \end{aligned} \tag{4.6}$$

Furthermore, $L_{\phi_1}(\tilde{W})$ and $L_{\phi_2}(\tilde{H})$ in (4.6) 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 $\Omega_W = \{W \mid W^\top e = e\}$ [21]. In that case, $\phi_1(W)$ satisfies this condition, as proved in Lemma 4.3; see Appendix A.3 for the proof.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Lemma 4.3. *The function $\phi_1(W) = \log\det(W^\top W + \delta I)$ with $\delta > 0$ satisfies the condition in (4.6) with $L_{\phi_1}(\tilde{W}) = 2\|(\tilde{W}^\top \tilde{W} + \delta I)^{-1}\|_2$, which is upper bounded by $2/\delta$.*

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

$$G_2^{(W)}(H,\tilde{H}) = \sum_{j=1}^{n} g_j^{(W)}(H_{:j},\tilde{H}_{:j}) + \lambda_1\phi_1(W) + \lambda_2\bar{\phi}_2(H,\tilde{H}), \tag{4.7}$$

where $\bar{\phi}_2(H,\tilde{H})$ is defined as in (4.6) and $g_j^{(W)}$ is defined as in (4.2). Similarly, we use the following majorizer for $W \mapsto f(W,H)$ while fixing H:

$$G_1^{(H)}(W,\tilde{W}) = \sum_{i=1}^{m} \mathbf{g}_i^{(H)}(W_{i:}^\top,\tilde{W}_{i:}^\top) + \lambda_1\bar{\phi}_1(W,\tilde{W}) + \lambda_2\phi_2(H), \tag{4.8}$$

where $\bar{\phi}_1(W,\tilde{W})$ is defined as in (4.6), and $\mathbf{g}_i^{(H)}$ is defined as in (4.3). Applying Algorithm 3.1 to solve problem (4.5), we get Algorithm B.1, a BMMe algorithm for regularized and constrained KL-NMF; see its detailed description in Appendix B. It works exactly as Algorithm 4.1, but the majorizers $G_1^t = G_1^{(H^t)}$ and $G_2^t = G_2^{(W^{t+1})}$ are defined as in (4.8) with $H = H^t$ and as in (4.7) with $W = W^{t+1}$, respectively. The convergence property of Algorithm B.1 is given in Theorem 4.4; see Appendix A.4 for the proof.

Theorem 4.4. *Suppose the extrapolation parameters α_W^t and α_H^t satisfy (4.4). Then BMMe applied to problem (4.5) (Algorithm B.1) generates a bounded sequence, and any one of its limit points is a coordinatewise minimizer of problem (4.5).*

BMMe for solving problem (4.5) (Algorithm B.1) 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 [21]:

$$\begin{aligned} \min_{W\geq\varepsilon, H\geq\varepsilon} \; & D_{KL}(X,WH) + \lambda_1 \log\det(W^\top W + \delta I) \\ \text{such that} \quad & e^\top W_{:j} = 1, j = 1,\ldots,r. \end{aligned} \tag{4.9}$$

The update of H is, as in (1.3), taking $\beta = 1$. The following lemma provides the update of W; see Appendix A.5 for the proof.

Lemma 4.5. *For notation succinctness, let $\hat{W} = W^t + \alpha_W^t[W^t - W^{t-1}]_+$ and $H = H^{t+1}$. BMMe for solving (4.9) (Algorithm B.1) updates W as follows:*

$$W \leftarrow \max\left(\varepsilon, \frac{1}{2}\left(-B_2 + \left[[B_2]^{.2} + 4\lambda_1 L_{\phi_1}(\hat{W})B_1\right]^{.1/2}\right)\right), \tag{4.10}$$

where

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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

$$
\begin{aligned}
&\underline{\mu}_k = \min_{j=1,\dots,m} \tilde{\mu}_{jk}, \quad \overline{\mu}_k = \max_{j=1,\dots,m} \tilde{\mu}_{jk}, \\
&\tilde{\mu}_{jk} = \frac{1}{\lambda_1}(4\lambda_1 L_{\phi_1}(\hat{W}) b_1 m - 1/m - \sum_{i=1}^n (H^\top)_{ik}) + L_{\phi_1}(\hat{W})\hat{W}_{jk} - A_{jk},
\end{aligned} \tag{4.11}
$$

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 (3.10). All experiments have been performed on a laptop computer with Intel Core i7-11800H @ 2.30 GHz and 16 GB memory with MATLAB R2021b. The code is available from https://github.com/vleplat/BMMe.

5.1. β-NMF for hyperspectral imaging.

We consider β-NMF (1.2) with $\beta = 3/2$, which is among the best NMF models for hyperspectral unmixing [9]. For this problem, the MU is the workhorse approach, and we compare it to MUe: Figure 2 provides the median evolution of objective function values for the Cuprite dataset ($m = 188$, $n = 47750$, $r = 20$); see the supplementary material section SM1 for more details and experiments on 3 other datasets with similar observations. There is a significant acceleration effect: on average, MUe requires only 41 iterations to obtain a smaller objective than MU with 100 iterations.

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 [17]. As reported in [14], MU and CCD are the state of the art for KL-NMF (sometimes one performs best, and sometimes the other does). Figure 3 shows the median evolution of the relative objective function for two datasets: a dense image dataset (ORL, $m = 10304$, $n = 400$), and

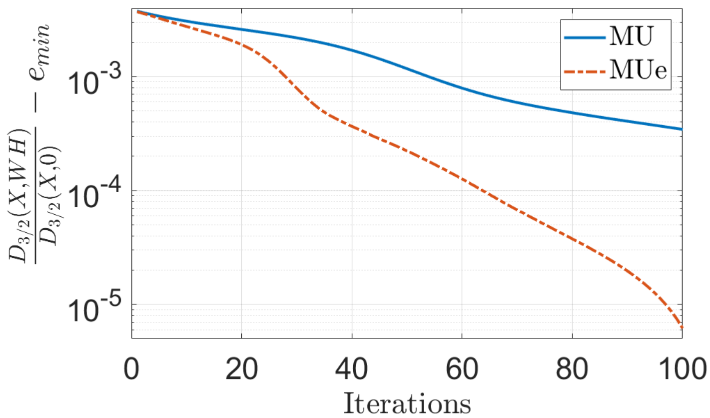

Figure 2. *Median relative objective function of* β*-NMF for* $\beta = 3/2$ *minus the smallest objective function found among* 10 *random initializations (denoted* e_{min}*).*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

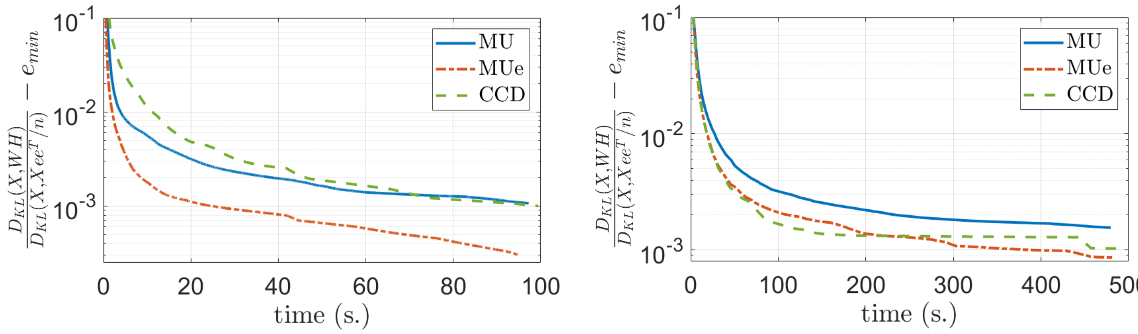

Figure 3. *Median relative objective function of KL-NMF minus the smallest objective function found among 10 runs (denoted e_{min}) w.r.t. CPU time: (left) ORL, (right) hitech, with $r = 10$ in both cases. Note that MU, MUe, and CCD, respectively, perform on average 7579, 6949, and 616 iterations for ORL, and 885, 886 and 343 iterations for hitech in the considered time intervals.*

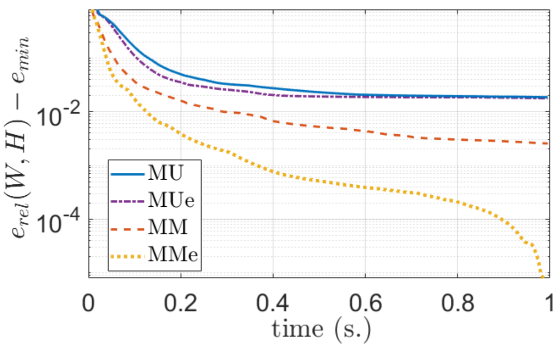

Figure 4. *Evolution of the median relative errors (see (5.1)), minus the smallest relative error found among the 20 random initializations, of min-vol KL-NMF (4.9) of the four algorithms. Note that MU, MUe, MM, and MMe, respectively, perform on average 94, 92, 304, and 278 iterations within one second.*

a sparse document dataset (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 datasets; see the supplementary material section SM2.

5.3. Min-vol KL-NMF for audio datasets. We address problem (4.9) in the context of blind audio source separation [21]. We compare the following algorithms: MU with H update from (1.3) and W update from Lemma 4.5, its extrapolated variant, MUe, a recent MM algorithm by [22], denoted MM, and MMe incorporating the BMMe extrapolation step in MM.

Figure 4 displays the median relative objective function values,

$$e_{rel}(W,H) = \frac{D_{KL}(X,WH) + \lambda \log\det(WW + \delta I)}{D_{KL}(X,(Xe/n)e^\top)}, \tag{5.1}$$

minus the smallest relative objective found, for the prelude from J.S. Bach ($m = 129$, $n = 2292$, $r = 16$); see the supplementary material section SM3 for two other datasets and more details. MUe exhibits accelerated convergence compared to MU. MM ranks second, while

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

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 satisfies neither condition (c) of Definition 2.1 nor Assumption 1 (A3) of [36], and hence Theorem 3.2 does not apply to MMe.

6. Conclusion and further work.

In this paper, we considered multiconvex optimization (1.1). 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 includes the use of BMMe for other NMF models and algorithms [26, 27] and 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 [36, Assumption 1(A3)], extending to the case when the subproblems in each block of variables are not convex,[1] or studying iteration complexity of BMMe (see the paragraph at the end of section 3).

Appendix A. Technical proofs.

A.1. Proof of Proposition A.1.

To prove our convergence result for BMMe in Theorem 3.2, we need the following useful proposition, which is an extension of Property 1 of [40].

Proposition A.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 superscript 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. The optimality condition gives us

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

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 $u \in \mathcal{Y}$,

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

∎

[1] We need the convexity assumption for the proof of Theorem 3.2 and Proposition A.1. Without convexity, we cannot establish the inequality (3.15), which is key to prove Theorem 3.2.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

A.2. Proof of Theorem 4.2. Let us first prove that the generated sequence of Algorithm 4.1 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 3.2. 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* 3.2. We see that $(W,H,\tilde{H}) \mapsto G_2^{(W)}(H,\tilde{H})$ is continuously differentiable over $\{(W,H,\tilde{H}) : W \ge \varepsilon, H \ge \varepsilon, \tilde{H} \ge \varepsilon\}$. Furthermore, suppose $(W^{t+1},H^t) \to (\bar{W},\bar{H})$; then we have $\bar{W} \ge \varepsilon$ and $\bar{H} \ge \varepsilon$ as $W^{t+1} \ge \varepsilon$ and $H^t \ge \varepsilon$. Hence, it is not difficult to verify that G_2^t satisfies condition (C1) of Theorem 3.2, and similarly for G_1^t.

Condition (C2) *of Theorem* 3.2. Considering condition (C2) of Theorem 3.2, 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 [31], we have

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

This implies that

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

and hence that condition (C2) is satisfied.

Now consider Algorithm 4.1. Note that $W^t \ge \varepsilon$, $H^t \ge \varepsilon$, $\hat{H}^t = H^t + \alpha_H^t[H^t - H^{t-1}]_+ \ge \varepsilon$, $\hat{W}^t = W^t + \alpha_W^t[W^t - W^{t-1}]_+ \ge \varepsilon$, and we have proved that $\{(W^t,H^t)\}_{t\ge 0}$ generated by Algorithm 4.1 is bounded. This implies that $\{\hat{W}^t\}_{t\ge 0}$ and $\{\hat{H}^t\}_{t\ge 0}$ are also bounded. We verify (C2) for block H, and it is similar for block W; recall that $G_2^{(W)}$ is defined in (4.2). Consider the compact set $\mathcal{C} = \{(H,\tilde{H}) : H \ge \varepsilon, \tilde{H} \ge \varepsilon, \|(H,\tilde{H})\| \le C_H\}$, where C_H is a positive constant such that $\mathcal{C}$ contains $(H^t,\hat{H}^t)$. Since $G_2^{(W)}$, with $W \ge \varepsilon$, 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 condition (C2) of Theorem 3.2 is satisfied.

Condition (C4) *of Theorem* 3.2. Finally, from (4.1), we see that $\nabla_{h_k}^2 g_j^{(W)}(h,\hat{h})$ is lower bounded by a positive constant when $h \ge \varepsilon$, $\hat{h} \ge \varepsilon$, $W \ge \varepsilon$, and h, $\hat{h}$, and W is upper bounded. Hence condition (C4) of Theorem 3.2 is satisfied.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

By Theorem 3.2, any limit point (W^*, H^*) of the generated sequence is a coordinatewise minimizer of problem (1.2). Hence,

$$\text{(A.1)} \qquad \begin{aligned} W^* &\geq \varepsilon, \quad \langle \nabla_W D_\beta(X, W^*H^*), W - W^* \rangle \geq 0 \quad \forall W \geq \varepsilon, \\ H^* &\geq \varepsilon, \quad \langle \nabla_H D_\beta(X, W^*H^*), H - H^* \rangle \geq 0 \quad \forall H \geq \varepsilon. \end{aligned}$$

By choosing $H = H^* + \mathbf{E}_{(i,j)}$ in (A.1) for each (i,j), where $\mathbf{E}_{(i,j)}$ 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 $H = \varepsilon ee^\top$ and $H = 2H^* - \varepsilon ee^\top$ in (A.1), 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 we conclude the proof.

A.3. Proof of Lemma 4.3. We have

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

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

A.4. Proof of Theorem 4.4. We verify the conditions of Theorem 3.2. It is similar to the case of standard β-NMF with $\beta \in [1,2]$; we have that $W \mapsto G_1^{(H)}(W, \tilde{W})$ and $H \mapsto G_2^{(W)}(H, \tilde{H})$ are convex and condition (C1) is satisfied.

Condition (C2) *of Theorem* 3.2. At iteration t, we verify (C2) for block H (recall that $G_2^{(W)}$ is defined in (4.7)), 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,

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

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 4.2, it is sufficient to prove that the norm of the Hessian $\nabla_h^2 (g_j^{(W)}(h, \hat{H}_{:j}) + \lambda_2 \bar{\phi}_2^j(h, \hat{H}))$, 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

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

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 (4.4), this implies that the generated sequence of Algorithm B.1 is bounded as the objective of (4.5) has bounded level sets.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

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

A.5. Proof of Lemma 4.5. The update of W is given by

$$W^{t+1} \leftarrow \arg\min_{W \geq \varepsilon} G_1(W, \hat{W}) = \sum_{i=1}^{m} g^i(W_{i:}^\top, \hat{W}_{i:}^\top) + \lambda_1 \bar{\phi}_1(W, \hat{W}) \quad \text{s.t.} \quad e^\top W_{:k} = 1, k = 1, \dots, r. \tag{A.2}$$

Problem (A.2) 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 [5, 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 $j = 1, \dots, n$, $k = 1, \dots, r$:

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

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

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

where

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

$$b_2(\mu_k) = \sum_{i=1}^{n} (H^\top)_{ik} + \lambda_1 (A_{jk} - L_{\phi_1}(\hat{W}) \hat{W}_{jk} + \mu_k).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

In matrix form, we have (4.10). 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 $\psi_{jk}(\mu_k) = \frac{1}{2}(-b_2 + \sqrt{b_2^2 + 4b_3b_1})$. Note that $\mu_k \mapsto W_{jk}(\mu_k)$ is a decreasing function since

$$\psi'_{jk}(\mu_k) = \frac{1}{2}\left(-\lambda_1 + \frac{\lambda_1 b_2}{\sqrt{b_2^2 + 4b_3b_1}}\right) < 0.$$

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 as in (4.11). Hence, $\mu_k \in [\underline{\mu}_k, \overline{\mu}_k]$, where $\underline{\mu}_k$ and $\overline{\mu}_k$ are defined as in (4.11).

Appendix B. BMMe for solving constrained and regularized KL-NMF (4.5). Algorithm B.1 is BMMe for the specific case of constrained and regularized KL-NMF.

Algorithm B.1. BMMe for solving constrained and regularized KL-NMF (4.5).

1: Choose initial points $W^{-1} \geq \varepsilon, W^0 \geq \varepsilon, H^{-1} \geq \varepsilon, H^0 \geq \varepsilon$.
2: **for** $t = 1, \dots$ **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 the condition of Theorem 4.2.
4: Update the two blocks of variables:

$$W^{t+1} \in \underset{W \in \bar{\Omega}_W}{\operatorname{argmin}} G_1^t(W, \hat{W}^t), \qquad \text{(B.1)}$$
$$H^{t+1} \in \underset{H \in \bar{\Omega}_H}{\operatorname{argmin}} G_2^t(H, \hat{H}^t),$$

where $G_1^t = G_1^{(H^t)}$ and $G_2^t = G_2^{(W^{t+1})}$ are the majorizers defined in (4.8) with $H = H^t$ and (4.7) with $W = W^{t+1}$, respectively.
5: **end for**

Acknowledgment. We are grateful to the anonymous reviewers who carefully read the manuscript. Their feedback helped us improve our paper.

REFERENCES

[1] A. M. S. Ang and N. Gillis, *Accelerating nonnegative matrix factorization algorithms using extrapolation*, Neural Comput., 31 (2019), pp. 417–439, https://doi.org/10.1162/neco_a_01157.
[2] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Program., 116 (2009), pp. 5–16, https://doi.org/10.1007/s10107-007-0133-5.
[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457, https://doi.org/10.1287/moor.1100.0449.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[4] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060, https://doi.org/10.1137/120887679.
[5] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 2016.
[6] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494, https://doi.org/10.1007/s10107-013-0701-9.
[7] P. Carbonetto, K. Luo, A. Sarkar, A. Hung, K. Tayeb, S. Pott, and M. Stephens, *GoM DE: Interpreting structure in sequence count data with differential expression analysis allowing for grades of membership*, Genome Biol., 24 (2023), 236, https://doi.org/10.1186/s13059-023-03067-9.
[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B Stat. Methodol., 39 (1977), pp. 1–22, https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
[9] C. Févotte and N. Dobigeon, *Nonlinear Hyperspectral Unmixing with Robust Nonnegative Matrix Factorization*, preprint, arXiv:1401.5649, 2014.
[10] C. Févotte and J. Idier, *Algorithms for nonnegative matrix factorization with the β-divergence*, Neural Comput., 23 (2011), pp. 2421–2456, https://doi.org/10.1162/NECO_a_00168.
[11] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, *Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain*, IEEE Trans. Signal Process., 63 (2015), pp. 2306–2320, https://doi.org/10.1109/TSP.2015.2404577.
[12] L. Grippo and M. Sciandrone, *On the convergence of the block nonlinear Gauss–Seidel method under convex constraints*, Oper. Res. Lett., 26 (2000), pp. 127–136, https://doi.org/10.1016/S0167-6377(99)00074-7.
[13] L. Hien, D. Phan, and N. Gillis, *Inertial alternating direction method of multipliers for non-convex non-smooth optimization*, Comput. Optim. Appl., 83 (2022), pp. 247–285, https://doi.org/10.1007/s10589-022-00394-8.
[14] L. T. K. Hien and N. Gillis, *Algorithms for nonnegative matrix factorization with the Kullback-Leibler divergence*, J. Sci. Comput., 87 (2021), 93, https://doi.org/10.1007/s10915-021-01504-0.
[15] L. T. K. Hien, N. Gillis, and P. Patrinos, *Inertial block proximal method for non-convex non-smooth optimization*, in Thirty-Seventh International Conference on Machine Learning (ICML), 2020.
[16] L. T. K. Hien, D. N. Phan, and N. Gillis, *An inertial block majorization minimization framework for nonsmooth nonconvex optimization*, J. Mach. Learn. Res., 24 (2023), pp. 1–41, https://www.jmlr.org/papers/v24/21-0571.html.
[17] C. J. Hsieh and I. S. Dhillon, *Fast coordinate descent methods with variable selection for non-negative matrix factorization*, in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.
[18] D. R. Hunter and K. Lange, *A tutorial on MM algorithms*, Amer. Statist., 58 (2004), pp. 30–37, https://doi.org/10.1198/0003130042836.
[19] K. Lange, D. R. Hunter, and I. Yang, *Optimization transfer using surrogate objective functions*, J. Comput. Graph. Stat., 9 (2000), pp. 1–20, https://doi.org/10.1080/10618600.2000.10474858.
[20] D. D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401 (1999), pp. 788–791, https://doi.org/10.1038/44565.
[21] V. Leplat, N. Gillis, and A. M. Ang, *Blind audio source separation with minimum-volume beta-divergence NMF*, IEEE Trans. Signal Process., 68 (2020), pp. 3400–3410, https://doi.org/10.1109/TSP.2020.2991801.
[22] V. Leplat, N. Gillis, and J. Idier, *Multiplicative updates for NMF with β-divergences under disjoint equality constraints*, SIAM J. Matrix Anal. Appl., 42 (2021), pp. 730–752, https://doi.org/10.1137/20M1377278.
[23] C. H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi, *Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case*, IEEE Trans. Geosci. Remote Sens., 53 (2015), pp. 5530–5546, https://doi.org/10.1109/TGRS.2015.2424719.
[24] H. Lyu and Y. Li, *Block majorization-minimization with diminishing radius for constrained non-smooth nonconvex optimization*, SIAM J. Optim., 35 (2025), pp. 842–871, https://doi.org/10.1137/23M1604515.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

[25] J. Mairal, *Optimization with first-order surrogate functions*, in 30th International Conference on Machine Learning, Vol. 28, 2013, pp. 783–791.
[26] A. Marmin, J. H. de Morais Goulart, and C. Févotte, *Joint majorization-minimization for nonnegative matrix factorization with the β-divergence*, Signal Process., 209 (2023), 109048, https://doi.org/10.1016/j.sigpro.2023.109048.
[27] A. Marmin, J. H. de Morais Goulart, and C. Févotte, *Majorization-minimization for sparse nonnegative matrix factorization with the β-divergence*, IEEE Trans. Signal Process., 71 (2023), pp. 1435–1447, https://doi.org/10.1109/TSP.2023.3266939.
[28] J. G. Melo and R. D. C. Monteiro, *Iteration-Complexity of a Jacobi-Type Non-Euclidean ADMM for Multi-block Linearly Constrained Nonconvex Programs*, https://arxiv.org/abs/1705.07229, 2017.
[29] L. Miao and H. Qi, *Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization*, IEEE Trans. Geosci. Remote Sens., 45 (2007), pp. 765–777, https://doi.org/10.1109/TGRS.2006.888466.
[30] R. M. Neal and G. E. Hinton, *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*, Springer, Dordrecht, The Netherlands, 1998, pp. 355–368.
[31] Y. Nesterov, *Lectures on Convex Optimization*, Springer, 2018.
[32] P. Ochs, *Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano*, SIAM J. Optim., 29 (2019), pp. 541–570, https://doi.org/10.1137/17M1124085.
[33] P. Ochs, Y. Chen, T. Brox, and T. Pock, *iPiano: Inertial proximal algorithm for nonconvex optimization*, SIAM J. Imaging Sci., 7 (2014), pp. 1388–1419, https://doi.org/10.1137/130942954.
[34] M. Q. Pham, J. Cohen, and T. Chonavel, *A Fast Multiplicative Updates Algorithm for Nonnegative Matrix Factorization*, preprint, arXiv:2303.17992, 2023.
[35] T. Pock and S. Sabach, *Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems*, SIAM J. Imaging Sci., 9 (2016), pp. 1756–1787, https://doi.org/10.1137/16M1064064.
[36] M. Razaviyayn, M. Hong, and Z.-Q. Luo, *A unified convergence analysis of block successive minimization methods for nonsmooth optimization*, SIAM J. Optim., 23 (2013), pp. 1126–1153, https://doi.org/10.1137/120891009.
[37] W. H. Richardson, *Bayesian-based iterative method of image restoration*, J. Optical Soc. Amer., 62 (1972), pp. 55–59, https://doi.org/10.1364/JOSA.62.000055.
[38] Y. Sun, P. Babu, and D. Palomar, *Majorization-minimization algorithms in signal processing, communications, and machine learning*, IEEE Trans. Signal Process., 65 (2017), pp. 794–816, https://doi.org/10.1109/TSP.2016.2601299.
[39] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl., 109 (2001), pp. 475–494, https://doi.org/10.1023/A:1017501703105.
[40] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, Technical report, 2008.
[41] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423, https://doi.org/10.1007/s10107-007-0170-0.
[42] O. Vu Thanh, A. Ang, N. Gillis, and L. T. K. Hien, *Inertial majorization-minimization algorithm for minimum-volume NMF*, in European Signal Processing Conference (EUSIPCO), 2021.
[43] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci., 6 (2013), pp. 1758–1789, https://doi.org/10.1137/120887795.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/17/25 to 109.136.75.42 by Nicolas Gillis (nicolas.gillis@umons.ac.be). Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy