Computational Optimization and Applications (2021) 79:681–715
https://doi.org/10.1007/s10589-021-00286-3

Multi-block Bregman proximal alternating linearized minimization and its application to orthogonal nonnegative matrix factorization

Masoud Ahookhosh[1] · **Le Thi Khanh Hien**[2] · **Nicolas Gillis**[2] · **Panagiotis Patrinos**[3]

Received: 16 February 2020 / Accepted: 28 May 2021 / Published online: 9 June 2021

Abstract
We introduce and analyze BPALM and A-BPALM, two multi-block proximal alternating linearized minimization algorithms using Bregman distances for solving structured nonconvex problems. The objective function is the sum of a multi-block relatively smooth function (i.e., relatively smooth by fixing all the blocks except one) and block separable (nonsmooth) nonconvex functions. The sequences generated by our algorithms are subsequentially convergent to critical points of the objective function, while they are globally convergent under the KL inequality assumption. Moreover, the rate of convergence is further analyzed for functions satisfying the Łojasiewicz's gradient inequality. We apply this framework to orthogonal nonnegative matrix factorization (ONMF) that satisfies all of our assumptions and the related subproblems are solved in closed forms, where some preliminary numerical results are reported.

Keywords Nonsmooth nonconvex optimization · Proximal alternating linearized minimization · Bregman distance · Multi-block relative smoothness · KL inequality · Orthogonal nonnegative matrix factorization

Mathematics Subject Classification 90C06 · 90C25 · 90C26 · 49J52 · 49J53

1 Introduction

For a nonempty, convex, and open set $C \subseteq \mathbb{R}^n$, we consider the structured nonsmooth nonconvex minimization problem

✉ Masoud Ahookhosh
masoud.ahookhosh@uantwerp.be

Extended author information available on the last page of the article

$$\underset{\boldsymbol{x}=(x_1,\ldots,x_N)\in\overline{C}}{\textbf{minimize}} \quad \varphi(\boldsymbol{x}) \equiv f(\boldsymbol{x}) + \sum_{i=1}^{N} g_i(x_i), \tag{1.1}$$

where we will assume the following hypotheses (see Sect. 3 for details):

Assumption 1 (**requirements for composite minimization** (1.1))

A1 $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ is proper and lower semicontinuous (lsc);
A2 $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, closed, and *multi-block* Legendre;
A3 $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $\mathcal{C}^1(\textbf{int dom}\, h)$ and $(L_1, \ldots, L_N)$-*smooth relative to* h; here $n = \sum_{i=1}^{N} n_i$;
A4 $\textbf{arg min}\, \varphi \neq \emptyset$, $\overline{C} := \overline{\textbf{dom}\, h}$, $C := \textbf{int dom}\, h$, $g := \sum_{i=1}^{N} g_i$, and $C \subseteq \textbf{dom}\, g$.

There is a huge number of algorithmic studies around solving the optimization problems of the form (1.1). Among all of such methodologies, we are interested in the class of *alternating minimization* algorithms such as block coordinate descent [15, 18, 42, 48, 53, 54, 61, 62], block coordinate [30, 32, 41], and Gauss-Seidel methods [11, 19, 36], which assume that all blocks are fixed except one and solve the corresponding auxiliary problem with respect to this block, update the latter block, and continue with the others. In particular, the proximal alternating minimization has received much attention in the last few years; see for example [6–10, 16]. Recently, the proximal alternating linearized minimization and its variants has been developed to handle (1.1); see for example [23, 51, 57].

Traditionally, the Lipschitz (Hölder) continuity of partial gradients of f in (1.1) is a necessary tool for providing the convergence analysis of optimization algorithms; see, e.g., [23, 51]. It is, however, well-known that it is not the Lipschitz (Hölder) continuity of gradients playing a key role in such analysis, but one of its consequences: an upper estimation of f including a Bregman distance called *descent lemma*; cf. [13, 46]. This idea is central to convergence analysis of many optimization schemes requiring such an upper estimation; see, e.g., [1–3, 12, 13, 24, 37, 38, 46, 59, 60]. In this paper, we propose a multi-block extension of the descent lemma given in [13, 46] and propose a Bregman proximal alternating linearized minimization (BPALM) algorithm and its adaptive version (A-BPALM) for (1.1).

1.1 Contribution

Our contribution is summarized as follows:

(i) (*Relative smoothness with possibly nonconvex kernel*) An extension of the relative smoothness [13, 46] for problems with block separable structure entails an application of multi-block convex kernel functions (see Definition 3.1) which is not necessarily a jointly convex function. This paves the way toward the development of non-Euclidean alternating minimization methodologies.

(ii) (*Bregman proximal alternating linearized minimization*) We introduce BPALM, a multi-block generalization of the proximal alternating linearized minimization (PALM) [23] using Bregman distances, and its adaptive version (A-BPALM). To do so, we extend the notion of relative smoothness [13, 46] to its multi-block counterpart to support a structured problem of the form (1.1). Owing to multi-block relative smoothness of f, our algorithm does not need to know the local Lipschitz moduli of partial gradients $\nabla_i f$ $(i = 1, \ldots, N)$ and their lower and upper bounds, which are hard to provide in practice.

(iii) (*Efficient framework for ONMF*) Exploiting a suitable kernel for Bregman distance, it turns out that the objective of orthogonal nonnegative matrix factorization (ONMF) is multi-block relatively smooth, and the subproblems of our algorithms are solved in closed forms making them suitable for large-scale data analysis problems. To the best of our knowledge, BPALM and A-BPALM are the first algorithms with rigorous convergence theory for ONMF.

1.2 Related works

Closely related to our framework, there are two papers [43, 63]. However, we notice that [63] uses a sum separable kernel function which is a special case of our multi-block kernel functions (see Example 3.2), and the paper provides a subsequential convergence theory. Regarding [43], an algorithm (named B-PALM) was proposed that is a special case of our BPALM when $N = 2$, $g_1 = g_2 = 0$, and $f \in \mathcal{C}^2$. Hence, the proposed algorithms in [43] are not applicable to ONMF, which requires nonnegativity constraints. We stress that involving the block separable nonsmooth nonconvex functions g_i and considering $N > 2$ make our analysis different from those of [43].

1.3 Organization

This paper has five sections, besides this introductory section. In Sect. 2, we describe some preliminaries, and in Sect. 3, we introduce the notion of multi-block relative smoothness, and verify the fundamental properties of Bregman proximal alternating linearized mapping. In Sect. 4, we introduce BPALM and A-BPALM and investigate their convergence analysis. In Sect. 5, we show that ONMF satisfies our assumptions, the related subproblems are solved in closed forms and report our numerical results. Finally, Sect. 6 delivers some conclusions.

2 Notation and preliminaries

We denote by $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ the extended-real line. For the identity matrix I_n, we set $U_i \in \mathbb{R}^{n \times n_i}$ such that $I_n = (U_1, \ldots, U_N) \in \mathbb{R}^{n \times n}$. For notation clarity, we will use bold lower-case letters (e.g., $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$) for vectors in $\mathbb{R}^{\sum n_i}$ and use normal lower-case letters (e.g., z, x_i, y_i) for vectors in $\mathbb{R}^{n_i}$. The open ball of radius $r > 0$ centered in $x \in \mathbb{R}^p$ is denoted as $\mathbf{B}(x;r)$. For a set $C \subseteq \mathbb{R}^p$, $\overline{C}$ denotes its closure. The set

of cluster points of $(x^k)_{k\in\mathbb{N}}$ is denoted as $\omega(x^0)$. A function $f:\mathbb{R}^p\to\overline{\mathbb{R}}$ is *proper* $f\not\equiv\infty$, in which case its *domain* is defined as the set $\mathbf{dom} f:=\{x\in\mathbb{R}^p\}[f(x)<\infty]$. For $\alpha\in\mathbb{R}$, $[f\leq\alpha]:=\{x\in\mathbb{R}^p\}[f(x)\leq\alpha]$ is the *α-(sub)level set* of f; $[f\geq\alpha]$ and $[f=\alpha]$ are defined similarly. We say that f is *level bounded* if $[f\leq\alpha]$ is bounded for all $\alpha\in\mathbb{R}$. A vector $v\in\partial f(x)$ is a *subgradient* of f at x, and the set of all such vectors is called the *subdifferential* $\partial f(x)$ [55, Definition 8.3], i.e.

$$\partial f(x)=\{v\in\mathbb{R}^p\}[\exists(x^k,v^k)_{k\in\mathbb{N}}\text{ s.t. } x^k\to x,\ f(x^k)\to f(x),\ \widehat{\partial} f(x^k)\ni v^k\to v],$$

and $\widehat{\partial} f(x)$ is the set of *regular subgradients* of f at x, namely

$$\widehat{\partial} f(x)=\{v\in\mathbb{R}^p\}[f(z)\geq f(x)+\langle v,z-x\rangle+o(\|z-x\|),\ \forall z\in\mathbb{R}^p].$$

The Fenchel conjugate function $f^*:\mathbb{R}^p\to\overline{\mathbb{R}}$ of $f:\mathbb{R}^p\to\overline{\mathbb{R}}$ is given by

$$f^*(z^*)=\mathbf{sup}\{\langle z^*,x\rangle-f(x)\}[x\in\mathbb{R}^p].$$

For a given nonempty closed convex set $S\subseteq\mathbb{R}^p$, the function $\delta_S:\mathbb{R}^p\to\overline{\mathbb{R}}$ denotes the *indicator function*, namely $\delta_S(x)=0$ if $x\in S$ and $\delta_S(x)=\infty$ otherwise. Moreover, $\mathbf{Proj}_S:\mathbb{R}^p\to\mathbb{R}^p$ denotes the *projection function* given by $\mathbf{Proj}_S(x)=\underset{z\in S}{\mathbf{arg\,min}}\,\|z-x\|$.

2.1 Bregman proximal mapping

For a kernel function $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ and a proper lower semicontinuous function $g:\mathbb{R}^n\to\overline{\mathbb{R}}$, the *Bregman proximal mapping* is given by

$$\mathbf{prox}^h_{\gamma g}(x):=\underset{z\in\mathbb{R}^n}{\mathbf{arg\,min}}\{g(z)+\tfrac{1}{\gamma}\mathbf{D}_h(z,x)\}. \tag{2.1}$$

which is a generalization of the classical one by using the Bregman distance (3.2) in place of the Euclidean distance; see, e.g., [28] and references therein. We note that

$$\mathbf{prox}^h_{\gamma g}(x)=\{y\in\mathbf{dom}\,g\cap\mathbf{dom}\,h\mid g(y)+\tfrac{1}{\gamma}\mathbf{D}_h(y,x)=\mathbf{inf}_z\{g(z)+\tfrac{1}{\gamma}\mathbf{D}_h(z,x)\}<+\infty\},$$

which implies $\mathbf{dom}\,\mathbf{prox}^h_{\gamma g}\subset\mathbf{intdom}\,h$, $\mathbf{rangeprox}^h_{\gamma g}\subset\mathbf{dom}\,g\cap\mathbf{dom}\,h$. The function g is *h-prox-bounded* if there exists $\gamma>0$ such that $\mathbf{inf}_z\{g(z)+\frac{1}{\gamma}\mathbf{D}_h(z,x)\}>-\infty$ for some $x\in\mathbb{R}^n$; see for example [3].

2.2 Kurdyka-Łojasiewicz (KL) function

Let us consider the class of functions that are satisfying the celebrated Kurdyka-Łojasiewicz inequality, which we present next.

Definition 2.1 (KL property) A proper and lsc function $\varphi:\mathbb{R}^{n_1}\times\ldots\times\mathbb{R}^{n_N}\to\overline{\mathbb{R}}$ has the *Kurdyka-Łojasiewicz* property (KL property) at $\boldsymbol{x}^\star\in\mathbf{dom}\,\varphi$ if there exist a

concave *desingularizing function* $\psi : [0,\eta] \to [0,+\infty[$ (for some $\eta > 0$) and neighborhood $\mathbf{B}(\boldsymbol{x}^\star;\varepsilon)$ with $\varepsilon > 0$, such that

(i) $\psi(0) = 0$;
(ii) ψ is of class $\mathcal{C}^1$ with $\psi > 0$ on $(0,\eta)$;
(iii) for all $\boldsymbol{x} \in \mathbf{B}(\boldsymbol{x}^\star;\varepsilon)$ such that $\varphi(\boldsymbol{x}^\star) < \varphi(\boldsymbol{x}) < \varphi(\boldsymbol{x}^\star) + \eta$ it holds that

$$\psi'(\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{x}^\star))\mathbf{dist}(0, \partial\varphi(\boldsymbol{x})) \geq 1. \tag{2.2}$$

If φ satisfies the KL property at each point of $\mathbf{dom}\,\partial\varphi$, then φ is called a KL function.

The first inequality of this type is given in the seminal work of Łojasiewicz [44, 45] for analytic functions, which we nowadays call Łojasiewicz's gradient inequality. Later, Kurdyka [40] showed that this inequality is valid for $\mathcal{C}^1$ functions whose graph belong to an *o-minimal structure* (see its definition in [31]). The first extensions of the KL property to nonsmooth functions were given by Bolte et al. [20–22].

The following two facts constitute the crucial steps for establishing the global convergence of the algorithms given in Sect. 4.

Fact 2.2 (uniformized KL property) [23, Lemma 6] Let Ω be a compact set and $\zeta : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper and lower semicontinuous function. Assume that ζ is constant on Ω and satisfies the KL property at each point of Ω. Then, there exist $\varepsilon > 0$, $\eta > 0$, and a desingularizing function ψ such that for $\overline{u} \in \Omega$ and all u in the intersection

$$\{u \in \mathbb{R}^d \mid \mathbf{dist}(u,\Omega) < \varepsilon\} \cap [\zeta(\overline{u}) < \zeta(u) < \zeta(\overline{u}) + \eta]$$

we have

$$\psi'(\zeta(u) - \zeta(\overline{u}))\mathbf{dist}(0, \partial\zeta(u)) \geq 1.$$

Fact 2.3 [25, Lemma 2.3] Let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be sequences in $[0,+\infty)$ such that $\sum_{k=1}^{\infty} b_k < \infty$ and $a_{k+1} \leq \alpha a_k + b_k$ for all $k \in \mathbb{N}$ and some $\alpha < 1$. Then, $\sum_{k=1}^{\infty} a_k < \infty$.

3 Multi-block Bregman proximal mapping

In this section, we first establish the notion of multi-block relative smoothness, which is an extension of the relative smoothness [13, 46] for problems with block structure. Afterwards, we introduce Bregman alternating linearized mapping and study some of its basic properties.

3.1 Multi-block relative smoothness

In order to extend the definition of Bregman distances for the multi-block problem (1.1), we first need to introduce the notion of *multi-block kernel* functions, which will coincide with the standard one (cf. [3, Definition 2.1]) if $N = 1$.

Definition 3.1 (multi-block convexity and kernel function) Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and lsc function with $\mathbf{int\,dom}\, h \neq \emptyset$ and such that $h \in \mathcal{C}^1(\mathbf{int\,dom}\, h)$. For a fixed vector $\boldsymbol{x} \in \mathbb{R}^n$, we define the function $h^i_{\boldsymbol{x}} : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ given by

$$h^i_{\boldsymbol{x}}(z) := h(\boldsymbol{x} + U_i(z - x_i)). \tag{3.1}$$

Then, we say that h is

(i) *multi-block (strongly/strictly) convex* if the function $h^i_{\boldsymbol{x}}(\cdot)$ is (strongly/strictly) convex for all $\boldsymbol{x} \in \mathbf{dom}\, h$ and $i = 1, \ldots, N$;
(ii) *multi-block locally strongly convex* around $\boldsymbol{x}^\star = (x_1^\star, \ldots, x_N^\star)$ if, for $i = 1, \ldots, N$, there exists $\delta > 0$ and $\sigma^i_h > 0$ such that

$$h^i_{\boldsymbol{x}}(x_i) \geq h^i_{\boldsymbol{x}}(y_i) + \langle \nabla_i h(\boldsymbol{x} + U_i(y_i - x_i)), x_i - y_i\rangle + \frac{\sigma^i_h}{2}\|x_i - y_i\|^2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{B}(\boldsymbol{x}^\star; \delta);$$

(iii) a *multi-block kernel function* if h is multi-block convex and $h^i_{\boldsymbol{x}}(\cdot)$ is 1-coercive for all $\boldsymbol{x} \in \mathbf{dom}\, h$ and $i = 1, \ldots, N$, i.e., $\mathbf{lim}_{\|z\|\to\infty} \frac{h^i_{\boldsymbol{x}}(z)}{\|z\|} = \infty$;
(iv) *essentially smooth*, if for every sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}} \subseteq \mathbf{int\,dom}\, h$ converging to a boundary point of $\mathbf{dom}\, h$, we have $\|\nabla h(\boldsymbol{x}^k)\| \to \infty$;
(v) of *multi-block Legendre* type if it is essentially smooth and multi-block strictly convex.

Example 3.2 (popular kernel functions) There are many kernel functions satisfying the conditions given in Definition 3.1(iii). For example, for $N = 1$, energy, Boltzmann-Shannon entropy, Fermi-Dirac entropy (cf. [14, Example 2.3]) and several examples in [46, Section 2]; and for $N = 2$ see two examples in [43, Section 2]. Two important classes of multi-block kernels are *sum separable kernels*, i.e., $h(x_1, \ldots, x_N) = h_1(x_1) + \ldots + h_N(x_N)$, and *product separable kernels*, i.e., $h(x_1, \ldots, x_N) = h_1(x_1) \times \ldots \times h_N(x_N)$, see such a kernel for ONMF in Proposition 5.1.

We now give the definition of *Bregman distances* (cf. [27]) for multi-block kernels.

Definition 3.3 (Bregman distance) For a multi-block kernel function h, the *Bregman distance* $\mathbf{D}_h : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is given by

$$\mathbf{D}_h(\boldsymbol{y}, \boldsymbol{x}) := \begin{cases} h(\boldsymbol{y}) - h(\boldsymbol{x}) - \langle \nabla h(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle & \boldsymbol{y} \in \mathbf{dom}\, h, x \in \mathbf{int\,dom}\, h \\ \infty & \text{otherwise.} \end{cases} \tag{3.2}$$

Fixing all blocks except the i-th one, the Bregman distance with respect to this block is given by

$$\begin{aligned} \mathbf{D}_h(\boldsymbol{x} + U_i(y_i - x_i), \boldsymbol{x}) &= h(\boldsymbol{x} + U_i(y_i - x_i)) - h(\boldsymbol{x}) - \langle \nabla h(\boldsymbol{x}), U_i(y_i - x_i)\rangle \\ &= h^i_{\boldsymbol{x}}(y_i) - h^i_{\boldsymbol{x}}(x_i) - \langle \nabla_i h(\boldsymbol{x}), y_i - x_i\rangle, \end{aligned}$$

which measures the proximity between $\boldsymbol{y}$ and $\boldsymbol{x}$ with respect to the i-th block of variables. Moreover, the kernel h is multi-block convex if and only if $\mathbf{D}_h(\boldsymbol{x}+U_i(y_i-x_i),\boldsymbol{x})\geq 0$ for all $\boldsymbol{y}\in\mathbf{dom}\,h$ and $\boldsymbol{x}\in\mathbf{int\,dom}\,h$ and $i=1,\ldots,N$. Note that if h is multi-block strictly convex, then $\mathbf{D}_h(\boldsymbol{x}+U_i(y_i-x_i),\boldsymbol{x})=0$ $(i=1,\ldots,N)$ if and only if $x_i=y_i$.

We are now in a position to present the notion of *multi-block relative smoothness*, which is the central tool for our analysis in Sect. 4.

Definition 3.4 (multi-block relative smoothness) Let $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a multi-block kernel and let $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a proper and lower semicontinuous function. If there exists $L_i>0$ $(i=1,\ldots,N)$ such that the functions $\phi^i_{\boldsymbol{x}}:\mathbb{R}^{n_i}\to\overline{\mathbb{R}}$ given by

$$\phi^i_{\boldsymbol{x}}(z):=L_ih(\boldsymbol{x}+U_i(z-x_i))-f(\boldsymbol{x}+U_i(z-x_i))$$

are convex for all $\boldsymbol{x},\boldsymbol{x}+U_i(z-x_i)\in\mathbf{int\,dom}\,h$ and $i=1,\ldots,N$, then f is called $(L_1,\ldots,L_N)$-*smooth relative to* h.

Note that if $N=1$, the multi-block relative smoothness is reduced to standard relative smoothness, which was introduced only recently in [13, 46]. In the Euclidean case, relative smoothness is equivalent to the one-sided descent lemma, which is implied by the Lipschitz continuity of gradients ∇f (i.e., $\frac{L}{2}\|\cdot\|^2-f$ is convex). Therefore, the relative smoothness of f generalizes the notions of Lipschitz smoothness using Bregman distances. If $N=2$, this definition will be reduced to the relative bi-smoothness given in [43] for $h,f\in\mathcal{C}^2$.

We next characterize the notion of multi-block relative smoothness.

Proposition 3.5 (characterization of multi-block relative smoothness) *Let $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a multi-block kernel and let $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a proper lower semicontinuous function and $f\in\mathcal{C}^1$. Then, the following statements are equivalent*:

(a) *f is $(L_1,\ldots,L_N)$-smooth relative to h;*
(b) *for all $(\boldsymbol{x},\boldsymbol{y})\in\mathbf{int\,dom}\,h\times\mathbf{int\,dom}\,h$ and $i=1,\ldots,N$,*

$$f(\boldsymbol{x}+U_i(y_i-x_i))\leq f(\boldsymbol{x})+\langle\nabla_i f(\boldsymbol{x}),y_i-x_i\rangle+L_i\mathbf{D}_h(\boldsymbol{x}+U_i(y_i-x_i),\boldsymbol{x});\quad(3.3)$$

(c) *for all $(\boldsymbol{x},\boldsymbol{y})\in\mathbf{int\,dom}\,h\times\mathbf{int\,dom}\,h$ and $i=1,\ldots,N$,*

$$\langle\nabla_i f(\boldsymbol{x})-\nabla_i f(\boldsymbol{y}),x_i-y_i\rangle\leq L_i\langle\nabla_i h(\boldsymbol{x})-\nabla_i h(\boldsymbol{y}),x_i-y_i\rangle;\quad(3.4)$$

(d) *if $f\in\mathcal{C}^2(\mathbf{int\,dom}\,f)$ and $h\in\mathcal{C}^2(\mathbf{int\,dom}\,h)$, then*

$$L_i\nabla^2_{x_ix_i}h(\boldsymbol{x})-\nabla^2_{x_ix_i}f(\boldsymbol{x})\succeq 0,\quad(3.5)$$

for $i=1,\ldots,N$ and for all $\boldsymbol{x}\in\mathbf{int\,dom}\,h$.

Proof Fixing all the blocks except one of them, the results can be concluded in the same way as [46, Proposition 1.1]. □

3.2 Bregman proximal alternating linearized mapping

The main objective of the remainder of this section is to provide a guarantee for existence of the *Bregman proximal alternating linearized minimization mapping* (see (3.9)) and to investigate its fundamental properties.

Let us begin with an extension of h-prox-boundedness to our multi-block setting, where both of them are extensions of the classical prox-boundedness condition; see, e.g., [55, Definition 1.23].

Definition 3.6 (multi-block h-prox-boundedness) A function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *multi-block h-prox-bounded* if for each $i \in \{1, \dots, N\}$ there exists $\gamma_i > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$ such that

$$g^{h/\gamma_i}(\boldsymbol{x}) := \mathbf{inf}_{z \in \mathbb{R}^{n_i}} \{g(\boldsymbol{x} + U_i(z - x_i)) + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x})\} > -\infty.$$

The supremum of the set of all such γ_i is the threshold $\gamma_{i,g}^h$ of the multi-block h-prox-boundedness, i.e.,

$$\gamma_{i,g}^h := \mathbf{sup}\{\gamma_i > 0 \mid \exists \boldsymbol{x} \in \mathbb{R}^n \text{ s.t. } g^{h/\gamma_i}(\boldsymbol{x}) > -\infty\}. \tag{3.6}$$

For the problem (1.1), we have $g = \sum_{i=1}^N g_i$ leading to

$$g^{h/\gamma_i}(\boldsymbol{x}) = \sum_{j \neq i} g_j(x_j) + \mathbf{inf}_{z \in \mathbb{R}^{n_i}} \{g_i(z) + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x})\}. \tag{3.7}$$

If g is multi-block h-prox-bounded for $\overline{\gamma}_i > 0$, so is for all $\gamma_i \in (0, \overline{\gamma}_i)$. We next present equivalent conditions to this notion.

Proposition 3.7 (characterization of multi-block h-prox-boundedness) *For a multi-block kernel function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and proper and lsc functions $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ with $i = 1, \dots, N$, the following statements are equivalent*:

(a) $g = \sum_{i=1}^N g_i$ *is multi-block h-prox-bounded*;
(b) *for all $i = 1, \dots, N$ and $h_{\boldsymbol{x}}^i$ given in (3.1), $g_i + r_i h_{\boldsymbol{x}}^i$ is bounded below on $\mathbb{R}^{n_i}$ for some $r_i \in \mathbb{R}$*;
(c) *for all $i = 1, \dots, N$, $\mathbf{liminf}_{\|z\| \to \infty} \frac{g_i(z)}{h_{\boldsymbol{x}}^i(z)} > -\infty$.*

Proof We first show $(a) \Leftrightarrow (b)$. Let us assume $g^{h/\gamma_i}(\boldsymbol{x}) > -\infty$, and let $r_i > \frac{1}{\gamma_i}$. Then, for all $i = 1, \dots, N$, it holds that

$$\begin{aligned} g_i(z) + r_i h_{\boldsymbol{x}}^i(z) &= g_i(z) + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x}) + r_i h_{\boldsymbol{x}}^i(z) - \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x}) \\ &\geq g^{h/\gamma_i}(\boldsymbol{x}) - \sum_{j \neq i} g_j(x_j) + \tfrac{r_i\gamma_i - 1}{\gamma_i} h_{\boldsymbol{x}}^i(z) + \tfrac{1}{\gamma_i}(h(\boldsymbol{x}) + \langle \nabla_i h(\boldsymbol{x}), z - x_i \rangle) =: \tilde{g}_i(z), \end{aligned}$$

for the function $\tilde{g}_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$. We notice that $\tilde{g}_i$ is convex and coercive, and as such is lower bounded. Conversely, suppose that $\alpha_i := \mathbf{inf} g_i + r_i h_{\boldsymbol{x}}^i > -\infty$. Then, from (3.7), we obtain

$$\begin{aligned} g^{h/\gamma_i}(\boldsymbol{x}) &= \sum_{j\neq i} g_j(x_j) + \min_{z\in\mathbb{R}^{n_i}} \{g_i(z) + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z-x_i), \boldsymbol{x})\}, \\ &\geq \sum_{j\neq i} g_j(x_j) + \alpha_i + \mathbf{inf}_z\{-r_i h_{\boldsymbol{x}}^i(z) + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z-x_i), \boldsymbol{x})\} \\ &\geq \sum_{j\neq i} g_j(x_j) + \alpha_i - \tfrac{1}{\gamma_i}h(\boldsymbol{x}) + \tfrac{1}{\gamma_i}\langle \nabla_i h(\boldsymbol{x}), x_i\rangle + \mathbf{inf}_z\{\tfrac{1-\gamma_i r_i}{\gamma_i} h_{\boldsymbol{x}}^i(z) - \tfrac{1}{\gamma_i}\langle \nabla_i h(\boldsymbol{x}), z\rangle\}, \end{aligned}$$

which is finite, owing to 1-coercivity of $z \mapsto \frac{1-\gamma_i r_i}{\gamma_i} h_{\boldsymbol{x}}^i(z) - \frac{1}{\gamma_i}\langle \nabla_i h(\boldsymbol{x}), z\rangle$.

We now show $(b) \Leftrightarrow (c)$. Suppose that $\alpha_i := \mathbf{inf} g_i + r_i h_{\boldsymbol{x}}^i > -\infty$. Since $h_{\boldsymbol{x}}^i(\cdot)$ is 1-coercive, we have

$$\mathbf{lim\,inf}_{\|z\|\to\infty} \tfrac{g_i(z)}{h_{\boldsymbol{x}}^i(z)} \geq -r_i + \mathbf{lim\,inf}_{\|z\|\to\infty} \tfrac{\alpha_i}{h_{\boldsymbol{x}}^i(z)} = -r_i > -\infty.$$

Conversely, suppose $\mathbf{lim\,inf}_{\|z\|\to\infty} \frac{g_i(z)}{h_{\boldsymbol{x}}^i(z)} > -\infty$. Then, there exist $\ell_i, M_i \in \mathbb{R}$ such that $\frac{g_i(z)}{h_{\boldsymbol{x}}^i(z)} \geq \ell_i$ whenever $\|z\| \geq M_i$. Thus, it holds that

$$\mathbf{inf}_{\|z\|\geq M_i} g_i(z) + r_i h_{\boldsymbol{x}}^i(z) \geq \mathbf{inf}_{\|x\|\geq M_i} h_{\boldsymbol{x}}^i(z)(\ell_i + r_i) > -\infty,$$

where the last inequality follows from coercivity of $h_{\boldsymbol{x}}^i$, i,e, $h_{\boldsymbol{x}}^i$ is level-bounded and consequently lower bounded (cf. [55, Corollary 1.10]). Since the function $g_i(z) + r_i h_{\boldsymbol{x}}^i(z)$ is lower semicontinuous, it follows from [55, Corollary 1.10] that $\mathbf{inf}_{\|z\|\leq M_i} g_i(z) + r_i h_{\boldsymbol{x}}^i(z) > -\infty$. Therefore, we conclude that $g_i + r_i h_{\boldsymbol{x}}^i$ is lower bounded on $\mathbb{R}^n$. □

Let us now define the function $\mathcal{M}_{h/\gamma} : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ as

$$\mathcal{M}_{h/\gamma}(\boldsymbol{z}, \boldsymbol{x}) := f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x}\rangle + \tfrac{1}{\gamma}\mathbf{D}_h(\boldsymbol{z}, \boldsymbol{x}) + \sum_{i=1}^N g_i(z_i) \tag{3.8}$$

and the set-valued *Bregman proximal alternating linearized mapping* $\mathbf{T}_{h/\gamma i}^i : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n_i}$ as

$$\mathbf{T}_{h/\gamma i}^i(\boldsymbol{x}) := \mathop{\mathbf{arg\,min}}_{z\in\mathbb{R}^{n_i}} \mathcal{M}_{h/\gamma_i}(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x}), \tag{3.9}$$

which reduces to the Bregman forward-backward splitting mapping if $N = 1$; cf. [3, 24].

Remark 3.8 (majorization model) Note that, for $\boldsymbol{x} \in \mathbf{int\,dom}\, h$, invoking Proposition 3.5(b) and Assumption 1, the multi-block $(L_1, \ldots, L_N)$-relative smoothness assumption of f entails a *block majorization model*

$$\begin{aligned}\varphi(\boldsymbol{x}+U_i(y_i-x_i)) &\le f(\boldsymbol{x})+\langle\nabla_i f(\boldsymbol{x}),y_i-x_i\rangle+L_i\mathbf{D}_h(\boldsymbol{x}+U_i(y_i-x_i),\boldsymbol{x})+g_i(y_i)+\sum_{j\neq i}g_j(x_j)\\ &\le f(\boldsymbol{x})+\langle\nabla_i f(\boldsymbol{x}),y_i-x_i\rangle+\tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}+U_i(y_i-x_i),\boldsymbol{x})+g_i(y_i)+\sum_{j\neq i}g_j(x_j),\end{aligned}$$

for $\gamma_i\in(0,1/L_i)$.

In the next lemma, we show that the cost function φ is monotonically decreasing by minimizing the model (3.8) with respect to each block of variables.

Lemma 3.9 (Bregman proximal alternating inequality) *Let the conditions in Assumption 1 hold, and let* $\bar{z}\in\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x})$ *with* $\gamma_i\in(0,1/L_i)$ *and* $\boldsymbol{x}\in\mathbf{int\,dom}\,h$. *Then,*

$$\varphi(\boldsymbol{x}+U_i(\bar{z}-x_i))\le\varphi(\boldsymbol{x})-\tfrac{1-\gamma_iL_i}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}+U_i(\bar{z}-x_i),\boldsymbol{x}),\tag{3.10}$$

for all $i=1,\dots,N$.

Proof For $i\in\{1,\dots,N\}$, (3.9) is simplified in the form

$$\begin{aligned}\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x}) &=\mathbf{arg\,min}_{z\in\mathbb{R}^{n_i}}\{\langle\nabla f(\boldsymbol{x}),U_i(z-x_i)\rangle+\tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}+U_i(z-x_i),\boldsymbol{x})+g_i(z)+\textstyle\sum_{j\neq i}^N g_j(x_j)\}\\ &=\mathbf{arg\,min}_{z\in\mathbb{R}^{n_i}}\{\langle\nabla_i f(\boldsymbol{x}),z-x_i\rangle+\tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}+U_i(z-x_i),\boldsymbol{x})+g_i(z)\}.\end{aligned}\tag{3.11}$$

Considering $\bar{z}\in\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x})$, we have

$$\langle\nabla_i f(\boldsymbol{x}),\bar{z}-x_i\rangle+\tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}+U_i(\bar{z}-x_i),\boldsymbol{x})+g_i(\bar{z})\le g_i(x_i).$$

Since f is $(L_1,\dots,L_N)$-smooth relative to h, it follows from Proposition 3.5(b) for $\boldsymbol{x}$ and $y_i=\bar{z}$ that

$$\begin{aligned}f(\boldsymbol{x}+U_i(\bar{z}-x_i)) &\le f(\boldsymbol{x})+\langle\nabla_i f(\boldsymbol{x}),\bar{z}-x_i\rangle+L_i\mathbf{D}_h(\boldsymbol{x}+U_i(\bar{z}-x_i),\boldsymbol{x})\\ &\le f(\boldsymbol{x})+L_i\mathbf{D}_h(\boldsymbol{x}+U_i(\bar{z}-x_i),\boldsymbol{x})+g_i(x_i)-g_i(\bar{z})-\tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}+U_i(\bar{z}-x_i),\boldsymbol{x})\\ &= f(\boldsymbol{x})+g_i(x_i)-g_i(\bar{z})-\tfrac{1-\gamma_iL_i}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}+U_i(\bar{z}-x_i),\boldsymbol{x}),\end{aligned}$$

giving (3.10).

Recall that a function $\vartheta:\mathbb{R}^n\times\mathbb{R}^m\to\overline{\mathbb{R}}$ with values $\vartheta(x,u)$ is *level-bounded* in x locally uniformly in u if for each $\bar{u}\in\mathbb{R}^m$ and $\alpha\in\mathbb{R}$ there is a neighborhood $\mathcal{U}$ of $\bar{u}$ along with a bounded set $B\subset\mathbb{R}^n$ such that $\{x\in\mathbb{R}^n\mid\vartheta(x,u)\le\alpha\}\subset B$ for all $u\in\mathcal{U}$, cf. [55, Definition 1.16]. Using this definition, we next investigate the fundamental properties of the mapping $\mathbf{T}^i_{h/\gamma i}$, which imply that if $\boldsymbol{x}\in\mathbf{int\,dom}\,h$, then the set $\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x})$ is nonempty. This is an essential assertion to show the well-definedness of the algorithm (BPALM) in the next section. Let us remind that such statement is a common hypothesis for methods assuming relative smoothness in nonconvex setting; see, e.g., [24].

Proposition 3.10 (properties of Bregman proximal alternating linearized mapping) *Under conditions given in Assumption 1 and setting* $\gamma_i \in (0, \gamma_{i,g}^h)$ *for* $i = 1, \ldots, N$, *the following statements are true*:

(i) $\mathbf{T}^i_{h/\gamma i}$ *is nonempty, compact, and outer semicontinuous* (*osc*) *for all* $\boldsymbol{x} \in \mathbf{int\,dom}\, h$;
(ii) $\mathbf{dom}\, \mathbf{T}^i_{h/\gamma i} = \mathbf{int\,dom}\, h$;

Proof Let us define the function $\Phi_i : \mathbb{R}^{n_i} \times \mathbf{int\,dom}\, h \times (0, \gamma_{i,g}^h) \to \overline{\mathbb{R}}$ given by

$$\Phi_i(z, \boldsymbol{x}, \gamma_i) := g_i(z) + \langle \nabla_i f(\boldsymbol{x}), z - x_i \rangle + \begin{cases} \frac{1}{\gamma_i} \mathbf{D}_h(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x}) & \text{if } \gamma_i \in (0, \gamma_i^0], \\ 0 & \text{if } \gamma_i = 0 \text{ and } z = x_i, \\ +\infty & \text{otherwise.} \end{cases}$$

Since f is $\mathcal{C}^1$, g_i is proper and lsc, and h is $\mathcal{C}^1$ proper and lsc, the function Φ_i is proper and lsc on the sets of the form $\{(z, \boldsymbol{x}, \gamma_i)\}[\|z - x_i\| \le \mu\gamma_i,\ 0 \le \gamma_i \le \gamma_i^0]$, for any $\mu > 0$ and a fixed $\gamma_i^0 \in (0, \gamma_{i,g}^h)$. We now show that Φ_i is level-bounded in z locally uniformly at every $(\boldsymbol{x}, \gamma_i) \in \mathbf{int\,dom}\, h \times (0, \gamma_i^0)$. If it is not true, then there exist $(\overline{\boldsymbol{x}}, \overline{\gamma}_i) \in \mathbf{int\,dom}\, h \times (0, \gamma_{i,g}^h)$, $(\boldsymbol{x}^k)_{k\in\mathbb{N}} \subset \mathbf{int\,dom}\, h$, $(z^k)_{k\in\mathbb{N}}$ with $\boldsymbol{x}^k + U_i(z^k - x_i^k) \subset \mathbf{int\,dom}\, h$, and $(\gamma_i^k)_{k\in\mathbb{N}} \subset (0, \gamma_i^0]$ such that $\Phi_i(z^k, \boldsymbol{x}^k, \gamma_i^k) \le \beta < \infty$ with $(\boldsymbol{x}^k, \gamma_i^k) \to (\overline{\boldsymbol{x}}, \overline{\gamma}_i)$ and $\|z^k\| \to \infty$. This guarantees that, for sufficiently large k, $z^k \ne x_i^k$, i.e., $\gamma_i^k \in (0, \gamma_i^0]$ and

$$g_i(z^k) + \langle \nabla_i f(\boldsymbol{x}^k), z^k - x_i^k \rangle + \tfrac{1}{\gamma_i^k} \mathbf{D}_h(\boldsymbol{x}^k + U_i(z^k - x_i^k), \boldsymbol{x}^k) \le \beta.$$

Setting $\tilde{\gamma}_i \in (\gamma_i^0, \gamma_i^h)$, Definition 3.6 and Proposition 3.7(b) ensure that there exists $\tilde{\beta} \in \mathbb{R}$ such that

$$g_i(z^k) + \tfrac{1}{\tilde{\gamma}_i} h(\boldsymbol{x}^k + U_i(z^k - x_i^k)) \ge \tilde{\beta}.$$

Subtracting the last two inequalities, it holds that

$$\langle \nabla_i f(\boldsymbol{x}^k), z^k - x_i^k \rangle + \tfrac{1}{\gamma_i^k} \mathbf{D}_h(\boldsymbol{x}^k + U_i(z^k - x_i^k), \boldsymbol{x}^k) - \tfrac{1}{\tilde{\gamma}_i} h(\boldsymbol{x}^k + U_i(z^k - x_i^k)) \le \beta - \tilde{\beta}.$$

By expanding $\mathbf{D}_h(\boldsymbol{x}^k + U_i(z^k - x_i^k), \boldsymbol{x}^k)$, dividing both sides by $\|z^k\|$, and taking limit from both sides of this inequality as $k \to \infty$, it can be deduced that

$$\mathbf{lim}_{k\to\infty} \left(\left\langle \nabla_i f(\boldsymbol{x}^k) - \tfrac{1}{\gamma_i^k} \nabla_i h(\boldsymbol{x}^k), \frac{z^k - x_i^k}{\|z^k\|} \right\rangle - \tfrac{1}{\gamma_i^k} \frac{h(\boldsymbol{x}^k)}{\|z^k\|} \right) + \left(\tfrac{1}{\gamma_i^k} - \tfrac{1}{\tilde{\gamma}_i} \right) \mathbf{lim}_{k\to\infty} \frac{h^i_{\boldsymbol{x}^k}(z^k)}{\|z^k\|} \le \mathbf{lim}_{k\to\infty} \frac{\beta - \tilde{\beta}}{\|z^k\|}.$$

This leads to the contradiction $+\infty \le 0$, which implies that Φ_i is level-bounded in z locally uniformly at every $(\boldsymbol{x}, \gamma_i) \in \mathbf{int\,dom}\, h \times (0, \gamma_i^0)$.

Considering the above result, all assumptions of the parametric minimization theorem given in [55, Theorem 1.17] are satisfied, i.e., Proposition 3.10(i) holds true. If

$\boldsymbol{x} \in \mathbf{int\,dom}\, h$, it follows from Proposition 3.10(i) that $\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x}) \neq \emptyset$, i.e., $\mathbf{int\,dom}\, h \subseteq \mathbf{dom}\, \mathbf{T}^i_{h/\gamma i}$. For $\boldsymbol{x} \in \mathbf{dom}\, \mathbf{T}^i_{h/\gamma i}$, the essential smoothness of h yields $\boldsymbol{x} \in \mathbf{int\,dom}\, h$, i.e., $\mathbf{dom}\, \mathbf{T}^i_{h/\gamma i} \subseteq \mathbf{int\,dom}\, h$, giving Proposition 3.10(ii).

Remark 3.11 (sum or product separable kernel) Let Assumption 1 hold, and let $h_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ $(i = 1, \ldots, N)$ be strictly convex, 1-coercive, and essentially smooth. Then, the following hold:

(i) If h is an *additive separable* function, i.e., $h(x_1, \ldots, x_N) = h_1(x_1) + \ldots + h_N(x_N)$, and $\nabla h_i(x_i) - \gamma_i \nabla_i f(\boldsymbol{x}) \in \mathbf{dom}\, \nabla h_i^*$, then (3.9) can be written in the form

$$\begin{aligned}
\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x}) &= \mathop{\mathbf{arg\,min}}_{z \in \mathbb{R}^{n_i}} \{g_i(z) + \langle \nabla_i f(\boldsymbol{x}), z - x_i\rangle + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x})\} \\
&= \mathop{\mathbf{arg\,min}}_{z \in \mathbb{R}^{n_i}} \{g_i(z) + \tfrac{1}{\gamma_i}(h_i(z) - h_i(x_i) - \langle \nabla h_i(x_i) - \gamma_i \nabla_i f(\boldsymbol{x}), z - x_i\rangle)\} \\
&= \mathop{\mathbf{arg\,min}}_{z \in \mathbb{R}^{n_i}} \{g_i(z) + \tfrac{1}{\gamma_i}\mathbf{D}_{h_i}(z, \nabla h_i^*(\nabla h_i(x_i) - \gamma_i \nabla_i f(\boldsymbol{x}))\} \\
&= \mathbf{prox}^{h_i}_{\gamma_i g_i}(\nabla h_i^*(\nabla h_i(x_i) - \gamma_i \nabla_i f(\boldsymbol{x}))).
\end{aligned}$$

(ii) If h is *product separable*, i.e., $h(x_1, \ldots, x_N) = h_1(x_1) \times \ldots \times h_N(x_N)$, each h_i $(i = 1, \ldots, N)$ is positive, and $\nabla h_i(x_i) - \gamma_i \nabla_i f(\boldsymbol{x}) \in \mathbf{dom}\, \nabla h_i^*$ holds, then

$$\begin{aligned}
\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x}) &= \mathop{\mathbf{arg\,min}}_{z \in \mathbb{R}^{n_i}} \{g_i(z) + \langle \nabla_i f(\boldsymbol{x}), z - x_i\rangle + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x} + U_i(z - x_i), \boldsymbol{x})\} \\
&= \mathop{\mathbf{arg\,min}}_{z \in \mathbb{R}^{n_i}} \{g_i(z) + \tfrac{\eta^i_{\boldsymbol{x}}}{\gamma_i}(h_i(z) - h_i(x_i) - \langle \nabla h_i(x_i) - \tfrac{\gamma_i}{\eta^i_{\boldsymbol{x}}} \nabla_i f(\boldsymbol{x}), z - x_i\rangle)\} \\
&= \mathop{\mathbf{arg\,min}}_{z \in \mathbb{R}^{n_i}} \{g_i(z) + \tfrac{1}{\mu_i}\mathbf{D}_{h_i}(z, \nabla h_i^*(\nabla h_i(x_i) - \mu_i \nabla_i f(\boldsymbol{x}))\} \\
&= \mathbf{prox}^{h_i}_{\mu_i g_i}(\nabla h_i^*(\nabla h_i(x_i) - \mu_i \nabla_i f(\boldsymbol{x}))),
\end{aligned}$$

where $\mu_i := \frac{\gamma_i}{\eta^i_{\boldsymbol{x}}}$ and $\eta^i_{\boldsymbol{x}} := \prod_{j \neq i} h_j(x_j) > 0$.

Note that in above-mentioned multi-block kernel function, h is not necessarily convex, but it is only multi-block convex that is a much weaker notion than the convexity.

4 Multi-block Bregman proximal alternating linearized minimization

We here introduce a multi-block proximal alternating linearized minimization algorithm and investigate its subsequential and global convergence, along with its convergence rate.

For given points $\boldsymbol{x}^k = (x_1^k, \ldots, x_N^k)$ and $\boldsymbol{x}^{k+1} = (x_1^{k+1}, \ldots, x_N^{k+1})$, we set

$$\boldsymbol{x}^{k,i} := (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_N^k) = \boldsymbol{x}^k + \sum_{j=1}^{i} U_j(x_j^{k+1} - x_j^k),$$

so that $\boldsymbol{x}^{k,0} = \boldsymbol{x}^k$ and $\boldsymbol{x}^{k,N} = \boldsymbol{x}^{k+1}$, where $U_j \in \mathbb{R}^{n\times n_j}$, $I_n = (U_1, \ldots, U_N) \in \mathbb{R}^{n\times n}$ for the identity matrix I_n. Using this notation and (3.9), we next introduce the multi-block Bregman proximal alternating linearized minimization (BPALM) algorithm.

Algorithm 1 (**BPALM**) Bregman Proximal Alternating Linearized Minimization

REQUIRE $\gamma_i \in (0, 1/L_i)$, $i = 1, \ldots, N$, $\boldsymbol{x}^0 \in \mathbf{int\,dom}\, h$, $I_n = (U_1, \ldots, U_N) \in \mathbb{R}^{n\times n}$ with $U_i \in \mathbb{R}^{n\times n_i}$ and the identity matrix I_n.

INITIALIZE $k = 0$.

1: **while** some stopping criterion is not met **do**

2: $\quad \boldsymbol{x}^{k,0} = \boldsymbol{x}^k$;

3: $\quad$ **for** $i = 1, \ldots, N$ **do**

$$x_i^{k,i} \in \mathbf{T}^i_{h/\gamma_i}(\boldsymbol{x}^{k,i-1}), \quad \boldsymbol{x}^{k,i} = \boldsymbol{x}^{k,i-1} + U_i(x_i^{k,i} - x_i^{k,i-1}); \tag{4.1}$$

4: $\quad \boldsymbol{x}^{k+1} = \boldsymbol{x}^{k,N}$, $k = k + 1$;

ENSURE A vector $\boldsymbol{x}^k$.

We note that each iteration of BPALM requires one call of the first-order oracle for the information needed in (4.1). In addition, notice that if $N = 1$, this algorithm reduces to the common (Bregman) proximal gradient (forward-backward) method [13, 17, 24]; if $N = 2$, $h, f \in \mathcal{C}^2$, and $g_1 = g_2 = 0$, then it reduces to B-PALM [43]; if $N = 2$ and $h(\boldsymbol{x}) = \frac{1}{2}(\|x_1\|^2 + \|x_2\|^2)$, it reduces to PALM [23]; if $h(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^N \|x_i\|^2$, then this algorithm is reduced to C-PALM [57].

From Proposition 3.10, we know that the operator $\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x}^{k,i-1})$ is nonempty and compact for all $\boldsymbol{x}^{k,i-1} \in \mathbf{int\,dom}\, h$; however, after the substitution $\boldsymbol{x}^{k,i} = \boldsymbol{x}^{k,i-1} + U_i(x_i^{k,i} - x_i^{k,i-1})$, we need to be sure $\boldsymbol{x}^{k,i} \in \mathbf{int\,dom}\, h$ to guarantee the nonemptymess of $\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x}^{k,i})$ in the next step of the algorithm. To do so, we require to make the following extra assumption:

Assumption 2 For $i = 1, \ldots, N$ and all $z \in \mathbf{range}\mathbf{T}^i_{h/\gamma i}(\boldsymbol{x})$, we have $\boldsymbol{x} + U_i(z - x_i) \in \mathbf{int\,dom}\, h$.

Let us emphasize that together with Proposition 3.10, the latter assumption implies that BPALM is well-defined. Now, we can begin with showing some basic properties of the sequence generated by BPALM, involving a *sufficient decrease condition*.

Proposition 4.1 **(sufficient decrease condition)** *Let Assumption 1 and 2 hold, and let* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *be generated by* BPALM. *Then, the following statements are true*:

(i) *the sequence* $(\varphi(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ *is nonincreasing and*

$$\rho \sum_{i=1}^N \mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1}) \le \varphi(\boldsymbol{x}^k) - \varphi(\boldsymbol{x}^{k+1}), \tag{4.2}$$

where $\rho := \min\left\{\frac{1-\gamma_1 L_1}{\gamma_1}, \ldots, \frac{1-\gamma_N L_N}{\gamma_N}\right\}$;

(ii) *we have*

$$\sum_{k=1}^{\infty}\sum_{i=1}^{N}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1}) < \infty, \tag{4.3}$$

i.e., $\lim_{k\to\infty}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1}) = 0$ for $i = 1,\ldots,N$.

Proof Plugging $\bar{z} = x_i^{k,i}$ and $\boldsymbol{x} = \boldsymbol{x}^{k,i-1}$ into Lemma 3.9, it holds that

$$\varphi(\boldsymbol{x}^{k,i}) \le \varphi(\boldsymbol{x}^{k,i-1}) - \tfrac{1-\gamma_i L_i}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1}). \tag{4.4}$$

Summing up both sides of (4.4) from $i = 1$ to N, it follows that

$$\sum_{i=1}^{N}\tfrac{1-\gamma_i L_i}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1}) \le \sum_{i=1}^{N}[\varphi(\boldsymbol{x}^{k,i-1}) - \varphi(\boldsymbol{x}^{k,i})] = \varphi(\boldsymbol{x}^{k}) - \varphi(\boldsymbol{x}^{k+1}),$$

giving (4.2). Let us sum up both sides of (4.2) from $k = 0$ to q:

$$\rho\sum_{k=0}^{q}\sum_{i=1}^{N}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1}) \le \sum_{k=0}^{q}\varphi(\boldsymbol{x}^{k}) - \varphi(\boldsymbol{x}^{k+1}) = \varphi(\boldsymbol{x}^{0}) - \varphi(\boldsymbol{x}^{q+1}) \le \varphi(\boldsymbol{x}^{0}) - \mathbf{inf}\,\varphi < \infty.$$

Taking the limit as $q \to +\infty$, (4.3) holds true. Together with $\mathbf{D}_h(\cdot,\cdot) \ge 0$, this proves the claim.

Let us consider the condition

$$\sum_{i=1}^{N}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1}) \le \varepsilon \tag{4.5}$$

as a *stopping criterion*, for the accuracy parameter $\varepsilon > 0$. Then, the first main consequence of Proposition 4.1 will provide us the *iteration complexity* of BPALM, which is the number of iterations needed for the stopping criterion (4.5) to be satisfied.

The boundedness of the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is a typical assumption in convergence analysis of proximal-type algorithms for solving general non-convex non-smooth composite optimization problem, see e.g., [7, 24]. We next provide the iteration complexity of BPALM and a sufficient condition guaranteeing the boundedness of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ as a simple consequence of Proposition 4.1.

Corollary 4.2 (iteration complexity and boundedness) *Let Assumption 1 and 2 hold, and let the sequence* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *be generated by* BPALM *with the stopping criterion (4.5). Then, for the constant* $\rho = \min\{\frac{1-\gamma_1 L_1}{\gamma_1}, \ldots, \frac{1-\gamma_N L_N}{\gamma_N}\}$,

(i) BPALM *will be terminated within* $k \le 1 + \frac{\varphi(x^0)-\mathbf{inf}\,\varphi}{\rho\varepsilon}$ *iterations*;

(ii) *If* φ *has bounded level sets, then the sequence* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *is bounded.*

Proof Summing both sides of (4.2) over the first $\mathcal{K} \in \mathbb{N}$ iterations and telescoping the right hand side, it holds that

$$\rho \sum_{k=0}^{\mathcal{K}-1} \sum_{i=1}^{N} \mathbf{D}_h(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) \leq \sum_{k=0}^{\mathcal{K}-1} \left(\varphi(\boldsymbol{x}^k) - \varphi(\boldsymbol{x}^{k+1})\right) = \varphi(\boldsymbol{x}^0) - \varphi(\boldsymbol{x}^{\mathcal{K}}) \leq \varphi(\mathbf{x}^0) - \mathbf{inf}\,\varphi.$$

Assuming that for all the first $(\mathcal{K}-1)$ iterations the stopping criterion (4.5) is not satisfied, i.e., $\sum_{i=1}^{N} \mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1}) > \varepsilon$, which leads to $\mathcal{K} \leq 1 + \frac{\varphi(x^0) - \mathbf{inf}\,\varphi}{\rho\varepsilon}$, giving the desired result.

Proposition 4.1(i) shows that $\varphi(\boldsymbol{x}^k)$ is non-increasing; hence, the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is encompassed within the lower level set $\left[\varphi \leq \varphi(\boldsymbol{x}^0)\right]$; i.e., $(\boldsymbol{x}^k)_{k\in\mathbb{N}} \subseteq \left[\varphi \leq \varphi(\boldsymbol{x}^0)\right]$. Since φ has bounded level sets, the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ would be bounded.

In order to show subsequential convergence of the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ generated by BPALM, the next proposition will provide a lower bound for iterations gap $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|$ using the subdifferential of $\partial\varphi(\boldsymbol{x}^{k+1})$. In the remainder of this section, we assume that the kernel h has a full domain, i.e., $\overline{C} = \mathbb{R}^n$.

Assumption 3 The multi-block Legendre kernel h has full domain, i.e., $\mathbf{dom}\, h = \mathbb{R}^n$.

Proposition 4.3 (subgradient lower bound for iterations gap) *Let Assumption 1, 2 and 3 hold, and let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ be generated by* BPALM *that we assume to be bounded. For a fixed $k \in \mathbb{N}$, we define*

$$\mathcal{G}_i^{k+1} := \tfrac{1}{\gamma_i}(\nabla_i h(\boldsymbol{x}^{k,i-1}) - \nabla_i h(\boldsymbol{x}^{k,i})) + \nabla_i f(\boldsymbol{x}^{k+1}) - \nabla_i f(\boldsymbol{x}^{k,i-1}) \quad i = 1, \ldots, N. \tag{4.6}$$

If $\nabla_i f$ and $\nabla_i h$ $(i = 1, \ldots, N)$ are locally Lipschitz on bounded sets involving all iterations, then $\left(\mathcal{G}_1^{k+1}, \ldots, \mathcal{G}_N^{k+1}\right) \in \partial\varphi(\boldsymbol{x}^{k+1})$ and

$$\|(\mathcal{G}_1^{k+1}, \ldots, \mathcal{G}_N^{k+1})\| \leq \overline{c} \sum_{i=1}^{N} \|x_i^{k+1} - x_i^k\|, \tag{4.7}$$

with $\overline{c} := \mathbf{max}\{\frac{\widetilde{L}+\gamma_1\widehat{L}}{\gamma_1}, \ldots, \frac{\widetilde{L}+\gamma_N\widehat{L}}{\gamma_N}\}$ in which $\widehat{L}$ and $\widetilde{L} > 0$ are Lipschitz moduli of $\nabla_i f$, $\nabla_i h$ $(i = 1, \ldots, N)$ on bounded sets involving all iterations.

Proof The optimality conditions for (4.1) ensure that there exist $q_i^{k,i} \in \partial g_i(x_i^{k,i})$ such that

$$\nabla_i f(\boldsymbol{x}^{k,i-1}) + \tfrac{1}{\gamma_i}\left(\nabla_i h(\boldsymbol{x}^{k,i}) - \nabla_i h(\boldsymbol{x}^{k,i-1})\right) + q^{k,i} = 0 \quad i = 1, \ldots, N,$$

leading to

$$q^{k,i} = \tfrac{1}{\gamma_i}\left(\nabla_i h(\boldsymbol{x}^{k,i-1}) - \nabla_i h(\boldsymbol{x}^{k,i})\right) - \nabla_i f(\boldsymbol{x}^{k,i-1}) \quad i = 1, \ldots, N. \tag{4.8}$$

On the other hand, owing to [7, Proposition 2.1], the subdifferential of φ is given by

$$\partial\varphi(\boldsymbol{x}) = (\partial_1\varphi(\boldsymbol{x}), \dots, \partial_N\varphi(\boldsymbol{x})) = (\nabla_1 f(\boldsymbol{x}) + \partial g_1(x_1), \dots, \nabla_N f(\boldsymbol{x}) + \partial g_N(x_N)),$$

i.e., for $\boldsymbol{x} = \boldsymbol{x}^{k+1}$,

$$\nabla_i f(\boldsymbol{x}^{k+1}) + \partial g_i(x_i^{k+1}) = \partial_i\varphi(\boldsymbol{x}^{k+1}) \quad i = 1, \dots, N,$$

which means $(\mathcal{G}_1^{k+1}, \dots, \mathcal{G}_N^{k+1}) \in \partial\varphi(\boldsymbol{x}^{k+1})$. It follows from Assumption 3 and the Lipschitz continuity of $\nabla_i f$, $\nabla_i h$ on bounded sets involving all iterations and the assumption of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ being bounded that there exist $\widehat{L}, \widetilde{L} > 0$ such that

$$\|\mathcal{G}_i^{k+1}\| \le \tfrac{1}{\gamma_i}\|\nabla_i h(\boldsymbol{x}^{k,i-1}) - \nabla_i h(\boldsymbol{x}^{k,i})\| + \|\nabla_i f(\boldsymbol{x}^{k+1}) - \nabla_i f(\boldsymbol{x}^{k,i-1})\| \le \tfrac{\widetilde{L}+\gamma_i\widehat{L}}{\gamma_i}\sum_{j=1}^{N}\|x_j^{k+1} - x_j^k\|,$$

for $i = 1, \dots, N$. Invoking the last two inequalities, it can be concluded that

$$\|(\mathcal{G}_1^{k+1}, \dots, \mathcal{G}_N^{k+1})\| \le \mathbf{max}\{\tfrac{\widetilde{L}+\gamma_1\widehat{L}}{\gamma_1}, \dots, \tfrac{\widetilde{L}+\gamma_N\widehat{L}}{\gamma_N}\}\sum_{i=1}^{N}\|x_i^{k+1} - x_i^k\|,$$

as claimed.

Next, we proceed to derive the *subsequential convergence* of the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ generated by BPALM: every cluster point of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is a critical point of φ. Further, we explain some fundamental properties of the set of all cluster points $\omega(\boldsymbol{x}^0)$ of the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$.

Theorem 4.4 (**subsequential convergence and properties of** $\omega(\boldsymbol{x}^0)$) *Let Assumption 1, 2 and 3 hold, let the kernel h be locally multi-block strongly convex, and let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ be generated by* BPALM. *If the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is bounded and $\nabla_i f$ and $\nabla_i h$ $(i = 1, \dots, N)$ are locally Lipschitz around its cluster points, then all the cluster points of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ are critical points φ, i.e., $\omega(\boldsymbol{x}^0) \subset$ **crit** φ.*

Proof For a limit point $\boldsymbol{x}^\star = (x_1^\star, \dots, x_N^\star)$ of the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$, it follows from the boundedness of this sequence that there exists an infinite index set $\mathcal{J} \subset \mathbb{N}$ such that the subsequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}[k \in \mathcal{J}]$ converges to $\boldsymbol{x}^\star$ as $k \to \infty$. From the lower semicontinuity of g_i $(i = 1, \dots, N)$ and for $j \in \mathcal{J}$, it can be deduced that

$$\mathbf{lim\,inf}_{j\to\infty} g_i(x_i^{k_j}) \ge g(x_i^\star) \quad i = 1, \dots, N. \tag{4.9}$$

By (4.1), we get

$$\begin{aligned}\langle\nabla_i f(\boldsymbol{x}^{k,i-1}), x_i^{k+1} - x_i^k\rangle + \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1}) + g_i(x_i^{k+1}) \quad &\le \langle\nabla_i f(\boldsymbol{x}^{k,i-1}), x_i^\star - x_i^k\rangle \\ &+ \tfrac{1}{\gamma_i}\mathbf{D}_h(\boldsymbol{x}^\star, \boldsymbol{x}^{k,i-1}) + g_i(x_i^\star).\end{aligned} \tag{4.10}$$

Using multi-block local strong convexity of h around $\boldsymbol{x}^\star$ and invoking Proposition 4.1(ii), there exist a neighborhood $\mathbf{B}(x_i^\star, \varepsilon_i^\star)$ for $\varepsilon_i^\star > 0$, $\sigma_i^\star > 0$, and $k_i^0 \in \mathbb{N}$ such that for $k \ge k_i^0$ and $k \in \mathcal{J}$

$$\mathbf{lim}_{k\to\infty}\tfrac{\sigma_i^\star}{2}\|x_i^{k+1}-x_i^k\|^2 \le \mathbf{lim}_{k\to\infty}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1})=0,\quad x_i^k\in\mathbf{B}(x_i^\star,\varepsilon_i^\star),\ i=1,\ldots,N. \tag{4.11}$$

This indicates that the distance between two successive iterations goes to zero for large enough k, i.e., for $\mathbf{lim}_{k\to\infty}x_i^{k+1}=\mathbf{lim}_{k\to\infty}x_i^k$. Together with the boundedness of the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$, the continuity of $\nabla_i f$ and $\nabla_i h$ around cluster points of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$, the substitution of $k=k_j-1$ in (4.10) for $j\in\mathcal{J}$, taking the limit in both sides of this inequality as $k\to\infty$, this implies that

$$\mathbf{lim\,sup}_{j\to\infty}g_i(x_i^{k_j})\le g_i(x_i^\star)\quad i=1,\ldots,N,$$

and consequently by lsc of g_i,

$$\mathbf{lim}_{j\to\infty}\varphi(\boldsymbol{x}^{k_j})=\mathbf{lim}_{j\to\infty}\left(f(x_1^{k_j},\ldots,x_N^{k_j})+\sum_{i=1}^N g_i(x_i^{k_j})\right)=f(x_1^\star,\ldots,x_N^\star)+\sum_{i=1}^N g_i(x_i^\star). \tag{4.12}$$

Further, Proposition 4.1(ii) and Proposition 4.3 ensure $\left(\mathcal{G}_1^{k+1},\ldots,\mathcal{G}_N^{k+1}\right)\in\partial\varphi(\boldsymbol{x}^{k+1})$ and

$$\mathbf{lim}_{k\to+\infty}\|\mathcal{G}_1^{k+1},\ldots,\mathcal{G}_N^{k+1}\|\le\bar{c}\mathbf{lim}_{k\to+\infty}\sum_{i=1}^N\|x_i^{k+1}-x_i^k\|\le\bar{c}\mathbf{lim}_{k\to+\infty}\left(\sum_{i=1}^N\sqrt{\tfrac{2}{\sigma_i^\star}\mathbf{D}_h(\boldsymbol{x}^{k,i},\boldsymbol{x}^{k,i-1})}\right)=0,$$

i.e., $\mathbf{lim}_{k\to\infty}\left(\mathcal{G}_1^{k+1},\ldots,\mathcal{G}_N^{k+1}\right)=(0_{n_1},\ldots,0_{n_N})$. Together with the closedness of the subdifferential mapping $\partial\varphi$ and (4.12), this implies that $(0_{n_1},\ldots,0_{n_N})\in\partial\varphi(x_1^\star,\ldots,x_N^\star)$, giving our desired result.

4.1 Global convergence under Kurdyka-Łojasiewicz inequality

Our following main result indicates that the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ generated by BPALM converges to a critical point $x^\star$ of φ if it satisfies the KL property; cf. Definition 2.1.

Theorem 4.5 (global convergence) *Let Assumption 1, 2 and 3 hold, let the kernels h be multi-block globally strongly convex with modulus σ_i $(i=1,\ldots,N)$, and let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ be generated by* BPALM *that we assume to be bounded. If φ is a KL function, then the following statements are true*:

(i) *The sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ has finite length, i.e.,*

$$\sum_{k=1}^\infty\|x_i^{k+1}-x_i^k\|<\infty\quad i=1,\ldots,N; \tag{4.13}$$

(ii) *The sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ converges to a stationary point $\boldsymbol{x}^\star$ of φ.*

Proof Let us set $\varphi^\star := \mathbf{lim}_{k\to\infty}\varphi(\boldsymbol{x}^k)$ and define the sequence $(\mathcal{S}_k)_{k\in\mathbb{N}}$ given by $\mathcal{S}_k := \varphi(\boldsymbol{x}^k) - \varphi^\star$, which is decreasing by Proposition 4.1(i), i.e., $(\mathcal{S}_k)_{k\in\mathbb{N}} \to 0$. We now consider two cases: (i) there exists $\bar{k} \in \mathbb{N}$ such that $\mathcal{S}_k = 0$; (ii) $\mathcal{S}_k > 0$ for all $k \geq 1$.

In Case (i), Proposition 4.1(i) implies that $\varphi(\boldsymbol{x}^k) = \varphi^\star$ for all $k \geq \bar{k}$. It follows from Proposition 4.1(ii) and multi-block strong convexity of h that

$$\tfrac{\sigma_i}{2}\|x_i^{k+1} - x_i^k\|^2 \leq \mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1}) = 0 \quad i = 1, \ldots, N,$$

implying $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k$ for all $k \geq \bar{k}$, which leads to Theorem 4.5(i).

In Case (ii), it holds that $\varphi(\boldsymbol{x}^k) > \varphi^\star$ for all $k \geq 1$. Thus, Lemma 8.1 yields that there exists $\varepsilon, \eta > 0$ and the desingularizing function ψ such that

$$\psi'(\varphi(\boldsymbol{x}^k) - \varphi^\star)\mathbf{dist}(0, \partial\varphi(\boldsymbol{x}^k)) \geq 1 \quad \text{for } k \geq k_0.$$

Let us define $\Delta_k := \psi(\varphi(\boldsymbol{x}^k) - \varphi^\star) = \psi(\mathcal{S}_k)$. Then, it follows from the concavity of ψ and Proposition 4.3 that

$$\begin{aligned}\Delta_k - \Delta_{k+1} &= \psi(\mathcal{S}_k) - \psi(\mathcal{S}_{k+1}) \geq \psi'(\mathcal{S}_k)(\mathcal{S}_k - \mathcal{S}_{k+1}) = \psi'(\mathcal{S}_k)(\varphi(\boldsymbol{x}^k) - \varphi(\boldsymbol{x}^{k+1}))\\ &\geq \frac{\varphi(\boldsymbol{x}^k) - \varphi(\boldsymbol{x}^{k+1})}{\mathbf{dist}(0, \partial\varphi(\boldsymbol{x}^k))} \geq \frac{\rho\sum_{i=1}^N \mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1})}{\bar{c}\sum_{i=1}^N \|x_i^k - x_i^{k-1}\|} \geq \frac{1}{\widehat{c}}\frac{\sum_{i=1}^N \|x_i^{k+1} - x_i^k\|^2}{\sum_{i=1}^N \|x_i^k - x_i^{k-1}\|},\end{aligned}$$

with $\widehat{c} := \frac{\bar{c}}{(\rho\min\{\sigma_1,\ldots,\sigma_N\})}$. Using the arithmetic and quadratic mean inequalities, and applying the arithmetic and geometric mean inequalities, it can be concluded that

$$\textstyle\sum_{i=1}^N \|x_i^{k+1} - x_i^k\| \leq \sqrt{\widehat{c}N(\Delta_k - \Delta_{k+1})\sum_{i=1}^N \|x_i^k - x_i^{k-1}\|} \leq \frac{1}{2}\sum_{i=1}^N \|x_i^k - x_i^{k-1}\| + \frac{\widehat{c}N}{2}(\Delta_k - \Delta_{k+1}). \tag{4.14}$$

We now define the sequences $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ as

$$a_{k+1} := \sum_{i=1}^N \|x_i^{k+1} - x_i^k\|, \quad b_k = \tfrac{\widehat{c}N}{2}\big(\Delta_k - \Delta_{k+1}\big), \quad \alpha := \tfrac{1}{2}, \tag{4.15}$$

where $\sum_{i=1}^\infty b_k = \frac{\widehat{c}N}{2}\sum_{i=1}^\infty \big(\Delta_i - \Delta_{i+1}\big) = \frac{\widehat{c}N}{2}(\Delta_1 - \Delta_\infty) = \frac{\widehat{c}N}{2}\Delta_1 < \infty$. According to Fact 2.3, we infer $\sum_{k=1}^\infty a_k < \infty$, which proves Theorem 4.5(i).

By (4.13), the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is a Cauchy sequence, i.e., it converges to a stationary point $\boldsymbol{x}^\star$, giving the desired result.

4.2 Convergence rate under Łojasiewicz-type inequality

We now investigate the convergence rate of the sequence generated by BPALM under KL inequality of Łojasiewicz type at $x^\star$ ($\psi(s) := \frac{\kappa}{1-\theta}s^{1-\theta}$ with $\theta \in [0, 1)$), i.e., there exists $\varepsilon > 0$ such that

$$|\varphi(\boldsymbol{x}) - \varphi^\star|^\theta \leq \kappa \mathbf{dist}(0, \partial\varphi(\boldsymbol{x})) \quad \forall \boldsymbol{x} \in \mathbf{B}(\boldsymbol{x}^\star; \varepsilon). \tag{4.16}$$

The following fact plays a key role in studying the convergence rate of the sequence generated by BPALM.

Fact 4.6 (convergence rate of a sequence with positive elements) [26, Lemma 15] Let $(s_k)_{k\in\mathbb{N}}$ be a monotonically decreasing sequence in $\mathbb{R}_+$ and let $\theta \in [0, 1)$ and $\beta > 0$. Suppose that $s_k^{2\theta} \leq \beta(s_k - s_{k+1})$ holds for all $k \in \mathbb{N}$. Then, the following assertions hold:

(i) If $\theta = 0$, the sequences $(s_k)_{k\in\mathbb{N}}$ converges in a finite time;
(ii) If $\theta \in (0, 1/2]$, there exist $\lambda > 0$ and $\tau \in [0, 1)$ such that for every $k \in \mathbb{N}$

$$0 \leq s_k \leq \lambda\tau^k;$$

(iii) If $\theta \in (1/2, 1)$, there exists $\mu > 0$ such that for every $k \in \mathbb{N}$

$$0 \leq s_k \leq \mu k^{-\frac{1}{2\theta-1}}.$$

We next derive the *convergence rates* of the sequences $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\varphi(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ under the additional assumption that the function φ satisfies the KL inequality of Łojasiewicz type.

Theorem 4.7 (convergence rate) *Let Assumption 1, 2 and 3 hold, let the kernel h be multi-block globally strongly convex with modulus $(\sigma_1, \ldots, \sigma_N)$, and let the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ generated by* BPALM *converging to $\boldsymbol{x}^\star$. If φ satisfies the KL inequality of Łojasiewicz type* (4.16)*, then the following assertions hold*:

(i) *If $\theta = 0$, then the sequences $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\varphi(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ converge in a finite number of steps to $\boldsymbol{x}^\star$ and $\varphi(\boldsymbol{x}^\star)$, respectively*;
(ii) *If $\theta \in (0, 1/2]$, then there exist $\lambda_1 > 0$, $\mu_1 > 0$, $\tau \in [0, 1)$, and $\overline{k} \in \mathbb{N}$ such that*

$$0 \leq \|\boldsymbol{x}^k - \boldsymbol{x}^\star\| \leq \lambda_1\tau^{\frac{k}{2}}, \quad 0 \leq \mathcal{S}_k \leq \mu_1\tau^k \quad \forall k \geq \overline{k};$$

(iii) *If $\theta \in (1/2, 1)$, then there exist $\lambda_2 > 0$, $\mu_2 > 0$, and $\overline{k} \in \mathbb{N}$ such that*

$$0 \leq \|\boldsymbol{x}^k - \boldsymbol{x}^\star\| \leq \lambda_2 k^{-\frac{1-\theta}{2\theta-1}}, \quad 0 \leq \mathcal{S}_k \leq \mu_2 k^{-\frac{1}{2\theta-1}} \quad \forall k \geq \overline{k} + 1.$$

Proof See the proof in the appendix.

4.3 Adaptive BPALM

The tightness of the i-th block upper bound of the function f given in Proposition 3.5(b) is dependent on the parameter $L_i > 0$; however, in general, this

parameter is a global information and it might not be tight locally, i.e., one may find a $L_i > \overline{L}_i(\boldsymbol{x}) \geq 0$ such that

$$f(\boldsymbol{x} + U_i(y_i - x_i)) \leq f(\boldsymbol{x}) + \langle \nabla_i f(\boldsymbol{x}), y_i - x_i \rangle + \overline{L}_i(\boldsymbol{x}) \mathbf{D}_h(\boldsymbol{x} + U_i(y_i - x_i), \boldsymbol{x})$$

for all $\boldsymbol{y} \in \mathbf{B}(\boldsymbol{x}; \varepsilon_1)$ with a small enough $\varepsilon_1 > 0$. Consequently, the block majorization model described by $\mathcal{M}_{h/\gamma_i}$ may not be tight enough, which will consequently lead to smaller step-sizes $\gamma_i \in (0, 1/L_i)$. In this case and in the case where $(L_1, \ldots, L_N)$ is not available, one can retrieve them adaptively by applying a line search starting from a lower estimate; see, e.g., [3, 5, 47, 49].

Putting together the above discussions, we propose an adaptive version of BPALM using a line search; see Algorithm 2.

Algorithm 2 (**A-BPALM**) adaptive BPALM

REQUIRE $\boldsymbol{x}^0 \in \mathbf{int}\,\mathbf{dom}\, h$, $\nu > 1$, $\overline{L}_i^0 > 0$ for $i = 1, \ldots, N$, $I_n = (U_1, \ldots, U_N) \in \mathbb{R}^{n \times n}$ with $U_i \in \mathbb{R}^{n \times n_i}$ and the identity matrix I_n.

INITIALIZE $k = 0$, $p = 0$, $\gamma_i^0 \in (0, 1/\overline{L}_i^0)$ for $i = 1, \ldots, N$.

1: **while** some stopping criterion is not met **do**
2: $\quad \boldsymbol{x}^{k,0} = \boldsymbol{x}^k$;
3: $\quad$ **for** $i = 1, \ldots, N$ **do**
4: $\quad\quad$ **repeat**
5: $\quad\quad\quad$ set $\overline{L}_i^{k+1} = \nu^p \overline{L}_i^k$, $\quad \gamma_i^{k+1} = \gamma_i^k / \nu^p$, $\quad p = p + 1$;
6: $\quad\quad\quad$ compute $x_i^{k,i} \in \mathbf{T}_{h/\gamma_i^{k+1}}(\boldsymbol{x}^{k,i-1})$, $\quad \boldsymbol{x}^{k,i} = \boldsymbol{x}^{k,i-1} + U_i(x_i^{k,i} - x_i^{k,i-1})$;
7: $\quad\quad$ **until** $f(\boldsymbol{x}^{k,i}) \leq f(\boldsymbol{x}^{k,i-1}) + \langle \nabla_i f(\boldsymbol{x}^{k,i-1}), \boldsymbol{x}_i^{k,i} - \boldsymbol{x}_i^{k,i-1} \rangle + \overline{L}_i^{k+1} \mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1})$
8: $\quad\quad p_i^k = p - 1$; $p = 0$;
9: $\quad \boldsymbol{x}^{k+1} = \boldsymbol{x}^{k,N}$, $k = k + 1$;

ENSURE A vector $\boldsymbol{x}^k$.

Let us assume that the solution of the subproblem $x_i^{k,i} \in \mathbf{T}^i_{h/\gamma_i^{k+1}}(\boldsymbol{x}^{k,i-1})$ can be computed exactly. Beside of the computational cost of solving the subproblem, we note that in each iteration of A-BPALM, we only needs two calls of the first-order oracle (one call for f and the other for h in Line 7). We next provide an upper bound on the *total number of calls of oracle* after k iterations of A-BPALM and those needed to satisfy (4.5).

Proposition 4.8 (worst-case oracle calls) *Let* $(\boldsymbol{x}^k)_{k \in \mathbb{N}}$ *be generated by* A-BPALM. *Then*,

(i) *after at most* $\mathbf{max}\left\{ \frac{1}{\ln \nu} \ln \frac{\nu L_i}{\overline{L}_i^0}, 0 \right\}$ *iterations the line search* (*Lines* 4 *to* 7 *of* A-BPALM) *will be satisfied*;
(ii) *the number of oracle calls* $\mathcal{N}_k$ *after k full cycle is bounded by*

$$\mathcal{N}_k \le \left(2N + 2\sum_{i=1}^{N} \max\left\{\tfrac{1}{\ln\nu}\ln\tfrac{\nu L_i}{\overline{L}_i^0}, 0\right\}\right)k;$$

(iii) *the worst-case number of oracle calls to satisfy* (4.5) *is given by*

$$\left(2N + 2\sum_{i=1}^{N} \max\left\{\tfrac{1}{\ln\nu}\ln\tfrac{\nu L_i}{\overline{L}_i^0}, 0\right\}\right)\left(1 + \tfrac{(\varphi(\boldsymbol{x}^0)-\inf\varphi)}{\overline{\rho}\epsilon}\right)$$

$$\text{with } \overline{\rho} := \min\left\{\tfrac{(1-\gamma_1^0\overline{L}_1^0)}{\gamma_1^0}, \ldots, \tfrac{(1-\gamma_N^0\overline{L}_N^0)}{\gamma_N^0}\right\}.$$

Proof According to step 5 and step 8 of A-BPALM, if $\ln\overline{L}_i^0 \ge \ln(\nu L_i)$ then from step 5 we have $\bar{L}_i^{k+1} \ge \bar{L}_i^k \ge \bar{L}_i^0 > L_i$ for all $k \ge 0$, hence $p_i^k = 0$ since the condition in step 7 is satisfied; otherwise, we have $\overline{L}_i^{k+1} = \nu^{p_i^k}\overline{L}_i^k$, i.e.,

$$p_i^k = \tfrac{1}{\ln\nu}\left(\ln\overline{L}_i^{k+1} - \ln\overline{L}_i^k\right) \le \tfrac{1}{\ln\nu}\left(\ln(\nu L_i) - \ln\overline{L}_i^0\right) \quad i = 1, \ldots, N,$$

giving Proposition 4.8(i). Hence, the total number of calls of oracle after k iterations is given by

$$\mathcal{N}_k = \sum_{j=0}^{k-1}\sum_{i=1}^{N} 2(p_i^j + 1) \le 2\sum_{i=1}^{N}\left[k + \sum_{j=0}^{k-1} \max\left\{\tfrac{1}{\ln\nu}\ln\tfrac{\nu L_i}{\overline{L}_i^0}, 0\right\}\right],$$

giving Proposition 4.8(ii).

The step 5 implies that the sequences $\left((1-\gamma_i^k\overline{L}_i^k)/\gamma_i^k\right)_{k\in\mathbb{N}}$ $(i = 1, \ldots, N)$ are increasing with respect to k, i.e., $(1-\gamma_i^{k+1}\overline{L}_i^{k+1})/\gamma_i^{k+1} \ge (1-\gamma_i^0\overline{L}_i^0)/\gamma_i^0$, $i = 1, \ldots, N$. Now, following the proof of Proposition 4.1, it is easy to see that

$$\begin{aligned}\overline{\rho}\textstyle\sum_{i=1}^{N}\mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1}) &= \min\{(1-\gamma_1^0\overline{L}_1^0)/\gamma_1^0, \ldots, (1-\gamma_N^0\overline{L}_N^0)/\gamma_N^0\}\textstyle\sum_{i=1}^{N}\mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1}) \\ &\le \min\{\tfrac{1-\gamma_1^{k+1}\overline{L}_1^{k+1}}{\gamma_1^{k+1}}, \ldots, \tfrac{1-\gamma_N^{k+1}\overline{L}_N^{k+1}}{\gamma_N^{k+1}}\}\textstyle\sum_{i=1}^{N}\mathbf{D}_h(\boldsymbol{x}^{k,i}, \boldsymbol{x}^{k,i-1}) \\ &\le \varphi(\boldsymbol{x}^k) - \varphi(\boldsymbol{x}^{k+1}).\end{aligned} \tag{4.17}$$

From the proof of Corollary 4.2, BPALM will be terminated within $k \le 1 + \frac{\varphi(x^0)-\inf\varphi}{\overline{\rho}\epsilon}$ iterations. Together with Proposition 4.8(ii), this implies that Proposition 4.8(iii) is true.

In light of (4.17), Proposition 4.1 holds true by replacing ρ with $\overline{\rho}$. Considering this replacement, all the results of Proposition 4.3, Theorem 4.4, Theorem 4.5, and Theorem 4.7 remain valid for A-BPALM.

Remark 4.9 **(A-BPALM variant)** We here notice that one may change Line 5 of A-BPALM as "set $\overline{L}_i^{k+1} = \nu^p\overline{L}_i^0, \gamma_i^{k+1} = \frac{\gamma_i^0}{\nu^p}, p = p + 1$;", which always start the procedure from $\overline{L}_i^0$ and γ_i^0. It is easy to see that the results of Proposition 4.8 are still valid for this variant of A-BPALM.

5 Application to orthogonal nonnegative matrix factorization

A natural way of analyzing large data sets is finding an effective way to represent them using dimensionality reduction methodologies. *Nonnegative matrix factorization* (NMF) is one such technique that has received much attention in the last few years; see, e.g., [29, 33, 34] and the references therein. In order to extract hidden and important features from data, NMF decomposes the data matrix into two factor matrices (usually much smaller than the original data matrix) by imposing componentwise nonnegativity and (possibly) sparsity constraints on these factor matrices. More precisely, let the data matrix be $X = (x_1, x_2, \dots, x_q) \in \mathbb{R}_+^{p\times q}$ where each x_i represents some data point. NMF seeks a decomposition of X into a nonnegative $p \times r$ basis matrix $U = (u_1, u_2, \dots, u_r) \in \mathbb{R}_+^{p\times r}$ and a nonnegative $r \times q$ coefficient matrix $V = (v_1, v_2, \dots, v_r)^T \in \mathbb{R}_+^{r\times q}$ such that

$$X \approx UV, \tag{5.1}$$

where $\mathbb{R}_+^{p\times q}$ is the set of $p \times q$ element-wise nonnegative matrices. Extensive research has been carried out on variants of NMF, and most studies in this area have focused on algorithmic developments, but with very limited convergence theory. This motivates us to study the application of BPALM and A-BPALM to a variant of NMF, namely orthogonal NMF (ONMF).

5.1 Orthogonal nonnegative matrix factorization

Besides the decomposition (5.1), the *orthogonal nonnegative matrix factorization* (ONMF) involves an additional orthogonality constraint $VV^T = I_r$ leading to the constrained optimization problem

$$\begin{array}{ll} \textbf{minimize} & \frac{1}{2}\|X - UV\|_F^2 \\ \textbf{subject to} & U \geq 0,\ V \geq 0,\ VV^T = I_r \end{array} \tag{5.2}$$

where $I_r \in \mathbb{R}^{r\times r}$ is the identity matrix. By imposing the matrix V to be orthogonal (as well as nonnegative), ONMF imposes that each data point is only associated with one basis vector, hence ONMF is closely related to clustering problems; see [52] and the references therein. Since the projection onto the set $\mathcal{C} := \{(U, V) \in \mathbb{R}^{p\times r} \times \mathbb{R}^{r\times q} \mid U \geq 0,\ V \geq 0,\ VV^T = I_r\}$ is costly, we here consider the penalized formulation

$$\begin{array}{ll} \textbf{minimize} & \frac{1}{2}\|X - UV\|_F^2 + \frac{\lambda}{2}\|I_r - VV^T\|_F^2 \\ \textbf{subject to} & U \geq 0,\ V \geq 0, \end{array} \tag{5.3}$$

for the penalty parameter $\lambda > 0$. By fixing U, the objective function of the problem (5.3) is not Lipschitz smooth with respect to V, and hence standard NMF block proximal gradient descent algorithms cannot be applied to solve (5.3). The unconstrained version of (5.3) is given by

$$\underset{(U,V)}{\textbf{minimize}} \ \ \tfrac{1}{2}\|X - UV\|_F^2 + \tfrac{\lambda}{2}\|I_r - VV^T\|_F^2 + \delta_{U\geq 0} + \delta_{V\geq 0}, \tag{5.4}$$

where $\delta_{U\geq 0}$ and $\delta_{V\geq 0}$ are the indicator functions of the sets $\mathcal{C}_1 := \{U \in \mathbb{R}^{p\times r} \mid U \geq 0\}$ and $\mathcal{C}_2 := \{V \in \mathbb{R}^{r\times q} \mid V \geq 0\}$, respectively. This problem can be put in the form of (1.1) using

$$f(U,V) := \tfrac{1}{2}\|X - UV\|_F^2 + \tfrac{\lambda}{2}\|I_r - VV^T\|_F^2, \quad g_1(U) := \delta_{U\geq 0}, \quad g_2(V) := \delta_{V\geq 0}, \tag{5.5}$$

in which both $g_1(U)$ and $g_2(V)$ are nonsmooth convex functions.

We now apply BPALM and A-BPALM to solve (5.4). Specifically, in Proposition 5.1, we provide a full-domain multi-block strongly convex kernel function h such that $f(U, V)$ is (L_1, L_2)-smooth relative to h, and then in Theorem 5.2 we give closed-form solutions of the subproblem (4.1). Let us emphasize that Assumption 1, 2 and 3 hold for the problem (5.4) and the kernel (5.6). Now, applying Theorem 4.5, we have that BPALM and A-BPALM *converges globally* to a stationary point of the objective function of (5.4). For the ONMF problem (5.4), since the function $f(U,V) = \frac{1}{2}\|X - UV\|_F^2 + \frac{\lambda}{2}\|I_r - VV^T\|_F^2$ is a polynomial and the functions $g_1(U) = \delta_{U\geq 0}$ and $g_2(V) = \delta_{V\geq 0}$ are indicator functions of semialgebraic sets, they are semialgebraic functions. Therefore, it follows from [20, Theorem 3.1] that $\varphi(U,V) = f(U,V) + g_1(U) + g_2(V)$ is a KL function with exponent $\theta \in [0, 1)$. We however could not provide a KL exponent (see (4.16)) for this problem, and so we are not able to derive the rate of convergence of our methods for this problem, which remains an open question.

Proposition 5.1 (multi-block relative smoothness of ONMF objective) *Let $\alpha, \beta, \varepsilon_1, \varepsilon_2 > 0$ and let the function $h : \mathbb{R}^{p\times r} \times \mathbb{R}^{r\times q} \to \overline{\mathbb{R}}$ be a kernel given by*

$$h(U,V) := \frac{\alpha}{2}\|U\|_F^2\|V\|_F^2 + \frac{\beta}{4}\|V\|_F^4 + \frac{\varepsilon_1}{2}\|U\|_F^2 + \frac{\varepsilon_2}{2}\|V\|_F^2. \tag{5.6}$$

Then the function f given in (5.5) is (L_1, L_2)-smooth relative to h with

$$L_1 \geq \tfrac{1}{\alpha}, \quad L_2 \geq \mathbf{max}\{\tfrac{6\lambda}{\beta}, \tfrac{1}{\alpha}\}. \tag{5.7}$$

Proof Using partial derivatives $\nabla_U f(U,V) = UVV^T - XV^T$, $\nabla^2_{UU} f(U,V)Z = ZVV^T$, and the Cauchy-Schwarz inequality, it can be concluded that $\langle Z, \nabla^2_{UU} f(U,V)Z\rangle \leq \|V\|_F^2\|Z\|_F^2$. On the other hand, $\nabla_U h(U,V) = \varepsilon_1 U + \alpha\|V\|_F^2 U$ and

$$\langle Z, \nabla^2_{UU} h(U,V)Z\rangle = (\varepsilon_1 + \alpha\|V\|_F^2)\langle Z, Z\rangle \geq \alpha\|V\|_F^2\|Z\|_F^2.$$

Together with (5.7), this yields $L_1\nabla^2_{UU} h(U,V) - \nabla^2_{UU} f(U,V) \succeq 0$.

From $\nabla_V f(U,\cdot)(V) = U^TUV - U^TX + 2\lambda(VV^TV - V)$ and the definition of directional derivative, we obtain

$$\begin{aligned}\nabla^2_{VV} f(U,V)Z &= \mathbf{lim}_{t\to 0}\frac{U^TU(V+tZ)-U^TX+2\lambda[(V+tZ)(V+tZ)^T(V+tZ)-(V+tZ)]}{t}\\ &\quad -\frac{U^TUV-U^TX+2\lambda(VV^TV-V)}{t}\\ &= U^TUZ+2\lambda(ZV^TV+VZ^TV+VV^TZ-Z)\quad \forall Z\in\mathbb{R}^{r\times q}.\end{aligned}$$

From this last equation, $\langle Y_1, Y_2\rangle := \mathbf{tr}(Y_1^TY_2)$, basic properties of the trace, the Cauchy-Schwarz inequality, and the submultiplicative property of the Frobenius norm, we obtain

$$\begin{aligned}\langle Z, \nabla^2_{VV} f(U,V)Z\rangle &= \langle Z, U^TUZ+2\lambda(ZV^TV+VZ^TV+VV^TZ-Z)\rangle\\ &\le 6\lambda\|Z\|_F^2\|V\|_F^2+\|U\|_F^2\|Z\|_F^2.\end{aligned}$$

We have $\nabla_V h(U,V) = (\varepsilon_2+\alpha\|U\|_F^2)V+\beta\|V\|^2V$. Hence,

$$\nabla^2_{VV} h(U,V)Z = \beta\big(\|V\|_F^2Z+2\langle V,Z\rangle V\big)+(\varepsilon_2+\alpha\|U\|_F^2)Z,$$

which implies

$$\begin{aligned}\langle Z, \nabla^2_{VV} h(U,V)Z\rangle &= \beta\|Z\|_F^2\|V\|_F^2+2\beta\langle V,Z\rangle^2+(\varepsilon_2+\alpha\|U\|_F^2)\|Z\|_F^2\\ &\ge \beta\|Z\|_F^2\|V\|_F^2+\alpha\|U\|_F^2\|Z\|_F^2.\end{aligned}$$

Hence, it follows from (5.7) that $L_2\nabla^2_{VV}h(U,V)-\nabla^2_{VV}f(U,V)\succeq 0$.

For given U^k and V^k, applying BPALM and A-BPALM to (5.4), U^{k+1} and V^{k+1} should be computed efficiently, which we study next.

Theorem 5.2 (closed-form solutions of the subproblem (4.1) for ONMF) *Let h be the kernel functions given in (5.6). For given U^k and V^k, the problem (5.4), and the subproblem (4.1), the following assertions hold*:

(i) *Let $\eta_1 = \alpha\|V^k\|_F^2+\varepsilon_1$. The iteration U^{k+1} is given by*

$$U^{k+1} = \mathbf{max}\{U^k-\tfrac{\gamma_1}{\eta_1}\big(U^kV^k(V^k)^T-X(V^k)^T\big),0\};\tag{5.8}$$

(ii) *Let $\eta_2 = \varepsilon_2+\alpha\|U^{k+1}\|_F^2$. The iteration V^{k+1} is given by*

$$V^{k+1} = \tfrac{1}{t_k}\mathbf{max}\{G^k,0\}\tag{5.9}$$

with $\nabla_V f(U^{k+1},V^k) = (U^{k+1})^TU^{k+1}V^k-(U^{k+1})^TX+2\lambda(V^k(V^k)^TV^k-V^k)$ and

$$\begin{aligned}G^k &= \eta_2V_k+\beta\|V_k\|_F^2V_k-\gamma_2\nabla_V f(U^{k+1},V^k),\\ t_k &= \tfrac{\eta_2}{3}+\sqrt[3]{\tfrac{c+\sqrt{\Delta}}{2}+\tfrac{\eta_2^3}{27}}+\sqrt[3]{\tfrac{c-\sqrt{\Delta}}{2}+\tfrac{\eta_2^3}{27}},\end{aligned}\tag{5.10}$$

where $c = \beta\|\mathbf{max}(0, G^k)\|_F^2$ and $\Delta = c^2 + \frac{4}{27}c\eta_2^3$.

Proof It follows from (3.11) that

$$\begin{aligned} U^{k+1} &= \mathop{\mathbf{arg\,min}}_{U\in\mathbb{R}^{p\times r}}\{\langle\nabla_U f(U^k, V^k), U - U^k\rangle + \tfrac{1}{\gamma_1}\mathbf{D}_h((U, V^k), (U^k, V^k)) + g_1(U)\} \\ &= \mathop{\mathbf{arg\,min}}_{U\geq 0}\{\langle\gamma_1\nabla_U f(U^k, V^k) - \nabla_U h(U^k, V^k), U\rangle + \tfrac{\eta_1}{2}\|U\|_F^2\} \\ &= \mathop{\mathbf{arg\,min}}_{U\geq 0}\|U - (\tfrac{1}{\eta_1}\nabla_U h(U^k, V^k) - \tfrac{\gamma_1}{\eta_1}\nabla_U f(U^k, V^k))\|_F^2 \\ &= \mathbf{Proj}_{U\geq 0}(U^k - \tfrac{\gamma_1}{\eta_1}\nabla_U f(U^k, V^k)), \end{aligned}$$

with $\nabla_U f(U^k, V^k) = U^k V^k (V^k)^T - X(V^k)^T$, giving (5.8).

We have

$$\begin{aligned} V^{k+1} &= \mathop{\mathbf{arg\,min}}_{V\in\mathbb{R}^{r\times q}}\{\langle\nabla_V f(U^{k+1}, V^k), V - V^k\rangle + \tfrac{1}{\gamma_2}\mathbf{D}_h((U^{k+1}, V), (U^{k+1}, V^k)) + g_2(V)\} \\ &= \mathop{\mathbf{arg\,min}}_{V\geq 0}\{h(U^{k+1}, V) - \langle\nabla_V h(U^{k+1}V^k) - \gamma_2\nabla_V f(U^{k+1}, V^k), V\rangle\}. \end{aligned} \tag{5.11}$$

Let us consider the normal cone $\mathcal{N}_{V\geq 0}(V^{k+1}) = \{P \in \mathbb{R}^{r\times q} \mid V^{k+1} \odot P = 0,\ P \leq 0\}$ (see [58, Corollary 3.5]), where $V \odot P$ denotes the *Hadamard product* given pointwise by $(V \odot P)_{ij} := V_{ij}P_{ij}$ for $i \in 1, \ldots, r$ and $j \in 1, \ldots, q$. The first-order optimality conditions for (5.11) lead to $G^k - (\beta\|V^{k+1}\|_F^2 + \eta_2)V^{k+1} \in \mathcal{N}_{V\geq 0}(V^{k+1})$ with $G^k := \nabla_V h(U^{k+1}V^k) - \gamma_2\nabla_V f(U^{k+1}, V^k)$.

Now we consider two cases: (i) $G_{ij}^k \leq 0$; (ii) $G_{ij}^k > 0$. In Case (i), $P_{ij} = G_{ij}^k - (\beta\|V^{k+1}\|_F^2 + \eta_2)V_{ij}^{k+1} \leq 0$, hence, $V_{ij}^{k+1} = 0$ (otherwise, $V_{ij}^{k+1} < 0$ which implies $P_{ij} < 0$; this contradicts to the condition $V_{ij}^{k+1}P_{ij} = 0$). In Case (ii), if $V_{ij}^{k+1} = 0$, then $P_{ij} = G_{ij}^k > 0$, which contradicts $P \leq 0$, hence, $G_{ij}^k - (\beta\|V^{k+1}\|_F^2 + \eta_2)V_{ij}^{k+1} = 0$. Combining both cases, we get

$$(\beta\|V^{k+1}\|_F^2 + \eta_2)V^{k+1} = \mathbf{Proj}_{V\geq 0}(G^k).$$

Denote $t_k = (\beta\|V^{k+1}\|_F^2 + \eta_2)$. Then we have $\|V^{k+1}\|_F^2 = (t_k - \eta_2)/\beta$. Therefore, t_k satisfies

$$t_k^3 - \eta_2 t_k^2 - \beta\|\mathbf{Proj}_{V\geq 0}(G^k)\|_F^2 = 0.$$

Note that the third-order polynomial equation $y^2(y - a) = c$ has the unique real solution $y = \frac{a}{3} + \sqrt[3]{\frac{c+\sqrt{\Delta}}{2} + \frac{a^3}{27}} + \sqrt[3]{\frac{c-\sqrt{\Delta}}{2} + \frac{a^3}{27}}$, where $\Delta = c^2 + \frac{4}{27}ca^3$, which gives (5.10).

5.2 Preliminary numerical experiments

In this section, we report preliminary numerical results of our experiments on BPALM and two variants of A-BPALM, namely,

- A-BPALM1: the algorithm A-BPALM;
- A-BPALM2: the variant of A-BPALM as described in Remark 4.9.

Since the unconstrained ONMF problem (5.4) involves the quadratic penalty term $\frac{\lambda}{2}\|I_r - VV^T\|_F^2$, we also consider a "continuation" variant of these algorithm that starts from some $\lambda > 0$, run one of the above-mentioned algorithms until some stopping criterion holds and save its best point, and then it increases the penalty parameter and run the algorithm with the starting point as the best point of the last call, and it continues the procedure until we stop the algorithm. We refer to this heuristic procedure as *continuation*; see Algorithm 3.

Algorithm 3 Continuation procedure

REQUIRE $x^0 \in \mathbf{int\,dom}\, h,\ \lambda > 0,\ c > 1$.
1: **repeat**
2: starting from x^0; run one of BPALM, A-BPALM1, or A-BPALM2 to attain an inexact solution $\overline{x}$ of (5.4);
3: set $x^0 \leftarrow \overline{x},\ \lambda \leftarrow c\lambda$;
4: **until** some stopping criterion holds
ENSURE $\overline{x}$

In our implementation, all the codes were written in MATLAB[1] and runs were performed on a laptop with 1.8 GHz Intel Core i7 CPU and 16 GB RAM. On the basis of our preliminary experiments, we here set $\alpha = \beta = 1$ and $\epsilon_1 = \epsilon_2 = 10^{-9}$ to provide the relative smoothness constants as described in (5.7), and the related stepsizes are computed by $\gamma_i = \frac{1}{L_i} - \epsilon$, when ϵ is set as the machine precision. For A-BPALM1 and A-BPALM2, we set $\nu = 2$, and we also set $\overline{L}_i^0 = 10^{-4}L_i$. For the continuation versions of our algorithms, we set $c = 2$ and consider the following notations:

- BPALM-c: BPALM with the continuation;
- A-BPALM1-c: A-BPALM1 with the continuation;
- A-BPALM2-c: A-BPALM2 with the continuation.

Moreover, we implement the following baseline algorithms in our experiments:

- MU: the Multiplicative Update method proposed in [56],
- HALS: the Hierarchical Alternating Least Squares Algorithm proposed in [39];

[1] The codes are publicly available at https://github.com/MasoudAhoo/BPALM

For all of these algorithms, we use the same initialization, namely the successive projection algorithm (SPA) [4]. SPA is guaranteed to recover correctly the factors U and V, given that $X = UV$ where $U \geq 0$, $V \geq 0$ and $VV^T = I_r$ (this is the noiseless case); in fact, ONMF is a special case of separable NMF in the absence of noise. Moreover, this recovery property holds true even in the presence of small bounded noise [35].

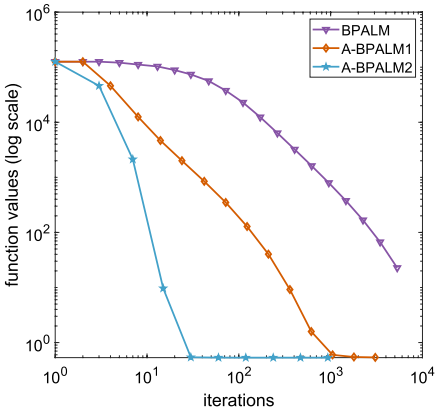
In the first experiment, we illustrate the evolution of the objective function (5.3) and the orthogonal error $O_{error} := \|I - V^k(V^k)^T\|_F$ obtained by our proposed methods BPALM, A-BPALM1 and A-BPALM2. We generate two synthetic data sets with $(p, q, r) = (500, 500, 10)$ and $(p, q, r) = (500, 2000, 10)$ as follows. We use the MATLAB command rand to generate random nonnegative matrices $U \in \mathbb{R}_+^{p\times r}$ and $R \in \mathbb{R}_+^{p\times q}$, then we generate a random orthogonal nonnegative matrix $V \in \mathbb{R}_+^{r\times q}$. Next, we set $X = UV$ to obtain the p-by-q orthogonal decomposable matrix X, and finally add 5% of noise by $X = X + 0.05\frac{\|X\|_F}{\|R\|_F}R$. We run our algorithms with the fixed penalty parameter $\lambda = 100$ and stopped the algorithms after 15 s. The results are illustrated in (5.3). We make two observations: (i) A-BPALM1 and A-BPALM2 outperform BPALM, A-BPALM2 being the best among them; (ii) within the same running time, BPALM can make more iterations than its line search variants A-BPALM1 and A-BPALM2, and A-BPALM2 runs the least number of iterations compared to the others (Fig. 1).

In the second experiment, we compare our algorithms (with and without continuation) against MU and HALS. We generate 50 data sets with $r = 10$, and p and q being uniformly chosen within the range [200,1000]. For each (p, q, r), we generate a data set in the same way as in the first experiment. For each data set, we use the same SPA initialization for all algorithms and run each algorithm for 20 s. For the continuation versions BPALM-c, A-BPALM1-c and A-BPALM2-c, we start with $\lambda = 10$, stop the inner algorithms every 2 s to increase λ by factor $c = 2$; and for BPALM, A-BPALM1 and A-BPALM2, we use a fixed $\lambda = 5120$, which is the last value of λ used for continuation versions.

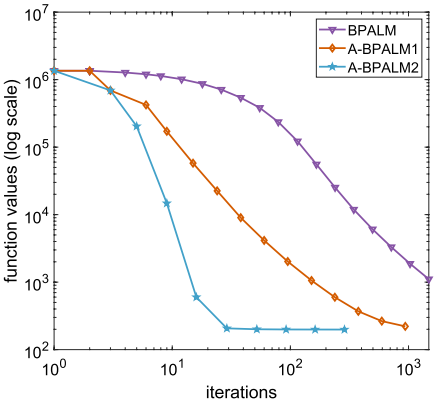
We report the mean and standard derivation of the final orthogonal error and the fitting error $F_{error} := \frac{\|X-U^kV^k\|_F}{\|X\|_F}$ obtained by each algorithm over 50 runs in Table 1.

We observe that, on average, the continuation versions outperform BPALM variants with fixed λ. Compared to HALS and MU, our algorithms A-BPALM2 and the continuation versions outperform HALS and MU in term of the orthogonal error and are competitive with HALS and MU in term of the fitting error. A-BPALM2 provides the best orthogonal error among the algorithms.

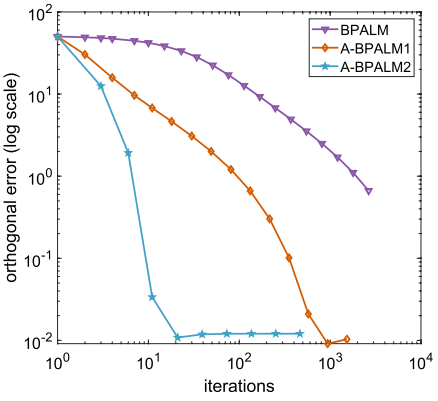
Finally, we report the performance of our algorithms on the Hubble telescope data set which is taken from [50]. In this problem, each row of the matrix X is a vectorized image of the Hubble telescope at a given wavelength for a total of $p = 100$ wavelengths. Each image contains $q = 128 \times 128$ pixels. Since each pixel in the image contains mostly a single material, it makes sense to use ONMF to cluster the pixel according to the material they contain (see Fig. 2 for an illustration). Since the continuation versions of our algorithms perform better, we here only apply the continuation versions of BPALM, A-BPALM1, and A-BPALM2. We use the SVD-based initialization as in



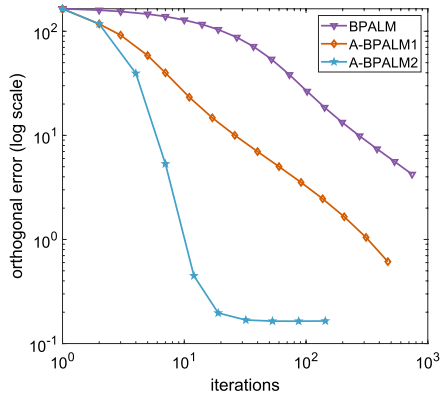
(a) *Function values evolutions (first dataset)*



(b) *Function values evolutions (second dataset)*



(c) *Orthogonal error evolutions (first dataset)*



(d) *Orthogonal error evolutions (second dataset)*

Fig. 1 A comparison among BPALM, A-BPALM1, and A-BPALM2: Subfigures **a** and **b** stand for function values vs. iterations for the 2 synthetic data with $(p, q, r) = (500, 500, 10)$ (first dataset) and $(p, q, r) = (500, 2000, 10)$ (second dataset), respectively; Subfigures **c** and **d** illustrate the orthogonal error vs. iterations for these data sets, respectively

[52]. We run each algorithm for 100 s. The final outputs of the algorithms, along with the ground true Hubble image, are illustrated in Fig. 2.

We observe on Fig. 2 that A-BPALM1-c and A-BPALM2-c provide slightly better quality images compared to BPALM-c (look for example at the fifth basis image), while HALS and MU fail to cluster the pixels properly as their solutions have a too large orthogonal error O_{error}, more than 10 times larger than A-BPALM1-c and A-BPALM2-c.

Table 1 Mean ± std of the final F_{error} and O_{error} obtained by each algorithm over 50 runs. In each column, the best results are displayed in bold

Algorithm	F_{error}	O_{error}
BPALM	$2.679\,10^{-2} \pm 1.992\,10^{-4}$	$1.269\,10^{0} \pm 9.191\,10^{-1}$
A-BPALM1	$2.654\,10^{-2} \pm 2.370\,10^{-4}$	$1.332\,10^{-2} \pm 2.614\,10^{-2}$
A-BPALM2	$2.573\,10^{-2} \pm 1.509\,10^{-4}$	$\mathbf{2.133\,10^{-3} \pm 4.067\,10^{-4}}$
BPALM-c	$2.582\,10^{-2} \pm 2.079\,10^{-4}$	$2.395\,10^{-3} \pm 3.787\,10^{-4}$
A-BPALM1-c	$2.568\,10^{-2} \pm 1.356\,10^{-4}$	$2.419\,10^{-3} \pm 4.641\,10^{-4}$
A-BPALM2-c	$2.568\,10^{-2} \pm 1.371\,10^{-4}$	$2.430\,10^{-3} \pm 4.921\,10^{-4}$
MU	$2.503\,10^{-2} \pm 1.540\,10^{-4}$	$9.566\,10^{-3} \pm 5.528\,10^{-4}$
HALS	$\mathbf{2.414\,10^{-2} \pm 1.433\,10^{-4}}$	$1.223\,10^{-2} \pm 5.324\,10^{-4}$

6 Final remarks

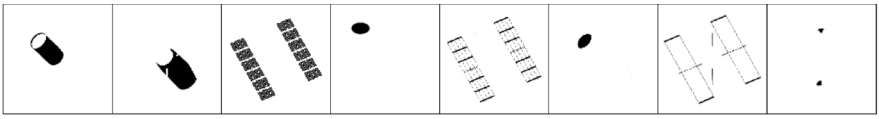
We have analyzed two new alternating linearized minimization algorithms called BPALM and A-BPALM for solving the popular nonconvex nonsmooth optimization problem (1.1). To do so, we first introduced the notion of multi-block relative smoothness and verified the fundamental properties of the Bregman proximal alternating linearized mapping. Convergence analysis including the subsequential convergence, global convergence and convergence rate of the proposed algorithms is studied under the framework of multi-block relative smoothness and multi-block kernel functions. We employ BPALM and A-BPALM to solve the orthogonal nonnegative matrix factorization (ONMF) problem. We emphasize that, to the best of our knowledge, BPALM and A-BPALM are the first algorithms with rigorous convergence guarantee for solving ONMF in the literature. Some preliminary numerical tests are provided to illustrate the performance of our algorithms.

Applying (A-)BPALM on other problems and comparing them with state-of-the-art algorithms is a topic for future work.

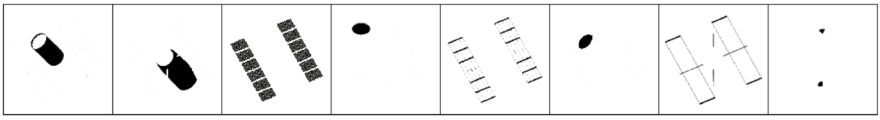
Appendix

Lemma 8.1 *Let all assumptions of Theorem 4.4 be valid. Then, the following assertions hold*:

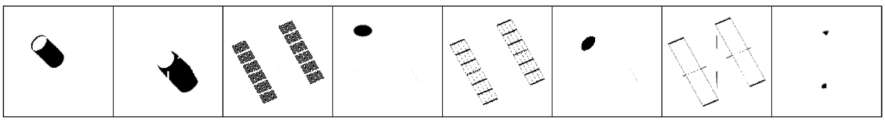
(i) $\mathbf{lim}_{k\to\infty}\mathbf{dist}\left(\boldsymbol{x}^k, \omega(\boldsymbol{x}^0)\right) = 0$;
(ii) $\omega(\boldsymbol{x}^0)$ *is a nonempty, compact, and connected set*;
(iii) *the objective function* φ *is finite and constant on* $\omega(\boldsymbol{x}^0)$.



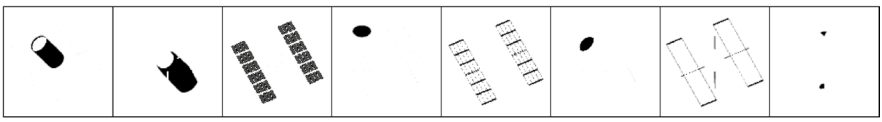
(a) *Ground truth factor V*



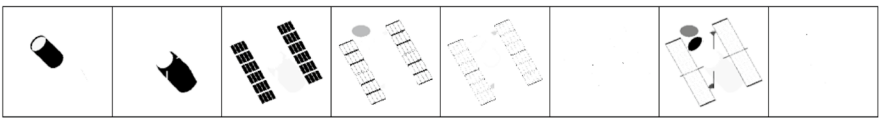
(b) *BPALM-c*, $F_{error} = 1.005 \times 10^{-1}$, $O_{error} = 9.992 \times 10^{-3}$



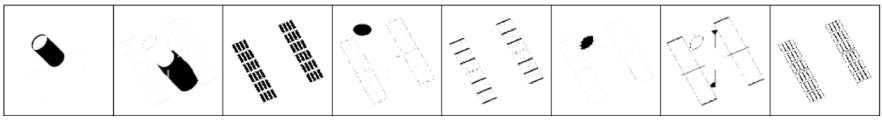
(c) *A-BPALM1-c*, $F_{error} = 6.396 \times 10^{-2}$, $O_{error} = 4.988 \times 10^{-3}$



(d) *A-BPALM2-c*, $F_{error} = 6.363 \times 10^{-2}$, $O_{error} = 7.348 \times 10^{-3}$



(e) *MU*, $F_{error} = 4.016 \times 10^{-2}$, $O_{error} = 4.4113 \times 10^{-1}$



(f) *HALS*, $F_{error} = 2.380 \times 10^{-2}$, $O_{error} = 1.274 \times 10^{-1}$

Fig. 2 Comparison of BPALM-c, A-BPALM1-c, A-BPALM2-c, MU and HALS on the Hubble image. Algorithms are run for 100 s. In each subfigure, each image corresponds to a row of V that has been reshaped as an image (since each entry corresponds to a pixel; see the description in the text)

Proof Lemma 8.1(i) is a direct consequence of Theorem 4.4, and Lemma 8.1(ii) and Lemma 8.1(iii) can be proved in the same way as [23, Lemma 5(iii)-(iv)].

Lemma 8.2 *Let all assumptions of Theorem 4.5 is satisfied. If* $\varphi(\boldsymbol{x}^k) > \varphi^\star$, *there exists* $\varepsilon, \eta > 0$ *and the desingularizing function* ψ *such that*

$$\psi'(\varphi(\boldsymbol{x}^k) - \varphi^\star)\mathbf{dist}(0, \partial\varphi(\boldsymbol{x}^k)) \geq 1 \quad \text{for } k \geq k_0. \tag{8.1}$$

Proof From Lemma 8.1(ii), the set of limit points $\omega(\boldsymbol{x}^0)$ of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is nonempty and compact and φ is finite and constant on $\omega(\boldsymbol{x}^0)$ due to Lemma 8.1(iii). Moreover,

$\varphi(\boldsymbol{x}^k) > \varphi^\star$ and the sequence $(\varphi(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ is decreasing (Proposition 4.1(i)), i.e., there exist $\eta > 0$ and $k_1 \in \mathbb{N}$ such that $\varphi^\star < \varphi(\boldsymbol{x}^k) < \varphi^\star + \eta$ for all $k \geq k_1$. For $\varepsilon > 0$, Proposition 4.1(i) implies that there exists $k_2 \in \mathbb{N}$ such that $\mathbf{dist}(\boldsymbol{x}^k, \omega(\boldsymbol{x}^0)) < \varepsilon$ for $k \geq k_2$. Setting $k_0 := \mathbf{max}\{k_1, k_2\}$ and according to Fact 2.2, there exist $\varepsilon, \eta > 0$ and a desingularization function ψ such that for any element in

$$\{\boldsymbol{x}^k \ \mid \ \mathbf{dist}(\boldsymbol{x}^k, \omega(\boldsymbol{x}^0)) < \varepsilon\} \cap [\varphi^\star < \varphi(\boldsymbol{x}^k) < \varphi^\star + \eta] \quad \text{for } k \geq k_0,$$

the inequality (8.1) is valid.

We next present the proof of Theorem 4.7.

Proof of Theorem 4.7. The proof has two key parts.

In the first part, we show that there exist $c > 0$ and $\bar{k} \in \mathbb{N}$ such that for all $k \geq \bar{k}$ the following inequalities hold for $i = 1, \ldots, N$:

$$\|x_i^k - x_i^\star\| \leq \begin{cases} c\mathbf{max}\{1, \frac{\kappa}{1-\theta}\}\sqrt{\mathcal{S}_{k-1}} & \text{if } \theta \in (0, 1/2], \\ c\frac{\kappa}{1-\theta}\mathcal{S}_{k-1}^{1-\theta} & \text{if } \theta \in (1/2, 1). \end{cases} \tag{8.2}$$

Let $\varepsilon > 0$ be as described in (4.16) and $x^k \in \mathbf{B}(x^\star;\varepsilon)$ for all $k \geq \tilde{k}$ and $\tilde{k} \in \mathbb{N}$. By the definitions of a_k and b_k in (4.15) and using (4.14), we get $a_{k+1} \leq \frac{1}{2}a_k + b_k$ for all $k \geq \tilde{k}$. Since $(\varphi(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ is nonincreasing,

$$\sum_{j=k}^{\infty} a_{j+1} \leq \frac{1}{2}\sum_{j=k}^{\infty}(a_j - a_{j+1} + a_{j+1}) + \frac{\hat{c}N}{2}\sum_{j=k}^{\infty}\left(\Delta_j - \Delta_{j+1}\right) = \frac{1}{2}\sum_{j=k}^{\infty} a_{j+1} + \frac{1}{2}a_k + \frac{\hat{c}N}{2}\Delta_k.$$

Together with the arithmetic and quadratic mean inequalities, $\psi(\mathcal{S}_k) \leq \psi(\mathcal{S}_{k-1})$, and Proposition 4.1(i), this lead to

$$\begin{aligned} \textstyle\sum_{j=k}^{\infty} a_{j+1} \ &\leq a_k + \hat{c}N\Delta_k = \textstyle\sum_{i=1}^{N}\|x_i^k - x_i^{k-1}\| + \hat{c}N\psi(\mathcal{S}_k) \leq \sqrt{N}\sqrt{\textstyle\sum_{i=1}^{N}\|x_i^k - x_i^{k-1}\|^2} + \hat{c}N\psi(\mathcal{S}_k) \\ &\leq \sqrt{2N}\mathbf{max}\{\tfrac{1}{\sqrt{\sigma_1}}, \ldots, \tfrac{1}{\sqrt{\sigma_N}}\}\sqrt{\textstyle\sum_{i=1}^{N}\mathbf{D}_h(\boldsymbol{x}^{k-1,i}, \boldsymbol{x}^{k-1,i-1})} + \hat{c}N\psi(\mathcal{S}_k) \\ &\leq \sqrt{\tfrac{2N}{\rho}}\mathbf{max}\{\tfrac{1}{\sqrt{\sigma_1}}, \ldots, \tfrac{1}{\sqrt{\sigma_N}}\}\sqrt{\mathcal{S}_{k-1} - \mathcal{S}_k} + \hat{c}N\psi(\mathcal{S}_{k-1}). \end{aligned} \tag{8.3}$$

On the other hand, for $i = 1, \ldots, N$, we have

$$\|x_i^k - x_i^\star\| \leq \|x_i^{k+1} - x_i^k\| + \|x_i^{k+1} - x_i^\star\| \leq \ldots \leq \sum_{j=k}^{\infty}\|x_i^{j+1} - x_i^j\|.$$

This inequality, together with (8.3), yields

$$\sum_{i=1}^{N}\|x_i^k - x_i^\star\| \leq \sqrt{\tfrac{2N}{\rho}}\mathbf{max}\{\tfrac{1}{\sqrt{\sigma_1}}, \ldots, \tfrac{1}{\sqrt{\sigma_N}}\}\sqrt{\mathcal{S}_{k-1} - \mathcal{S}_k} + \hat{c}N\psi(\mathcal{S}_{k-1}),$$

leading to

$$\|x_i^k - x_i^\star\| \leq c\mathbf{max}\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} \quad i = 1, \ldots, N, \tag{8.4}$$

where $c := \sqrt{\frac{2N}{\rho}}\mathbf{max}\left\{\frac{1}{\sqrt{\sigma_1}},\dots,\frac{1}{\sqrt{\sigma_N}}\right\} + \widehat{c}N$ and $\psi(s) := \frac{\kappa}{1-\theta}s^{1-\theta}$. Let us consider the nonlinear equation

$$\sqrt{\mathcal{S}_{k-1}} - \frac{\kappa}{1-\theta}\mathcal{S}_{k-1}^{1-\theta} = 0,$$

which has a solution at $\mathcal{S}_{k-1} = \left(\frac{(1-\theta)}{\kappa}\right)^{\frac{2}{1-2\theta}}$. Form the monotonicity of $\mathcal{S}_k$, there exists $\hat{k} \in \mathbb{N}$ such that for $k \geq \hat{k}$ (8.4) holds and

$$\mathcal{S}_{k-1} \leq \left(\frac{1-\theta}{\kappa}\right)^{\frac{2}{1-2\theta}}.$$

We now consider two cases: (a) $\theta \in (0, 1/2]$; (b) $\theta \in (1/2, 1)$. In Case (a), if $\theta \in (0, 1/2)$, then $\psi(\mathcal{S}_{k-1}) \leq \sqrt{\mathcal{S}_{k-1}}$. If $\theta = 1/2$, then $\psi(\mathcal{S}_{k-1}) = \frac{\kappa}{1-\theta}\sqrt{\mathcal{S}_{k-1}}$, i.e.,

$$\mathbf{max}\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} = \mathbf{max}\{1, \tfrac{\kappa}{1-\theta}\}\sqrt{\mathcal{S}_{k-1}}.$$

Therefore, it holds that $\mathbf{max}\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} \leq \mathbf{max}\{1, \frac{\kappa}{1-\theta}\}\sqrt{\mathcal{S}_{k-1}}$. In Case (b), we have that

$$\psi(\mathcal{S}_{k-1}) \geq \sqrt{\mathcal{S}_{k-1}},$$

i.e., $\mathbf{max}\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} = \frac{\kappa}{1-\theta}\mathcal{S}_{k-1}^{1-\theta}$. Then, it follows from (8.4) that (8.2) holds for all $k \geq \bar{k} := \mathbf{max}\{\tilde{k}, \hat{k}\}$.

In the second part of the proof, we will show the assertions in the statement of the theorem. For $(\mathcal{G}_i^k, \dots, \mathcal{G}_N^k) \in \partial\varphi(\boldsymbol{x}^k)$ as defined in Proposition 4.3, by Proposition 4.1(i), we infer

$$\begin{aligned}
\mathcal{S}_{k-1} - \mathcal{S}_k = \varphi(x^{k-1}) - \varphi(x^k) &\geq \rho\sum_{i=1}^{N}\mathbf{D}_h(x^{k-1,i}, x^{k-1,i-1}) \geq \frac{\rho}{2}\sum_{i=1}^{N}\sigma_i\|x_i^k - x_i^{k-1}\|^2 \\
&\geq \frac{\rho}{2N}\min\{\sigma_1,\dots,\sigma_N\}\left(\sum_{i=1}^{N}\|x_i^k - x_i^{k-1}\|\right)^2 \geq \frac{\rho}{2N\overline{c}^2}\min\{\sigma_1,\dots,\sigma_N\}\|(\mathcal{G}_i^k,\dots,\mathcal{G}_N^k)\|^2 \\
&\geq \frac{\rho}{2N\overline{c}^2}\min\{\sigma_1,\dots,\sigma_N\}\mathbf{dist}(0,\partial\varphi(x^k))^2 \geq \frac{\rho}{2N\overline{c}^2\kappa^2}\min\{\sigma_1,\dots,\sigma_N\}\mathcal{S}_{k-1}^{2\theta} = \widetilde{c}\,\mathcal{S}_{k-1}^{2\theta},
\end{aligned}$$

with $\widetilde{c} := \frac{\rho}{2N\overline{c}^2\kappa^2}\min\{\sigma_1,\dots,\sigma_N\}$ and for all $k \geq \bar{k}$. Hence, all assumptions of Fact 4.6 hold with $\alpha = 2\theta$. Therefore, our results follows from this fact and (8.2).

□

Acknowledgements We would like to thank the anonymous reviewers for their insightful comments that helped improve the paper; in particular, one of the reviewers gave a suggestion that leads to the kernel function in Proposition 5.1. The first author is grateful to Andreas Themelis for his useful comments and discussions on the paper. MA and PP acknowledge the support by the Research Foundation Flanders (FWO) research projects G086518N andG086318N; Research Council KU Leuven C1 project No. C14/18/068; Fonds de la Recherche Scientifique - FNRS andthe Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO) under EOS project no 30468160 (SeLMA). LTKH andNG also acknowledge the support by the European Research Council (ERC starting grant no 679515).

References

1. Ahookhosh, M.: Accelerated first-order methods for large-scale convex optimization: nearly optimal complexity under strong convexity. Math. Methods Oper. Res. **89**(3), 319–353 (2019)
2. Ahookhosh, M., Hien, L.T.K., Gillis, N., Patrinos, P.: A block inertial bregman proximal algorithm for nonsmooth nonconvex problems with application to symmetric nonnegative matrix tri-factorization. J. Optim. Theory Appl. (2021)
3. Ahookhosh, M., Themelis, A., Patrinos, P.: A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. SIAM J. Optim. **31**(1), 653–685 (2021)
4. Araújo, U., Saldanha, B., Galvão, R., Yoneyama, T., Chame, H., Visani, V.: The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr. Intell. Lab. Syst. **57**(2), 65–73 (2001)
5. Armijo, L.: Minimization of functions having Lipschitz continuous first partial derivatives. Pac. J. Math. **16**(1), 1–3 (1966)
6. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Alternating proximal algorithms for weakly coupled convex minimization problems. applications to dynamical games and PDE's. J. Convex Anal. **15**(3), 485 (2008)
7. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
8. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Program. **137**(1), 91–129 (2013)
9. Attouch, H., Redont, P., Soubeyran, A.: A new class of alternating proximal minimization algorithms with costs-to-move. SIAM J. Optim. **18**(3), 1061–1081 (2007)
10. Attouch, H., Soubeyran, A.: Inertia and reactivity in decision making as cognitive variational inequalities. J. Conv. Anal. **13**(2), 207 (2006)
11. Auslender, A.: Optimisation méthodes numériques. Mason, Paris (1976)
12. Bauschke, H.H., Bolte, J., Chen, J., Teboulle, M., Wang, X.: On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. J. Optim. Theory Appl. **182**, 1068–1087 (2019)
13. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2016)
14. Bauschke, H.H., Dao, M.N., Lindstrom, S.B.: Regularizing with Bregman–Moreau envelopes. SIAM J. Optim. **28**(4), 3208–3228 (2018)
15. Beck, A., Pauwels, E., Sabach, S.: The cyclic block conditional gradient method for convex optimization problems. SIAM J. Optim. **25**(4), 2024–2049 (2015)
16. Beck, A., Sabach, S., Teboulle, M.: An alternating semiproximal method for nonconvex regularized structured total least squares problems. SIAM J. Matrix Anal. Appl. **37**(3), 1129–1150 (2016)
17. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
18. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM J. Optim. **23**(4), 2037–2060 (2013)
19. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Prentice-Hall, Inc., Hoboken (1989)
20. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2007)
21. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
22. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**(6), 3319–3363 (2010)
23. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1–2), 459–494 (2014)
24. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. **28**(3), 2131–2151 (2018)

25. Boţ, R.I., Csetnek, E.R.: An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems. J. Optim. Theory Appl. **171**(2), 600–616 (2016)
26. Boţ, R.I., Nguyen, D.K.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. Math. Oper. Res. **45**(2), 682–712 (2020)
27. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
28. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM J. Optim. **3**(3), 538–543 (1993)
29. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. John Wiley & Sons, Hoboken (2009)
30. Combettes, P.L., Pesquet, J.C.: Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. **25**(2), 1221–1248 (2015)
31. Van den Dries, L.: Tame Topology and o-Minimal Structures, vol. 248. Cambridge University Press, Cambridge (1998)
32. Fercoq, O., Bianchi, P.: A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. SIAM J. Optim. **29**(1), 100–134 (2019)
33. Fu, X., Huang, K., Sidiropoulos, N.D., Ma, W.K.: Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. IEEE Signal Process. Mag. **36**(2), 59–80 (2019)
34. Gillis, N.: The why and how of nonnegative matrix factorization. Regular. Optim. Kernels Support Vector Mach. **12**(257), 257–291 (2014)
35. Gillis, N., Vavasis, S.A.: Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**(4), 698–714 (2013)
36. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. Operat. Res. Lett. **26**(3), 127–136 (2000)
37. Hanzely, F., Richtárik, P.: Fastest rates for stochastic mirror descent methods. arXiv:1803.07374 (2018)
38. Hanzely, F., Richtarik, P., Xiao, L.: Accelerated bregman proximal gradient methods for relatively smooth convex optimization. Comput Optim Appl **22**, 1–36 (2021)
39. Kimura, K., Tanaka, Y., Kudo, M.: A fast hierarchical alternating least squares algorithm for orthogonal nonnegative matrix factorization. In: D. Phung, H. Li (eds.) Proceedings of the Sixth Asian Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 39, pp. 129–141. PMLR, Nha Trang City, Vietnam (2015). http://proceedings.mlr.press/v39/kimura14.html
40. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l'institut Fourier **48**(3), 769–783 (1998)
41. Latafat, P., Freris, N.M., Patrinos, P.: A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. IEEE Trans. Autom. Cont. **64**(10), 4050–4065 (2019)
42. Latafat, P., Themelis, A., Patrinos, P.: Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems. Math. Program. 1–30. arxiv.org/abs/1906.10053 (2021)
43. Li, Q., Zhu, Z., Tang, G., Wakin, M.B.: Provable Bregman-divergence based methods for nonconvex and non-Lipschitz problems. arXiv preprint arXiv:1904.09712 (2019)
44. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. Les équations aux dérivées partielles pp. 87–89 (1963)
45. Łojasiewicz, S.: Sur la géométrie semi- et sous- analytique. Annales de l'institut Fourier **43**(5), 1575–1595 (1993)
46. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Optim. **28**(1), 333–354 (2018)
47. Mukkamala, M.C., Ochs, P., Pock, T., Sabach, S.: Convex-concave backtracking for inertial bregman proximal gradient algorithms in nonconvex optimization. SIAM J. Math. Data Sci. **2**(3), 658–682 (2020)
48. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
49. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
50. Pauca, V.P., Piper, J., Plemmons, R.J.: Nonnegative matrix factorization for spectral data analysis. Linear Algebra Appl. **416**(1), 29–47 (2006)

51. Pock, T., Sabach, S.: Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM J. Imag .Sci. **9**(4), 1756–1787 (2016)
52. Pompili, F., Gillis, N., Absil, P.A., Glineur, F.: Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. Neurocomputing **141**, 15–25 (2014)
53. Razaviyayn, M., Hong, M., Luo, Z.Q.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Optim. **23**(2), 1126–1153 (2013)
54. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**(1–2), 1–38 (2014)
55. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer Science & Business Media, Berlin (2011)
56. Choi, S.: Algorithms for orthogonal nonnegative matrix factorization. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1828–1832 (2008)
57. Shefi, R., Teboulle, M.: On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. EURO J. Comput. Optim. **4**(1), 27–46 (2016)
58. Tam, M.K.: Regularity properties of non-negative sparsity sets. J. Math. Anal. Appl. **447**(2), 758–777 (2017)
59. Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. **170**(1), 67–96 (2018)
60. Themelis, A., Ahookhosh, M., Patrinos, P.: On the acceleration of forward-backward splitting via an inexact Newton method. In: Luke, R., Bauschke, H., Burachik, R. (eds.) Splitting Algorithms, Modern Operator Theory, and Applications, pp. 363–412. Springer, Berlin (2019)
61. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**(3), 475–494 (2001)
62. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**(1–2), 387–423 (2009)
63. Wang, X., Yuan, X., Zeng, S., Zhang, J., Zhou, J.: Block coordinate proximal gradient method for nonconvex optimization problems: convergence analysis. http://www.optimization-online.org/DB_HTML/2018/04/6573.html (2018)

Authors and Affiliations

Masoud Ahookhosh[1] · Le Thi Khanh Hien[2] · Nicolas Gillis[2] · Panagiotis Patrinos[3]

Le Thi Khanh Hien
ThiKhanhHien.LE@umons.ac.be

Nicolas Gillis
nicolas.gillis@umons.ac.be

Panagiotis Patrinos
panos.patrinos@esat.kuleuven.be

[1] Department of Mathematics, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium

[2] Department of Mathematics and Operational Research, Faculté polytechnique, Université de Mons. Rue de Houdain 9, 7000 Mons, Belgium

[3] Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium