

# A block inertial Bregman proximal algorithm for nonsmooth nonconvex problems with application to symmetric nonnegative matrix tri-factorization

Masoud Ahooghosh · Le Thi Khanh Hien · Nicolas Gillis · Panagiotis Patrinos

the date of receipt and acceptance should be inserted later

**Abstract** We propose **BIBPA**, a block inertial Bregman proximal algorithm for minimizing the sum of a block relatively smooth function (that is, relatively smooth concerning each block) and block separable nonsmooth nonconvex functions. We show that the cluster points of the sequence generated by BIBPA are critical points of the objective under standard assumptions, and this sequence converges globally when a regularization of the objective function satisfies the Kurdyka-Łojasiewicz (KL) property. We also provide the convergence rate when a regularization of the objective function satisfies the Łojasiewicz inequality. We apply BIBPA to the symmetric nonnegative matrix tri-factorization (SymTriNMF) problem, where we propose kernel functions for SymTriNMF and provide closed-form solutions for subproblems of **BIBPA**.

Communicated by Hedy Attouch

**Keywords** Nonsmooth nonconvex optimization, Block Bregman proximal algorithm, Inertial effects, Block relative smoothness, Symmetric nonnegative matrix tri-factorization

**Mathematics Subject Classification (2010)** 90C06, 90C25, 90C26, 49J52, 49J53.

## 1 Introduction

This paper is concerned with the minimization of the sum of a block relatively smooth (see [Definition 2.2](#)), and a block separable (nonsmooth) nonconvex function. Although this problem has a simple structure, it covers a broad range of optimization problems arising in signal and image processing, machine learning, and inverse problems. In our block-structured nonconvex setting, the most common class of methodologies is *first-order* ones, where the central to their convergence analysis is the so-called *descent lemma* in both the Euclidean (e.g.,

---

M. Ahooghosh  
Department of Mathematics, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium .  
masoud.ahookhosh@uantwerp.be

L. T. K. Hien and N. Gillis  
Department of Mathematics and Operational Research, Faculté polytechnique, Université de Mons. Rue de Houdain 9, 7000 Mons, Belgium. . {ThiKhanhHien.LE,nicolas.gillis}@umons.ac.be

P. Patrinos  
Department of Electrical Engineering (ESAT-STADIUS) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. panos.patrinos@esat.kuleuven.be

[?,?,?,?]) and the non-Euclidean setting (e.g., [?,?,?]). While for the Euclidean case, the descent lemma is guaranteed if the function has Lipschitz continuous gradients, in the non-Euclidean setting it holds for *relatively smooth* functions encompassing the class of smooth functions with Lipschitz gradients.

In the *Euclidean setting*, there is a large number of alternating minimization algorithms for handling our structured problem such as block coordinate methods [?,?,?,?] and Gauss-Seidel methods [?,?,?], proximal alternating minimization [?,?], and proximal alternating linearized minimization [?,?,?]. In the *non-Euclidean setting*, several algorithms have been proposed, namely, Bregman forward-backward splitting [?,?,?,?,?], accelerated Bregman forward-backward splitting [?,?], stochastic mirror descent methods [?], and Bregman proximal alternating linearized minimization [?].

In order to establish the global convergence of generic algorithms for (nonsmooth) nonconvex problems, one needs to assume that the celebrated Kurdyka-Łojasiewicz inequality (see Definition 3.11) is satisfied as a feature of the underlying problem's class. The earliest abstract convergence theorem was introduced by Attouch et al. [?] and by Bolte et al. [?], relying on the following conditions that an algorithm should satisfy: (i) *sufficient decrease condition of the cost function*; (ii) *relative error conditions*; (iii) *subsequential convergence*. These conditions are shown to be satisfied by many algorithms [?]. In [?], these conditions were extended for proximal alternating linearized minimization. In the case of inertial proximal point algorithms [?,?,?], it was shown that some Lyapunov function satisfies the sufficient decrease condition, which leads to a generalization of the abstract convergence theorem. A generalization of this theorem was introduced for variable metric algorithms in [?], which has been recently extended for inertial variable metric algorithms [?]. In this paper, we show that the results of [?] can cover the global convergence of algorithms in non-Euclidean settings.

#### 1.1 Contribution

Our contribution is twofold:

1) (*Block inertial Bregman proximal algorithm*) We introduce **BIBPA**, a block generalization of the Bregman proximal gradient method [?] *with an inertial force*. We extend the notion of relative smoothness [?,?,?] to its block version (with different kernels for each block) to support our structured nonconvex problems. Notably, these kernel functions are block-wise convex, a property that does not necessarily imply their joint convexity for all blocks. Unlike the global convergence theorem in [?,?] that verifies the *sufficient decrease condition* and *relative error condition* on the cost function, for **BIBPA** these properties hold for a *Lyapunov function* including Bregman terms (see the equation (3.8)). Then, the global convergence of **BIBPA** is studied under the KL property, and its convergence rate is studied for Łojasiewicz-type KL functions.

2) (*Globally convergent scheme for solving the SymTriNMF problem*) With appropriate selection of kernel functions for Bregman distances, it turns out that the objective of the *symmetric nonnegative matrix tri-factorization* (SymTriNMF) problem is block relatively smooth, and the corresponding subproblems can be solved in closed form, an important property when dealing with machine learning problems that include a large number of variables. To the best of our knowledge, **BIBPA** is the first scheme with a rigorous theoretical guarantee of convergence for the SymTriNMF problem.

#### 1.2 Related works

There are three papers [?,?,?] that are closely related to this paper. In [?], we introduced a multi-block relative smoothness condition that exploits a single kernel function for all

blocks, while in the current paper we assume a block relative smoothness condition allowing a different kernel function for each block. Moreover, our algorithm **BIBPA** involves dynamic step-sizes and inertial terms for each block that makes our derivation and analysis different from those of [?]. Beside of the algorithmic differences with [?], we use nonseparable (nonconvex) kernels as apposed to the separable convex kernel used in [?] for each block. An inertial Bregman proximal gradient algorithm was presented in [?] for composite minimization that does not support our block structure nonconvex problems and therefore is different in derivation and analysis concerning our work.

### 1.3 Organization

The remainder of this paper is organized as follows. While **Section 2** discusses the problem statement and the block relative smoothness, **Section 3** introduces and analyzes a block inertial Bregman proximal algorithm (**BIBPA**). In **Section 4**, it is shown the **BIBPA**'s subproblems are solved in closed form. Some conclusion are delivered in **Section 5**.

## 2 Problem statement and block relative smoothness

We consider the structured nonsmooth nonconvex minimization problem

$$\underset{x \in \bar{C}}{\text{minimize}} \quad \Phi(x) \equiv f(x) + \sum_{i=1}^N g_i(x_i), \quad (2.1)$$

where  $C$  is a nonempty, convex, and open set in  $\mathbb{R}^n$  and  $\bar{C}$  denotes its closure. Setting  $n = \sum_{i=1}^N n_i$  and  $i \in [N]$ , we assume the following hypotheses:

**Assumption I** (requirements for composite minimization (2.1)).

- A1  $g_i : \mathbb{R}^{n_i} \rightarrow \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$  is proper and lower semicontinuous (lsc);
- A2  $h_i : \mathbb{R}^{n_i} \rightarrow \bar{\mathbb{R}}$  is  $i$ -th block Legendre,  $\mathbf{int\,dom} \, h_1 = \dots = \mathbf{int\,dom} \, h_N$ ,  $\bar{C} \subseteq \overline{\mathbf{dom} \, h_1}$ , and  $\mathbf{dom} \, g \cap C \neq \emptyset$  with  $g := \sum_{i=1}^N g_i$ ;
- A3  $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  is  $C^1(\mathbf{int\,dom} \, h_1)$  and  $(L_1, \dots, L_N)$ -smooth relative to  $(h_1, \dots, h_N)$ ;
- A4  $\mathbf{arg\,min} \{ \Phi(x) \mid x \in \bar{C} \} \neq \emptyset$ .

### 2.1 Notation

We denote by  $\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$  the extended-real line. We use boldface lower-case letters (e.g.,  $\mathbf{x}, \mathbf{y}, \mathbf{z}$ ) for vectors in  $\mathbb{R}^n$  and use normal lower-case letters (e.g.,  $z_i, x_i, y_i$ ) for vectors in  $\mathbb{R}^{n_i}$ , for  $n_i \in \mathbb{N}$ . For the identity matrix  $I_n$ , we set  $U_i \in \mathbb{R}^{n \times n_i}$  such that  $I_n = (U_1, \dots, U_N) \in \mathbb{R}^{n \times n}$ . The set of cluster points of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  is denoted as  $\omega(\mathbf{x}^k)$ . A function  $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  is proper if  $f > -\infty$  and  $f \not\equiv \infty$ , in which case its domain is defined as the set  $\mathbf{dom} \, f := \{ \mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) < \infty \}$ . A vector  $\mathbf{v} \in \partial f(\mathbf{x})$  is a subgradient of  $f$  at  $\mathbf{x}$ , and the set of all such vectors is called the subdifferential  $\partial f(\mathbf{x})$  [?, Definition 8.3], i.e.

$$\partial f(\mathbf{x}) = \{ \mathbf{v} \in \mathbb{R}^n \mid \exists (\mathbf{x}^k, \mathbf{v}^k)_{k \in \mathbb{N}} \text{ s.t. } \mathbf{x}^k \rightarrow \mathbf{x}, f(\mathbf{x}^k) \rightarrow f(\mathbf{x}), \widehat{\partial} f(\mathbf{x}^k) \ni \mathbf{v}^k \rightarrow \mathbf{v} \},$$

where  $\widehat{\partial} f(\mathbf{x}) = \{ \mathbf{v} \in \mathbb{R}^n \mid f(\mathbf{z}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{z} - \mathbf{x} \rangle + o(\|\mathbf{z} - \mathbf{x}\|), \forall \mathbf{z} \in \mathbb{R}^n \}$ .

## 2.2 Block relative smoothness

We first describe the notion of *block relative smoothness*, which is an extension of the relative smoothness [?,?]. To this end, we introduce the notion of *block kernel* functions, which coincides with the classical one (cf. [?, Definition 2.1]) for  $N = 1$ .

**Definition 2.1** (*i*-th block convexity and kernel function). *Let  $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  be a proper and lower semicontinuous (lsc) function with  $\text{int dom } h \neq \emptyset$  and such that  $h \in C^1(\text{int dom } h)$ . For a fixed vector  $\mathbf{x} \in \mathbb{R}^n$  and  $i \in \{1, \dots, N\}$ , we say that  $h$  is*

- (i)  *$i$ -th block (strongly/strictly) convex if the function  $h(\mathbf{x} + U_i(\cdot - x_i))$  is (strongly/strictly) convex for all  $\mathbf{x} \in \text{dom } h$ ;*
- (ii) *a  $i$ -th block kernel function if  $h$  is  $i$ -th block convex and  $h(\mathbf{x} + U_i(\cdot - x_i))$  is 1-coercive for all  $\mathbf{x} \in \text{dom } h$ , i.e.,  $\lim_{\|z\| \rightarrow \infty} \frac{h(\mathbf{x} + U_i(z - x_i))}{\|z\|} = \infty$ ;*
- (iii)  *$i$ -th block essentially smooth, if for every sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}} \subseteq \text{int dom } h$  converging to a boundary point of  $\text{dom } h$ , we have  $\|\nabla_i h(\mathbf{x}^k)\| \rightarrow \infty$ , where  $\nabla_i h$  is the partial derivative of  $h$  with respect to the  $i$ th block of variables;*
- (iv)  *$i$ -th block Legendre if it is  $i$ -th block essentially smooth and  $i$ -th block strictly convex.*

Let  $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  be a Legendre function. Then, the classical definition of *Bregman distances* (cf. [?]) leads to the function  $\mathbf{D}_h : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  given by

$$\mathbf{D}_h(\mathbf{y}, \mathbf{x}) := \begin{cases} h(\mathbf{y}) - h(\mathbf{x}) - \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle & \text{if } \mathbf{y} \in \text{dom } h, \mathbf{x} \in \text{int dom } h, \\ \infty & \text{otherwise.} \end{cases} \quad (2.2)$$

However, in the remainder of this paper, we extend this definition for the cases that  $h$  is only an  $i$ -th block Legendre function. Fixing all blocks except the  $i$ -th one, the Bregman distance (2.2) will reduce to  $\mathbf{D}_h(\mathbf{x} + U_i(y_i - x_i), \mathbf{x}) = h(\mathbf{x} + U_i(y_i - x_i)) - h(\mathbf{x}) - \langle \nabla_i h(\mathbf{x}), y_i - x_i \rangle$ , which measures the proximity between  $\mathbf{x} + U_i(y_i - x_i)$  and  $\mathbf{x}$  with respect to the  $i$ -th block of variables. Moreover, the kernel  $h$  is  $i$ -th block convex if and only if  $\mathbf{D}_h(\mathbf{x} + U_i(y_i - x_i), \mathbf{x}) \geq 0$  for all  $\mathbf{x} + U_i(y_i - x_i) \in \text{dom } h$  and  $\mathbf{x} \in \text{int dom } h$ . Note that if  $h$  is  $i$ -th block strictly convex, then  $\mathbf{D}_h(\mathbf{x} + U_i(y_i - x_i), \mathbf{x}) = 0$  if and only if  $x_i = y_i$ .

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

**Definition 2.2** (block relative smoothness). *For  $i \in [N] = \{1, \dots, N\}$ , let  $h_i : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  be  $i$ -th block kernel functions and let  $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  be a proper and lsc function. If there exists  $L_i > 0$ ,  $i \in [N]$ , such that  $L_i h_i(\mathbf{x} + U_i(\cdot - x_i)) - f(\mathbf{x} + U_i(\cdot - x_i))$  are convex for all  $\mathbf{x}, \mathbf{x} + U_i(\cdot - x_i) \in \text{int dom } h_i$ , then,  $f$  is called  $(L_1, \dots, L_N)$ -smooth relative to  $(h_1, \dots, h_N)$ .*

Note that if  $N = 1$ , the block relative smoothness is reduced to standard relative smoothness [?,?]. If  $f$  is  $L_f$ -Lipschitz continuous, then both  $L_f/2 \|\cdot\|^2 - f$  and  $L_f/2 \|\cdot\|^2 + f$  are convex, i.e., the relative smoothness of  $f$  generalizes the notions of Lipschitz continuity.

**Proposition 2.3** (characterization of block relative smoothness). *For  $i \in [N]$ , let  $h_i : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  be  $i$ -th block kernels and let  $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  be a proper lsc function and  $f \in C^1$ . Then, the following statements are equivalent:*

- (a)  *$f$  is  $(L_1, \dots, L_N)$ -smooth relative to  $(h_1, \dots, h_N)$ ;*
- (b) *for all  $(\mathbf{x}, \mathbf{y}) \in \text{int dom } h_i \times \text{int dom } h_i$  and  $i \in [N]$ ,*

$$f(\mathbf{x} + U_i(y_i - x_i)) \leq f(\mathbf{x}) + \langle \nabla_i f(\mathbf{x}), y_i - x_i \rangle + L_i \mathbf{D}_{h_i}(\mathbf{x} + U_i(y_i - x_i), \mathbf{x}); \quad (2.3)$$

(c) for all  $(\mathbf{x}, \mathbf{y}) \in \text{int dom } h_i \times \text{int dom } h_i$  and  $i \in [N]$ ,

$$\langle \nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{y}), \mathbf{x}_i - \mathbf{y}_i \rangle \leq L_i \langle \nabla_i h_i(\mathbf{x}) - \nabla_i h_i(\mathbf{y}), \mathbf{x}_i - \mathbf{y}_i \rangle; \quad (2.4)$$

(d) if  $f \in C^2(\text{int dom } f)$  and  $h_i \in C^2(\text{int dom } h_i)$ , then denoting  $\nabla_{ii}^2 h_i = \frac{\partial^2 h_i}{\partial \mathbf{x}_i^2}$ , we have

$$L_i \nabla_{ii}^2 h_i(\mathbf{x}) - \nabla_{ii}^2 f(\mathbf{x}) \geq 0, \quad \forall \mathbf{x} \in \text{int dom } h_i, \quad i \in [N]. \quad (2.5)$$

*Proof.* Fixing all the blocks except one, it is a direct extension of [?, Proposition 1.1].  $\square$

### 2.3 Motivating example: symmetric nonnegative matrix tri-factorization

We consider a symmetric matrix  $X \in \mathbb{R}^{m \times m}$  and aim to decompose it in the form  $X = UVU^T$ , where  $U \in \mathbb{R}_+^{m \times r}$  and  $V \in \mathbb{R}_+^{r \times r}$ . This translates to the minimization of  $\frac{1}{2} \|X - UVU^T\|_F^2$  for  $U, V \geq 0$ , leading to the unconstrained problem

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{r \times r}} \frac{1}{2} \|X - UVU^T\|_F^2 + \delta_{U \geq 0} + \delta_{V \geq 0}. \quad (2.6)$$

**Proposition 2.4** (block relative smoothness of SymTriNMF objective). *Let functions  $h_1 : \mathbb{R}^{m \times r} \times \mathbb{R}^{r \times r} \rightarrow \overline{\mathbb{R}}$  and  $h_2 : \mathbb{R}^{m \times r} \times \mathbb{R}^{r \times r} \rightarrow \overline{\mathbb{R}}$  be strongly convex kernel functions as*

$$h_1(U, V) := \frac{a_1}{4} \|V\|_F^2 \|U\|_F^4 + \frac{b_1}{2} (\|X\|_F \|V\|_F + \varepsilon_1) \|U\|_F^2, \quad (2.7)$$

$$h_2(U, V) := \frac{a_2}{2} (\|U\|_F^4 + \varepsilon_2) \|V\|_F^2. \quad (2.8)$$

with  $\varepsilon_1, \varepsilon_2 > 0$ . Then the function  $f : \mathbb{R}^{m \times r} \times \mathbb{R}^{r \times r} \rightarrow \overline{\mathbb{R}}$  given by  $f(U, V) := \frac{1}{2} \|X - UVU^T\|_F^2$  is  $(L_1, L_2)$ -smooth relative to  $(h_1, h_2)$  with

$$L_1 \geq \max \left\{ \frac{6}{a_1}, \frac{2}{b_1} \right\}, \quad \text{and} \quad L_2 \geq \frac{1}{a_2}. \quad (2.9)$$

*Proof.* Plugging the partial derivative  $\nabla_U f(U, V) = -XUV^T - X^T UV + UVU^T UV^T + UV^T U^T UV$  into the definition of directional derivative, we obtain

$$\begin{aligned} \nabla_{UU}^2 f(U, V)Z &= -2XZV^T + UVU^T ZV + UVZ^T UV + ZVU^T UV^T \\ &\quad + UV^T U^T ZV + UV^T Z^T UV + ZV^T U^T UV, \end{aligned}$$

which consequently leads to

$$\begin{aligned} \langle Z, \nabla_{UU}^2 f(U, V)Z \rangle &= -2\langle Z, XZV^T \rangle + \langle Z, UVU^T ZV \rangle + \langle Z, UVZ^T UV \rangle + \langle Z, ZVU^T UV^T \rangle \\ &\quad + \langle Z, UV^T U^T ZV \rangle + \langle Z, UV^T Z^T UV \rangle + \langle Z, ZV^T U^T UV \rangle \\ &\leq (2\|X\| \|V\| + 6\|U\|^2 \|V\|^2) \|Z\|_F^2. \end{aligned}$$

On the other hand, from  $\nabla_V h_1(U, V) = (a_1 \|U\|_F^2 \|V\|_F^2 + b_1 (\|X\|_F \|V\|_F + \varepsilon_1))U$ , we have

$$\nabla_{UU}^2 h_1(U, V)Z = (2a_1 \|V\|^2 \langle U, Z \rangle)U + (a_1 \|V\|_F^2 \|U\|_F^2 + b_1 (\|X\|_F \|V\|_F + \varepsilon_1))Z,$$

implying that

$$\begin{aligned} \langle Z, \nabla_{UU}^2 h_1(U, V)Z \rangle &\geq (a_1 \|V\|_F^2 \|U\|_F^2 + b_1 (\|X\|_F \|V\|_F + \varepsilon_1)) \|Z\|_F^2 \\ &\geq (a_1 \|V\|^2 \|U\|^2 + b_1 (\|X\| \|V\| + \varepsilon_1)) \|Z\|_F^2. \end{aligned}$$

Therefore, the inequality

$$\begin{aligned} \langle Z, (L_1 \nabla_{UU}^2 h_1(U, V) - \nabla_{UU}^2 f(U, V))Z \rangle \\ \geq \left( (L_1 a_1 - 6) \|V\|^2 \|U\|^2 + (L_1 b_1 - 2) \|X\| \|V\| + \varepsilon_1 L_1 \right) \|Z\|_F^2 \geq 0 \end{aligned}$$

holds if  $L_1 a_1 - 6 \geq 0$  and  $L_1 b_1 - 2 \geq 0$ , as claimed.

It follows from  $\nabla_V f(U, V) = U^T X U + U^T U V U^T U$  that

$$\nabla_{VV}^2 f(U, V)Z = \lim_{t \rightarrow 0} \frac{\nabla_U f(U + tZ, V) - \nabla_U f(U, V)}{t} = U^T U Z U^T U,$$

leading to the inequality  $\langle Z, \nabla_{VV}^2 f(U, V)Z \rangle = \langle Z, U^T U Z U^T U \rangle \leq \|U\|^4 \|Z\|_F^2$ . Now, using  $\nabla_V h_2(U, V) = a_2(\|U\|^4 + \varepsilon_2)V$ , we get  $\langle Z, \nabla_{VV}^2 h_2(U, V)Z \rangle = a_2(\|U\|^4 + \varepsilon_2) \|Z\|_F^2$ , i.e.,

$$\langle Z, (L_2 \nabla_{VV}^2 h_2(U, V) - \nabla_{VV}^2 f(U, V))Z \rangle = \left( (L_2 a_2 - 1) \|U\|^4 + \varepsilon_2 L_2 \right) \|Z\|_F^2 \geq 0$$

if  $L_2 a_2 - 1 \geq 0$ , giving our desired results.  $\square$

### 3 Block inertial Bregman proximal algorithm

This section discusses our algorithm, starting from the prox-boundedness extension [?].

**Definition 3.1** (block prox-boundedness). *A function  $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is block prox-bounded if for each  $i \in \{1, \dots, N\}$  there exists  $\gamma_i > 0$  and  $\mathbf{x} \in \mathbb{R}^n$  such that*

$$\inf_{z \in \mathbb{R}^{n_i}} \left\{ g(\mathbf{x} + U_i(z - x_i)) + \frac{1}{\gamma_i} \mathbf{D}_{h_i}(\mathbf{x} + U_i(z - x_i), \mathbf{x}) \right\} > -\infty.$$

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

$$\gamma_{i,g}^h := \sup_{\gamma_i > 0} \left\{ \gamma_i : \exists \mathbf{x} \in \mathbb{R}^n, \inf_{z \in \mathbb{R}^{n_i}} \left\{ g(\mathbf{x} + U_i(z - x_i)) + \frac{1}{\gamma_i} \mathbf{D}_{h_i}(\mathbf{x} + U_i(z - x_i), \mathbf{x}) \right\} > -\infty \right\}. \quad (3.1)$$

**Proposition 3.2** (characteristics of block prox-boundedness). *For  $h_i : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  and proper and lsc functions  $g_i : \mathbb{R}^{n_i} \rightarrow \overline{\mathbb{R}}$  ( $i \in [N]$ ), the following statements are equivalent:*

- (a)  $g = \sum_{i=1}^N g_i$  is block prox-bounded;
- (b) for all  $i \in [N]$ ,  $g_i + r_i h_i(\mathbf{x} + U_i(z - x_i))$  is bounded below on  $\mathbb{R}^{n_i}$  for some  $r_i \in \mathbb{R}$ ;
- (c) for all  $i \in [N]$ ,  $\liminf_{\|z\| \rightarrow \infty} g_i(z)/h_i(\mathbf{x} + U_i(z - x_i)) > -\infty$ .

*Proof.* The proof is a straightforward adaptation of [?, Proposition 2.7].  $\square$

For a given points  $\mathbf{x}^k, \mathbf{x}^{k-1} \in \mathbb{R}^n$  and  $\alpha_i^k \in \mathbb{R}$ , let us define the function  $\mathcal{M}_{h_i/\gamma_i^k} : \mathbf{int dom } h_i \times \mathbf{int dom } h_i \rightarrow \overline{\mathbb{R}}$  given by

$$\mathcal{M}_{h_i/\gamma_i^k}(\mathbf{x}, \mathbf{x}^k, \mathbf{x}^{k-1}) := \langle \nabla f(\mathbf{x}^k) - \frac{\alpha_i^k}{\gamma_i^k} (\mathbf{x}^k - \mathbf{x}^{k-1}), \mathbf{x} - \mathbf{x}^k \rangle + \frac{1}{\gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}, \mathbf{x}^k) + \sum_{i=1}^N g_i(x_i) \quad (3.2)$$

and the block inertial Bregman proximal mapping  $\mathbf{T}_{h_i/\gamma_i^k} : \mathbf{int dom } h_i \times \mathbf{int dom } h_i \rightrightarrows \mathbb{R}^{n_i}$  as

$$\begin{aligned} \mathbf{T}_{h_i/\gamma_i^k}(\mathbf{x}^k, \mathbf{x}^{k-1}) &:= \arg \min_{z \in \mathbb{R}^{n_i}} \mathcal{M}_{h_i/\gamma_i^k}(\mathbf{x}^k + U_i(z - x_i^k), \mathbf{x}^k, \mathbf{x}^{k-1}) \\ &= \arg \min_{z \in \mathbb{R}^{n_i}} \langle \nabla f(\mathbf{x}^k) - \frac{\alpha_i^k}{\gamma_i^k} (\mathbf{x}^k - \mathbf{x}^{k-1}), z - \mathbf{x}_i^k \rangle + \frac{1}{\gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^k + U_i(z - x_i^k), \mathbf{x}^k) + g_i(z), \end{aligned} \quad (3.3)$$

which can be set-valued by nonconvexity of  $g_i$  ( $i \in [N]$ ), and it reduces to the inertial Bregman forward-backward mapping for  $N = 1$ ; cf. [?]. For a given sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}}$ , we introduce the following notation

$$\mathbf{x}^{k,i} := (x_1^{k+1}, \dots, x_i^{k+1}, x_{i+1}^k, \dots, x_N^k), \quad (3.4)$$

i.e.,  $\mathbf{x}^{k,0} = \mathbf{x}^k$  and  $\mathbf{x}^{k,N} = \mathbf{x}^{k+1}$ . Using this notation and the mapping (3.3), we next introduce the *block inertial Bregman proximal algorithm* (BIBPA); see Algorithm 1.

---

**Algorithm 1 (BIBPA)** Block Inertial Bregman Proximal Algorithm

---

INPUT  $\mathbf{x}^0 \in \text{int dom } h_1$ ,  $I_n = (U_1, \dots, U_N) \in \mathbb{R}^{n \times n}$  with  $U_i \in \mathbb{R}^{n \times n_i}$  and the identity matrix  $I_n$ ,  $k = 0$ .

1: **while** some stopping criterion is not met (e.g., see Remark 3.10) **do**

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

3: **for**  $i \in [N]$  **do** choose  $\gamma_i^k$  and  $\alpha_i^k$  as Prop. 3.5 and compute

$$x_i^{k,i} \in \mathbf{T}_{h_i/\gamma_i^k}(\mathbf{x}^{k,i-1}, \mathbf{x}^{k-1}), \quad \mathbf{x}^{k,i} = \mathbf{x}^{k,i-1} + U_i(x_i^{k,i} - x_i^{k,i-1}); \quad (3.5)$$

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

OUTPUT A vector  $\mathbf{x}^k$ .

---

In order to verify the well-definedness of the iterations generated by BIBPA, we next investigate some important properties of the mapping  $\mathbf{T}_{h_i/\gamma_i^k}$ . To do so, we require the following assumption, which consequently implies  $\omega(\mathbf{x}^k) \subseteq \bar{C}$ .

**Assumption II.** For  $i \in [N]$  and  $\gamma_i^k \in (0, 1/L_i)$ , the inclusion  $\mathbf{x} + U_i(z - x_i) \in C$  holds true with  $z \in \mathbf{T}_{h_i/\gamma_i^k}(\mathbf{x}, \mathbf{y})$  and for all  $\mathbf{x}, \mathbf{y} \in \text{int dom } h_i$ .

**Proposition 3.3** (properties of the mapping  $\mathbf{T}_{h_i/\gamma_i^k}$ ). Under Assumption I and Assumption II,  $\gamma_i^k \in (0, \gamma_{i,g}^{h_i})$  for  $i \in [N]$ , and  $\mathbf{x}^k, \mathbf{x}^{k-1} \in \text{int dom } h_i$ , the following statements are true:

- (i)  $\mathbf{T}_{h_i/\gamma_i^k}(\mathbf{x}^k, \mathbf{x}^{k-1})$  is nonempty, compact, and outer semicontinuous;
- (ii)  $\text{dom } \mathbf{T}_{h_i/\gamma_i^k} = \text{int dom } h_i \times \text{int dom } h_i$ ;
- (iii) If  $x_i^{k,i} \in \mathbf{T}_{h_i/\gamma_i^k}(\mathbf{x}^{k,i-1}, \mathbf{x}^{k-1})$  for  $\gamma_i^k \in (0, 1/L_i)$ , then  $\mathbf{x}^{k,i} \in \text{int dom } h_i$ .

*Proof.* The proof follows from [?, Proposition 2.10] and Assumption II.  $\square$

In the subsequent lemma, we show that the cost function  $\Phi$  satisfies some necessary inequality that will be needed in the next result.

**Lemma 3.4** (cyclic inequality of the cost). Let Assumption I and Assumption II hold, and let  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  be generated by BIBPA. If  $h_i$  ( $i \in [N]$ ) is  $\sigma_i$ -block strongly convex, then we have

$$\Phi(\mathbf{x}^{k+1}) - \Phi(\mathbf{x}^k) \leq \sum_{i=1}^N \left( \left( \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \frac{1 - \gamma_i^k L_i}{\gamma_i^k} \right) \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}) \right). \quad (3.6)$$

*Proof.* For  $i \in \{1, \dots, N\}$  and  $x_i^{k,i} \in \mathbf{T}_{h_i/\gamma_i^k}(\mathbf{x}^{k,i-1}, \mathbf{x}^{k-1})$ , it holds that

$$\langle \nabla_i f(\mathbf{x}^{k,i-1}) - \frac{\alpha_i^k}{\gamma_i^k} (x_i^k - x_i^{k-1}), x_i^{k,i} - x_i^k \rangle + \frac{1}{\gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i}) \leq \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i-1}).$$

Together with [Assumption IA3](#) and [Proposition 2.3\(b\)](#), this implies

$$\begin{aligned}
f(\mathbf{x}^{k,i}) &\leq f(\mathbf{x}^{k,i-1}) + \langle \nabla_i f(\mathbf{x}^{k,i-1}), \mathbf{x}_i^{k,i} - \mathbf{x}_i^{k,i-1} \rangle + L_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) \\
&\leq f(\mathbf{x}^{k,i-1}) + \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i-1}) - \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i}) - \frac{1-\gamma_i^k L_i}{\gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \frac{\alpha_i^k}{\gamma_i^k} \langle \mathbf{x}_i^k - \mathbf{x}_i^{k-1}, \mathbf{x}_i^{k,i} - \mathbf{x}_i^k \rangle \\
&\leq f(\mathbf{x}^{k,i-1}) + \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i-1}) - \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i}) - \frac{1-\gamma_i^k L_i}{\gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) \\
&\quad + \frac{|\alpha_i^k|}{2\gamma_i^k} (\|\mathbf{x}_i^k - \mathbf{x}_i^{k-1}\|^2 + \|\mathbf{x}_i^{k,i} - \mathbf{x}_i^k\|^2) \\
&\leq f(\mathbf{x}^{k,i-1}) + \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i-1}) - \sum_{j=1}^N g_j(\mathbf{x}_j^{k,i}) + \left( \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \frac{1-\gamma_i^k L_i}{\gamma_i^k} \right) \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) \\
&\quad + \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}),
\end{aligned}$$

which yields

$$\Phi(\mathbf{x}^{k,i}) \leq \Phi(\mathbf{x}^{k,i-1}) + \left( \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \frac{1-\gamma_i^k L_i}{\gamma_i^k} \right) \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}). \quad (3.7)$$

Now, let us sum up both sides of (3.7) for  $i \in [N]$ , i.e.,

$$\begin{aligned}
\Phi(\mathbf{x}^{k+1}) - \Phi(\mathbf{x}^k) &= \sum_{i=1}^N (\Phi(\mathbf{x}^{k,i}) - \Phi(\mathbf{x}^{k,i-1})) \\
&\leq \sum_{i=1}^N \left( \left( \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \frac{1-\gamma_i^k L_i}{\gamma_i^k} \right) \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}) \right).
\end{aligned}$$

□

We notice that [Lemma 3.4](#) does not guarantee the monotonicity of the sequence  $(\Phi(\mathbf{x}^k))_{k \in \mathbb{N}}$ . For  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  and  $\delta_i \geq 0$ , we define the *Lyapunov function*  $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ ,

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x}) + \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}((x_1, \dots, x_i, y_{i+1}, \dots, y_N), (x_1, \dots, x_{i-1}, y_i, \dots, y_N)), \quad (3.8)$$

Note that  $\mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k) := \Phi(\mathbf{x}^{k+1}) + \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})$ . We denote by  $\mathcal{L}^{k+1}$  and  $\mathcal{L}^k$  the terms  $\mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)$  and  $\mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1})$ , respectively. We next prove the monotonicity of  $(\mathcal{L}^k)_{k \in \mathbb{N}}$ .

**Proposition 3.5** (descent property of the Lyapunov function). *Let [Assumption I](#) and [Assumption II](#) hold, let  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  be generated by [BIBPA](#), and let  $h_i$  ( $i \in [N]$ ) be  $\sigma_i$ -block strongly convex. If  $\lim_{k \rightarrow \infty} \alpha_i^k = \alpha_i$  and  $0 < \gamma_i \leq \frac{\sigma_i - 2|\alpha_i|}{\sigma_i L_i}$  and*

$$|\alpha_i^k| < \frac{\sigma_i}{2}, \quad 0 < \gamma_i \leq \gamma_i^k \leq \frac{\sigma_i - 2|\alpha_i^k|}{\sigma_i L_i}, \quad \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \leq \delta_i \leq \frac{1-\gamma_i^k L_i}{\gamma_i^k} - \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \quad i \in [N], \quad (3.9)$$

then, setting  $a_i := \frac{1-\gamma_i^k L_i}{\gamma_i^k} - \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \delta_i$  and  $b_i := \delta_i - \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k}$  for  $i \in [N]$ , we get

$$\mathcal{L}^{k+1} - \mathcal{L}^k \leq - \sum_{i=1}^N (a_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + b_i \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})), \quad (3.10)$$

i.e., the sequence  $(\mathcal{L}^k)_{k \in \mathbb{N}}$  is non-increasing and consequently  $\lim_{k \rightarrow \infty} \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) = 0$ , i.e.,  $\lim_{k \rightarrow \infty} \|\mathbf{x}^{k,i} - \mathbf{x}^{k,i-1}\| = 0$ , for all  $i \in [N]$ .



*Proof.* Using (3.6) and applying the Lyapunov function (3.8), we have

$$\begin{aligned} \mathcal{L}^{k+1} - \mathcal{L}^k &= \Phi(\mathbf{x}^{k+1}) - \Phi(\mathbf{x}^k) + \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) - \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}) \\ &\leq \sum_{i=1}^N \left( \left( \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \frac{1-\gamma_i^k L_i}{\gamma_i^k} + \delta_i \right) \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \left( \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \delta_i \right) \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}) \right), \end{aligned}$$

as claimed in (3.10). In order to guarantee the non-increasing property of the sequence  $(\mathcal{L}^k)_{k \in \mathbb{N}}$ , the inequalities  $a_i = \frac{1-\gamma_i^k L_i}{\gamma_i^k} - \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} - \delta_i \geq 0$ ,  $b_i = \delta_i - \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \geq 0$  should be satisfied, for  $i \in [N]$ , i.e.,  $\frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \leq \delta_i \leq \frac{1-\gamma_i^k L_i}{\gamma_i^k} - \frac{|\alpha_i^k|}{\sigma_i \gamma_i^k} \leq \frac{1-\gamma_i L_i}{\gamma_i}$   $i \in [N]$ , which is guaranteed by (3.9), i.e.,  $\mathcal{L}^{k+1} \leq \mathcal{L}^k$ . Together with (3.10), this yields that

$$\begin{aligned} \sum_{k=0}^p \sum_{i=1}^N a_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + b_i \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}) &\leq \sum_{k=0}^p (\mathcal{L}^k - \mathcal{L}^{k+1}) \\ &= \mathcal{L}^0 - \mathcal{L}^{p+1} \leq \mathcal{L}^0 - \inf \mathcal{L} < +\infty. \end{aligned}$$

Let  $p \rightarrow +\infty$ , the result follows from  $\mathbf{D}_{h_i}(\cdot, \cdot) \geq 0$  and block strong convexity of  $h_i$ .  $\square$

In convergence analysis of proximal algorithms, one usual assumption is the boundedness of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$ ; cf. [?,?]. A sufficient condition for this is given next.

**Corollary 3.6** (boundedness of iterations). *Suppose that all assumptions of Proposition 3.5 hold. Further, if  $\Phi$  has bounded level sets, then the sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  is bounded.*

*Proof.* It follows from Proposition 3.5 that  $\mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)$  is non-increasing, hence

$$\begin{aligned} \Phi(\mathbf{x}^{k+1}) &\leq \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k) = \Phi(\mathbf{x}^{k+1}) + \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) \\ &\leq \mathcal{L}(\mathbf{x}^1, \mathbf{x}^0) = \Phi(\mathbf{x}^1) + \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}(\mathbf{x}^{0,i}, \mathbf{x}^{0,i-1}) < \infty. \end{aligned}$$

Hence,  $\mathcal{N}(\mathbf{x}^1, \mathbf{x}^0) := \{\mathbf{x} \in \mathbb{R}^n \mid \Phi(\mathbf{x}) \leq \Phi(\mathbf{x}^1) + \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}(\mathbf{x}^{0,i}, \mathbf{x}^{0,i-1})\}$  encompasses  $(\mathbf{x}^k)_{k \in \mathbb{N}}$ , i.e.,  $(\mathbf{x}^k)_{k \in \mathbb{N}} \subseteq \mathcal{N}(\mathbf{x}^1, \mathbf{x}^0)$ . Since  $\Phi$  has bounded level sets, we have  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  are bounded.  $\square$

The next proposition provides a lower bound for  $\sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})}$ .

**Proposition 3.7** (relative error condition). *Let Assumption I and Assumption II hold, let  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  be generated by BIBPA, and let  $h_i$  ( $i \in [N]$ ) be  $\sigma_i$ -block strongly convex. Furthermore, suppose that  $\nabla_i f$ ,  $\nabla_i h$ , ( $i \in [N]$ ) are locally Lipschitz on bounded sets with Lipschitz moduli  $\bar{L}$  and  $\bar{L}_i > 0$ ,  $\nabla_{ii}^2 h_i$  is bounded on bounded set with constants  $\bar{L}_i$  ( $i \in [N]$ ) and that the sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  is bounded. For a fixed  $k \in \mathbb{N}$  and  $j \in [N]$ , we define*

$$\mathcal{G}_j^{k+1} := (\mathcal{V}_j^{k+1}, \mathcal{W}_j^{k+1}), \quad (3.11)$$

where

$$\begin{aligned} \mathcal{V}_j^{k+1} &:= \sum_{i=j}^N \delta_i (\nabla_j h_i(\mathbf{x}^{k,i}) - \nabla_j h_i(\mathbf{x}^{k,i-1})) + \frac{1}{\gamma_j^k} (\nabla_j h_j(\mathbf{x}^{k,j-1}) - \nabla_j h_j(\mathbf{x}^{k,j})) \\ &\quad + \frac{\alpha_j^k}{\gamma_j^k} (\mathbf{x}_j^k - \mathbf{x}_j^{k-1}) + \nabla_j f(\mathbf{x}^{k+1}) - \nabla_j f(\mathbf{x}^{k,j-1}) \\ \mathcal{W}_j^{k+1} &:= \sum_{i=1}^{j-1} \delta_i (\nabla_j h_i(\mathbf{x}^{k,i}) - \nabla_j h_i(\mathbf{x}^{k,i-1})) - \nabla_{jj}^2 h_j(\mathbf{x}^{k,j-1}) (\mathbf{x}_j^{k+1} - \mathbf{x}_j^k). \end{aligned}$$

If  $h_i$ ,  $i \in [N]$ , is block strongly convex, then  $\mathcal{G}^{k+1} := (\mathcal{G}_1^{k+1}, \dots, \mathcal{G}_N^{k+1}) \in \partial \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)$  and

$$\|\mathcal{G}^{k+1}\| \leq \bar{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \hat{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})}, \quad (3.12)$$

with

$$\begin{aligned}\bar{c} &:= \max \left\{ \sqrt{2/\sigma_1}, \dots, \sqrt{2/\sigma_N} \right\} \left( N(\bar{L} + \max \{ \delta_1 \bar{L}_1, \dots, \delta_N \bar{L}_N \}) + \max \left\{ \frac{\bar{L}_1}{\gamma_1} + \bar{L}_1, \dots, \frac{\bar{L}_N}{\gamma_N} + \bar{L}_N \right\} \right), \\ \widehat{c} &:= \max \left\{ \sqrt{2/\sigma_1}, \dots, \sqrt{2/\sigma_N} \right\} \max \left\{ \frac{\sigma_1(1-\gamma_1 \bar{L}_1)}{\gamma_1}, \dots, \frac{\sigma_N(1-\gamma_N \bar{L}_N)}{\gamma_N} \right\}.\end{aligned}$$

*Proof.* Following [?, Chapter 10], the subdifferential of  $\mathcal{L}$  at  $(\mathbf{x}^{k+1}, \mathbf{x}^k)$  is given by

$$\partial \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k) = \left( \partial_{\mathbf{x}^{k+1}} \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k), \partial_{\mathbf{x}^k} \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k) \right), \quad (3.13)$$

where, for  $j \in [N]$ , by applying [?, Exercise 8.8] we have

$$\partial_{\mathbf{x}_j^{k+1}} \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k) = \nabla_j f(\mathbf{x}^{k+1}) + \partial g_j(\mathbf{x}_j^{k+1}) + \sum_{i=j}^N \delta_i (\nabla_j h_i(\mathbf{x}^{k,i}) - \nabla_j h_i(\mathbf{x}^{k,i-1})); \quad (3.14)$$

$$\partial_{\mathbf{x}_j^k} \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k) = \sum_{i=1}^{j-1} \delta_i (\nabla_j h_i(\mathbf{x}^{k,i}) - \nabla_j h_i(\mathbf{x}^{k,i-1})) - \nabla_{jj}^2 h_j(\mathbf{x}^{k,j-1})(\mathbf{x}_j^{k+1} - \mathbf{x}_j^k). \quad (3.15)$$

Writing the first-order optimality conditions for the subproblem (3.3) implies that there exists a subgradient  $\eta_j^{k+1} \in \partial g_j(\mathbf{x}_j^{k+1})$  such that

$$\nabla_j f(\mathbf{x}^{k,j-1}) - \frac{\alpha_j^k}{\gamma_j^k} (\mathbf{x}_j^k - \mathbf{x}_j^{k-1}) + \frac{1}{\gamma_j^k} (\nabla_j h_j(\mathbf{x}^{k,j}) - \nabla_j h_j(\mathbf{x}^{k,j-1})) + \eta_j^{k+1} = 0 \quad j \in [N],$$

which implies  $\eta_j^{k+1} = \frac{1}{\gamma_j^k} (\nabla_j h_j(\mathbf{x}^{k,j-1}) - \nabla_j h_j(\mathbf{x}^{k,j})) + \frac{\alpha_j^k}{\gamma_j^k} (\mathbf{x}_j^k - \mathbf{x}_j^{k-1}) - \nabla_j f(\mathbf{x}^{k,j-1})$ ,  $j \in [N]$ .

Therefore, we have  $\mathcal{V}_j^{k+1} = \nabla_j f(\mathbf{x}^{k+1}) + \eta_j^{k+1} + \sum_{i=j}^N \delta_i (\nabla_j h_i(\mathbf{x}^{k,i}) - \nabla_j h_i(\mathbf{x}^{k,i-1})) \in \partial_{\mathbf{x}_j^{k+1}} \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)$ , which implies  $\mathcal{G}^{k+1} \in \partial \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)$ . Together with the Lipschitz continuity of  $\nabla_j f$ ,  $\nabla_j h_i$  and the boundedness of  $\nabla_{ii}^2 h_i$  on bounded sets, the boundedness of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$ , and the triangle inequality, this implies that there exist constants  $\widehat{L}$ ,  $\widehat{L}_i$ ,  $\bar{L}_i > 0$  (for  $i \in [N]$ ) such that

$$\begin{aligned}\|\mathcal{G}_j^{k+1}\| &= \|\mathcal{V}_j^{k+1}\| + \|\mathcal{W}_j^{k+1}\| \leq \frac{\alpha_j^k}{\gamma_j^k} \|\mathbf{x}_j^k - \mathbf{x}_j^{k-1}\| + \|\nabla_j f(\mathbf{x}^{k+1}) - \nabla_j f(\mathbf{x}^{k,j-1})\| \\ &\quad + \sum_{i=1}^N \delta_i \|\nabla_j h_i(\mathbf{x}^{k,i}) - \nabla_j h_i(\mathbf{x}^{k,i-1})\| + \frac{1}{\gamma_j^k} \|\nabla_j h_j(\mathbf{x}^k) - \nabla_j h_j(\mathbf{x}^{k,1})\| + \|\nabla_{jj}^2 h_j(\mathbf{x}^{k,j-1})\| \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^k\| \\ &\leq \frac{\alpha_j^k}{\gamma_j^k} \|\mathbf{x}_j^k - \mathbf{x}_j^{k-1}\| + \widehat{L} \sum_{i=1}^N \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| + \sum_{i=1}^N \delta_i \bar{L}_i \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| + \left( \frac{\bar{L}_j}{\gamma_j^k} + \bar{L}_j \right) \|\mathbf{x}_j^{k+1} - \mathbf{x}_j^k\|.\end{aligned}$$

Combining the last two inequalities with (3.9), it can be deduced that

$$\begin{aligned}\|\mathcal{G}^{k+1}\| &\leq \left( N(\bar{L} + \max \{ \delta_1 \bar{L}_1, \dots, \delta_N \bar{L}_N \}) + \max \left\{ \frac{\bar{L}_1}{\gamma_1} + \bar{L}_1, \dots, \frac{\bar{L}_N}{\gamma_N} + \bar{L}_N \right\} \right) \sum_{i=1}^N \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| \\ &\quad + \max \left\{ \frac{\alpha_1^k}{\gamma_1^k}, \dots, \frac{\alpha_N^k}{\gamma_N^k} \right\} \sum_{i=1}^N \|\mathbf{x}_i^k - \mathbf{x}_i^{k-1}\| \\ &\leq \left( N(\bar{L} + \max \{ \delta_1 \bar{L}_1, \dots, \delta_N \bar{L}_N \}) + \max \left\{ \frac{\bar{L}_1}{\gamma_1} + \bar{L}_1, \dots, \frac{\bar{L}_N}{\gamma_N} + \bar{L}_N \right\} \right) \sum_{i=1}^N \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| \\ &\quad + \max \left\{ \frac{\sigma_1(1-\gamma_1 \bar{L}_1)}{\gamma_1}, \dots, \frac{\sigma_N(1-\gamma_N \bar{L}_N)}{\gamma_N} \right\} \sum_{i=1}^N \|\mathbf{x}_i^k - \mathbf{x}_i^{k-1}\| \\ &\leq \bar{c} \sum_{i=1}^N \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| + \widehat{c} \sum_{i=1}^N \|\mathbf{x}_i^k - \mathbf{x}_i^{k-1}\|.\end{aligned}$$

Hence, it follows from the block strong convexity of  $h_i$  ( $i \in [N]$ ) that

$$\begin{aligned}\|\mathcal{G}^{k+1}\| &\leq \bar{c} \sum_{i=1}^N \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| + \widehat{c} \sum_{i=1}^N \|\mathbf{x}_i^k - \mathbf{x}_i^{k-1}\| \\ &\leq \bar{c} \sum_{i=1}^N \sqrt{\mathbf{D}_h(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \widehat{c} \sum_{i=1}^N \sqrt{\mathbf{D}_h(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})},\end{aligned}$$

giving our desired result.  $\square$

**Remark 3.8.** Note that a uniformly continuous function maps bounded sets to bounded sets. Therefore, in Proposition 3.7, if the function  $\nabla_{ii}^2 h_i$  ( $i \in [N]$ ) is uniformly continuous, it is bounded on bounded sets.  $\square$

Applying Proposition 3.7, the *subsequential convergence* of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  generated by BIBPA is presented next. On top of that we explain some basic properties of  $\omega(\mathbf{x}^k)$ .

**Assumption III.**  $C \subseteq \text{int dom } h_1$ .

**Theorem 3.9** (subsequential convergence and properties of  $\omega(\mathbf{x}^k)$ ). *Let all assumptions of Proposition 3.7 and Assumption III hold. Then, the following assertions are satisfied:*

- (i) every cluster point of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  is a critical point of  $\Phi$ , i.e.,  $\omega(\mathbf{x}^k) \subset \text{crit } \Phi$ ;
- (ii)  $\lim_{k \rightarrow \infty} \text{dist}(\mathbf{x}^k, \omega(\mathbf{x}^k)) = 0$ ;
- (iii)  $\omega(\mathbf{x}^k)$  is a nonempty, compact, and connected set;
- (iv) the Lyapunov function  $\mathcal{L}$  is finite and constant on  $\omega(\mathbf{x}^k)$ .

*Proof.* Let us assume  $\mathbf{x}^* = (x_1^*, \dots, x_N^*) \in \omega(\mathbf{x}^k)$ . The boundedness of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  implies that there exists an infinite index set  $\mathcal{J} \subset \mathbb{N}$  such that the subsequence  $(\mathbf{x}^k)_{k \in \mathcal{J}} \rightarrow \mathbf{x}^*$  as  $k \rightarrow \infty$ . It follows from (3.5) that

$$\begin{aligned} & \langle \nabla_i f(\mathbf{x}^{k,i-1}) - \frac{\alpha_i^k}{\gamma_i^k} (x_i^k - x_i^{k-1}), x_i^{k+1} - x_i^k \rangle + \frac{1}{\gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + g_i(x_i^{k+1}) \\ & \leq \langle \nabla_i f^k(\mathbf{x}^{k,i-1}) - \frac{\alpha_i^k}{\gamma_i^k} (x_i^k - x_i^{k-1}), x_i^* - x_i^k \rangle + \frac{1}{\gamma_i^k} \mathbf{D}_{h_i}(\mathbf{x}^*, \mathbf{x}^{k,i-1}) + g_i(x_i^*). \end{aligned} \quad (3.16)$$

Invoking Proposition 3.5 and using block strong convexity of  $h_i$ , there exist  $\varepsilon_i^* > 0$ ,  $k_i^0 \in \mathbb{N}$ , and a neighborhood  $\mathbf{B}(x_i^*, \varepsilon_i^*)$  such that  $\lim_{k \rightarrow \infty} \frac{\alpha_i^k}{2} \|x_i^{k+1} - x_i^k\|^2 \leq \lim_{k \rightarrow \infty} \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) = 0$ ,  $x_i^k \in \mathbf{B}(x_i^*, \varepsilon_i^*)$ ,  $i \in [N]$ , for  $k \geq k_i^0$  and  $k \in \mathcal{J}$ , i.e.,  $\lim_{k \rightarrow \infty} (x_i^{k+1} - x_i^k) = 0$ . Hence, substituting  $k = k_j - 1$  for  $k_j \in \mathcal{J}$  into (3.16) and taking the limit from both sides of (3.16), we derive  $\limsup_{j \rightarrow \infty} g_i(x_i^{k_j}) \leq g_i(x_i^*)$ ,  $i \in [N]$ . Furthermore, since  $g_i$  is lsc, this yields that  $\lim_{j \rightarrow \infty} g_i(x_i^{k_j}) = g_i(x_i^*)$ , then

$$\lim_{j \rightarrow \infty} \mathcal{L}(\mathbf{x}^{k_j+1}, \mathbf{x}^{k_j}) = \lim_{j \rightarrow \infty} \left( f(x_1^{k_j}, \dots, x_N^{k_j}) + \sum_{i=1}^N g_i(x_i^{k_j}) + \sum_{i=1}^N \delta_i \mathbf{D}_{h_i}(\mathbf{x}^{k_j,i}, \mathbf{x}^{k_j,i-1}) \right) = \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*).$$

Hence, from (3.12) and Proposition 3.5, we obtain

$$\lim_{k \rightarrow +\infty} \|\mathcal{G}^{k+1}\| \leq \lim_{k \rightarrow +\infty} \left( \widehat{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \widehat{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} \right) = 0,$$

which consequently yields  $\lim_{k \rightarrow \infty} \mathcal{G}^{k+1} = 0$ . As a result, we have  $0 \in \partial \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*)$ , owing to the closedness of the subdifferential mapping  $\partial \mathcal{L}$ . The result of Theorem 3.9(i) follows from the fact  $\partial \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*) = (\partial \Phi(\mathbf{x}^*), 0)$ . Moreover, Theorem 3.9(ii) is a straightforward consequence of Theorem 3.9(i), and Theorem 3.9(iii) and Theorem 3.9(iv) can be proved in the same way as [?, Lemma 5(iii)-(iv)].  $\square$

**Remark 3.10** (stopping criterion for BIBPA (Alg. 1)). One natural way to stop BIBPA is to make sure that either  $\text{dist}(0, \partial \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*)) = \text{dist}(0, \partial \Phi(\mathbf{x}^*)) \leq \|\mathcal{G}^{k+1}\| \leq \varepsilon$  or the *relative error*  $\sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} \leq \varepsilon$  holds true, for  $\varepsilon > 0$ ; see [?, H2].  $\square$

### 3.1 Global convergence for KL functions

In this section, we consider the class of Kurdyka-Łojasiewicz (KL) functions (see [?,?]) and show that for such functions the sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  converges to a critical point  $\mathbf{x}^*$ .

**Definition 3.11** (KL property). *A proper and lsc function  $\varphi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  has the KL property at  $\mathbf{x}^* \in \text{dom } \varphi$  if there exist a concave function  $\psi : [0, \eta] \rightarrow [0, +\infty[$  (with  $\eta > 0$ ) and neighborhood  $\mathbf{B}(\mathbf{x}^*; \varepsilon)$  with  $\varepsilon > 0$ , such that (i)  $\psi(0) = 0$ ; (ii)  $\psi$  is of class  $C^1$  with  $\psi > 0$  on  $(0, \eta)$ ; (ii) for all  $\mathbf{x} \in \mathbf{B}(\mathbf{x}^*; \varepsilon)$  such that  $\varphi(\mathbf{x}^*) < \varphi(\mathbf{x}) < \varphi(\mathbf{x}^*) + \eta$  it holds that*

$$\psi'(\varphi(\mathbf{x}) - \varphi(\mathbf{x}^*)) \mathbf{dist}(0, \partial\varphi(\mathbf{x})) \geq 1. \quad (3.17)$$

If this property holds for each point of  $\text{dom } \partial\varphi$ , the  $\varphi$  is a KL function.

In [?,?], Stanisław Łojasiewicz showed for the first time that every real analytic function<sup>1</sup> satisfies (3.17) with  $\psi(s) := \frac{\kappa}{1-\theta} s^{1-\theta}$  with  $\theta \in [0, 1)$ . In 1998, Kurdyka [?] proved that this inequality is valid for  $C^1$  functions whose graph belong to an *o-minimal structure* (see its definition in [?]). Later, (3.17) was extended for nonsmooth functions in [?,?,?].

The KL property (3.17) of the underlying objective function plays a key role in establishing the global convergence of a generic algorithm for nonconvex problems; however, this is not sufficient and one also needs some additional conditions to be guaranteed by the algorithm (see below). In particular, for several algorithms the cost functions satisfy the sufficient decrease condition (cf. [?,?,?]), while for some others the sufficient decrease condition is satisfied for some Lyapunov functions (cf. [?,?,?,?]).

As shown in Proposition 3.5, Proposition 3.7, and Theorem 3.9 (see its proof), the sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  generated by BIBPA satisfies the following conditions that are non-Euclidean extension of those given in [?,?] for the structured problem (2.1):

- 1) (*sufficient descent condition*) For each  $k \in \mathbb{N}$  and  $a_i, b_i \geq 0$  ( $i \in [N]$ ),

$$\sum_{i=1}^N (a_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + b_i \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})) \leq \mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}) - \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k);$$

- 2) (*relative error condition*) For each  $k \in \mathbb{N}$ , there exists a subgradient  $\mathcal{G}^{k+1} \in \partial\mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)$  and  $\bar{c}, \bar{d} \geq 0$  such that

$$\|\mathcal{G}^{k+1}\| \leq \bar{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \bar{d} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})};$$

- 3) (*continuity condition*) The function  $\mathcal{L}$  is a KL function, and each cluster point  $\mathbf{x}^*$  of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  ( $\mathbf{x}^* \in \omega(\mathbf{x}^k)$ ) satisfies  $(\mathbf{x}^*, \mathbf{x}^*) \in \text{crit } \mathcal{L}$

We now use the above three conditions to prove that the whole sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  converges.

**Theorem 3.12** (global convergence). *Let all assumptions of Proposition 3.7 and Assumption III hold. If  $\mathcal{L}$  is a KL function, then the following statements are true:*

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

$$\sum_{k=1}^{\infty} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| < \infty \quad i \in [N]; \quad (3.18)$$

<sup>1</sup> A function  $\varphi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  said to be real analytic if it can be represented by a convergent power series.

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

*Proof.* Define the sequence  $(d_k)_{k \in \mathbb{N}}$  as  $d_k := \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})}$ . From [Proposition 3.7](#) for  $\bar{c} := \max\{\bar{c}, \widehat{c}\}$ , we obtain

$$\|\mathcal{G}^{k+1}\| \leq \bar{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \widehat{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} \leq \bar{c} d_k. \quad (3.19)$$

Applying twice the root-mean square and arithmetic mean inequality we come to

$$\begin{aligned} d_k &\leq \sqrt{N \sum_{i=1}^N \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \sqrt{N \sum_{i=1}^N \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} \\ &\leq \sqrt{2N \sum_{i=1}^N (\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}))}. \end{aligned} \quad (3.20)$$

Then, it can be concluded from [Proposition 3.5](#) and (3.20) that

$$\mathcal{L}^k - \mathcal{L}^{k+1} \geq \sum_{i=1}^N (a_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + b_i \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})) \geq \frac{\varrho}{2N} d_k^2,$$

where  $\varrho := \min\{a_1, b_1, \dots, a_N, b_N\}$ . Together with (3.19) and [Theorem 3.9\(i\)](#), this implies that [[?](#), Assumption H] holds true with  $a_k = \frac{\varrho}{2N}$ ,  $b_k = 1$ ,  $b = \bar{c}$ ,  $I = \{1\}$ ,  $\varepsilon_k = 0$ . Therefore, since  $\mathcal{L}$  is a proper lower semicontinuous KL function, [[?](#), Theorem 10] yields that [Theorem 3.12\(i\)](#) holds true and the sequence  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  converges to  $\mathbf{x}^*$  in which  $(\mathbf{x}^*, \mathbf{x}^*)$  is a stationary point of the Lyapunov function  $\mathcal{L}$  (3.8), i.e.,  $0 \in \partial \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*)$ . Finally, the result follows from the fact  $\partial \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*) = (\partial \Phi(\mathbf{x}^*), 0)$ .  $\square$

### 3.2 Rate of convergence for Łojasiewicz-type KL functions

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

$$|\varphi(\mathbf{x}) - \varphi^*|^\theta \leq \kappa \mathbf{dist}(0, \partial \varphi(\mathbf{x})) \quad \forall \mathbf{x} \in \mathbf{B}(\mathbf{x}^*; \varepsilon). \quad (3.21)$$

**Fact 3.13** (convergence rate of a sequence with positive elements). [[?](#), 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  $0 \leq s_k \leq \lambda \tau^k$  for every  $k \in \mathbb{N}$ .
- (iii) If  $\theta \in (1/2, 1)$ , there exists  $\mu > 0$  such that  $0 \leq s_k \leq \mu k^{-\frac{1}{2\theta-1}}$  for every  $k \in \mathbb{N}$ .

Let  $(\mathcal{S}_k)_{k \in \mathbb{N}}$  given by  $\mathcal{S}_k := \mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}) - \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*)$ . We next derive the convergence rates of  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  and  $(\mathcal{S}_k)_{k \in \mathbb{N}}$  when  $\mathcal{L}$  satisfies the KL inequality of Łojasiewicz type.

**Theorem 3.14** (convergence rate). *Let all assumptions of [Proposition 3.7](#) and [Assumption III](#) hold, and  $(\mathbf{x}^k)_{k \in \mathbb{N}}$  converges to  $\mathbf{x}^*$ . If  $\mathcal{L}$  satisfies the KL inequality of Łojasiewicz type, then the following assertions hold:*

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

$$0 \leq \|\mathbf{x}^k - \mathbf{x}^*\| \leq \lambda_1 \tau^k, \quad 0 \leq \mathcal{S}_k \leq \mu_1 \bar{\tau}^k \quad \forall k \geq \bar{k};$$

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

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

*Proof.* We first set  $\varepsilon > 0$  to be that a constant described in (3.21) and  $\mathbf{x}^k \in \mathbf{B}(\mathbf{x}^*; \varepsilon)$  for all  $k \geq \bar{k}$  and  $\bar{k} \in \mathbb{N}$ . Let us define  $\Delta_k := \psi(\mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}) - \mathcal{L}(\mathbf{x}^*, \mathbf{x}^*)) = \psi(\mathcal{S}_k)$ . Then, it follows from the concavity of  $\psi$  and 2) that

$$\begin{aligned} \Delta_k - \Delta_{k+1} &= \psi(\mathcal{S}_k) - \psi(\mathcal{S}_{k+1}) \geq \psi'(\mathcal{S}_k)(\mathcal{S}_k - \mathcal{S}_{k+1}) \\ &= \psi'(\mathcal{S}_k)(\mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}) - \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)) \geq \frac{\mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}) - \mathcal{L}(\mathbf{x}^{k+1}, \mathbf{x}^k)}{\text{dist}(0, \partial \mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}))} \\ &\geq \frac{\sum_{i=1}^N (a_i \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + b_i \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}))}{\widehat{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \widehat{c} \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})}} \\ &\geq \frac{1}{c} \frac{\sum_{i=1}^N (\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}))}{\sum_{i=1}^N (\sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})})}, \end{aligned}$$

with  $c := \max\{\widehat{c}, \widehat{c}\} / \min\{a_1, b_1, \dots, a_N, b_N\}$ . Using (3.20) and applying the arithmetic mean and geometric mean inequality, it can be concluded that

$$\begin{aligned} d_k &\leq \sqrt{2N \sum_{i=1}^N (\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}))} \\ &\leq \sqrt{2cN(\Delta_k - \Delta_{k+1}) \sum_{i=1}^N (\sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})})} \quad (3.22) \\ &\leq cN(\Delta_k - \Delta_{k+1}) + \frac{1}{2} \sum_{i=1}^N (\sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})}) \end{aligned}$$

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

$$p_{k+1} := \sum_{i=1}^N \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})}, \quad q_k = cN(\Delta_k - \Delta_{k+1}), \quad \alpha := \frac{1}{2}, \quad (3.23)$$

where  $\sum_{i=1}^{\infty} q_k = 2cN \sum_{i=1}^{\infty} (\Delta_i - \Delta_{i+1}) = \Delta_1 - \Delta_{\infty} = \Delta_1 < \infty$ . This and (3.22) yield  $p_{k+1} \leq \frac{1}{2} p_k + q_k$  for all  $k \geq \bar{k}$ . Since  $(\Phi)_{k \in \mathbb{N}}$  is non-increasing,

$$\sum_{j=k}^{\infty} p_{j+1} \leq \frac{1}{2} \sum_{j=k}^{\infty} (p_j - p_{j+1} + p_{j+1}) + 2cN \sum_{j=k}^{\infty} (\Delta_j - \Delta_{j+1}) = \frac{1}{2} \sum_{j=k}^{\infty} p_{j+1} + \frac{1}{2} p_k + 2cN \Delta_k.$$

From the root-mean square, the arithmetic mean inequality,  $\psi(\mathcal{S}_k) \leq \psi(\mathcal{S}_{k-1})$ , and Proposition 3.5, this lead to

$$\begin{aligned} \sum_{j=k}^{\infty} p_{j+1} &\leq p_k + 4cN \Delta_k = \sum_{i=1}^N (\sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})}) + 4cN \psi(\mathcal{S}_k) \\ &\leq \sqrt{N \sum_{i=1}^N \mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1})} + \sqrt{N \sum_{i=1}^N \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + 4cN \psi(\mathcal{S}_k) \\ &\leq \sqrt{2N \sum_{i=1}^N (\mathbf{D}_{h_i}(\mathbf{x}^{k,i}, \mathbf{x}^{k,i-1}) + \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}))} + 4cN \psi(\mathcal{S}_k) \\ &\leq \sqrt{2N/\varrho} \sqrt{\mathcal{S}_{k-1} - \mathcal{S}_k} + 4cN \psi(\mathcal{S}_{k-1}), \end{aligned} \quad (3.24)$$

with  $\varrho := \min\{a_1, b_1, \dots, a_N, b_N\}$ . Since  $\mathbf{D}_{h_i}(\cdot, \cdot) \geq 0$ , for  $i \in [N]$ , it holds that

$$\begin{aligned} \|x_i^k - x_i^*\| &\leq \|x_i^{k+1} - x_i^k\| + \|x_i^{k+1} - x_i^*\| \leq \dots \leq \sum_{j=k}^{\infty} \|x_i^{j+1} - x_i^j\| \\ &\leq \sum_{j=k}^{\infty} \sqrt{\frac{2}{\sigma_i} \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} \leq \sqrt{\frac{2}{\sigma_i} \sum_{j=k}^{\infty} (\sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})})}. \end{aligned}$$

Combining this with (3.24) and setting  $\rho := \max\{\sqrt{2/\sigma_1}, \dots, \sqrt{2/\sigma_N}\}$ , we come to

$$\begin{aligned} \sum_{i=1}^N \|x_i^k - x_i^*\| &\leq \rho \sum_{j=k}^{\infty} \sum_{i=1}^N \left( \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})} \right) \\ &\leq \rho \sqrt{2N/\varrho} \sqrt{\mathcal{S}_{k-1} - \mathcal{S}_k} + 4c\rho N\psi(\mathcal{S}_{k-1}), \end{aligned}$$

which consequently yields

$$\|x_i^k - x_i^*\| \leq \nu \max\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} \quad i \in [N], \quad (3.25)$$

with  $\nu := \rho \sqrt{2N/\varrho} + 4c\rho N$  and  $\psi(s) := \frac{\kappa}{1-\theta} s^{1-\theta}$ . Furthermore, the nonlinear equation  $\sqrt{\mathcal{S}_{k-1}} - \frac{\kappa}{1-\theta} \mathcal{S}_{k-1}^{1-\theta} = 0$  has a solution at  $\mathcal{S}_{k-1} = ((1-\theta)/\kappa)^{\frac{2}{1-2\theta}}$ . For  $\hat{k} \in \mathbb{N}$  and  $k \geq \hat{k}$ , we assume (3.25) holds and  $\mathcal{S}_{k-1} \leq \left(\frac{\kappa}{1-\theta}\right)^{\frac{2}{1-2\theta}}$ . Two cases are recognized: (a)  $\theta \in (0, 1/2]$ ; (b)  $\theta \in (1/2, 1)$ . In Case (a), if  $\theta \in (0, 1/2)$ , then  $\psi(\mathcal{S}_{k-1}) \leq \sqrt{\mathcal{S}_{k-1}}$ . For  $\theta = 1/2$ , we get  $\psi(\mathcal{S}_{k-1}) = \frac{\kappa}{1-\theta} \sqrt{\mathcal{S}_{k-1}}$ , which implies  $\max\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} = \max\{1, \frac{\kappa}{1-\theta}\} \sqrt{\mathcal{S}_{k-1}}$ . Then,  $\max\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} \leq \max\{1, \frac{\kappa}{1-\theta}\} \sqrt{\mathcal{S}_{k-1}}$ . In Case (b), it holds that  $\psi(\mathcal{S}_{k-1}) \geq \sqrt{\mathcal{S}_{k-1}}$ , i.e.,  $\max\{\sqrt{\mathcal{S}_{k-1}}, \psi(\mathcal{S}_{k-1})\} = \frac{\kappa}{1-\theta} \mathcal{S}_{k-1}^{1-\theta}$ . Combining both cases, for all  $k \geq \bar{k} := \max\{\bar{k}, \hat{k}\}$ , we end up with

$$\|x_i^k - x_i^*\| \leq \begin{cases} \nu \max\{1, \frac{\kappa}{1-\theta}\} \sqrt{\mathcal{S}_{k-1}} & \text{if } \theta \in (0, 1/2], \\ \nu \frac{\kappa}{1-\theta} \mathcal{S}_{k-1}^{1-\theta} & \text{if } \theta \in (1/2, 1). \end{cases} \quad (3.26)$$

On the other hand, it follows from Proposition 3.5 that

$$\begin{aligned} \mathcal{S}_{k-1} - \mathcal{S}_k &= \mathcal{L}(\mathbf{x}^{k-1}, \mathbf{x}^{k-2}) - \mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}) \geq \varrho \sum_{i=1}^N \left( \mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1}) + \mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1}) \right) \\ &\geq \frac{\varrho}{2N} \left( \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-1,i}, \mathbf{x}^{k-1,i-1})} + \sqrt{\mathbf{D}_{h_i}(\mathbf{x}^{k-2,i}, \mathbf{x}^{k-2,i-1})} \right)^2 \\ &\geq \frac{\varrho}{2Nc^2} \|(\mathcal{G}_1^k, \dots, \mathcal{G}_N^k)\|^2 \geq \frac{\varrho}{2Nc^2} \text{dist}(0, \partial \mathcal{L}(\mathbf{x}^k, \mathbf{x}^{k-1}))^2 \geq \frac{\varrho}{2Nc^2\kappa^2} \mathcal{S}_{k-1}^\theta = c_2 \mathcal{S}_{k-1}^\theta, \end{aligned}$$

where  $c_2 := \frac{\varrho}{2Nc^2\kappa^2}$ . The results then follow from  $\mathcal{S}_k \rightarrow 0$ , (3.26) and Fact 3.13.  $\square$

#### 4 Application to symmetric nonnegative matrix tri-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., [?, ?] 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) other constraints such as sparsity to take prior information into account. More precisely, let the data matrix be  $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}_+^{m \times n}$  where each  $x_i$  represents some data point. NMF seeks a decomposition of  $X$  into a nonnegative  $n \times r$  basis matrix  $U = [u_1, u_2, \dots, u_r] \in \mathbb{R}_+^{n \times r}$  and a nonnegative  $r \times m$  coefficient matrix  $V = [v_1, v_2, \dots, v_r]^T \in \mathbb{R}_+^{r \times m}$  such that

$$X \approx UV, \quad (4.1)$$

where  $\mathbb{R}_+^{m \times n}$  is the set of  $m \times n$  nonnegative matrices. Extensive research has been carried out on variants of NMF, and most studies have focused on algorithmic developments, but with very limited convergence theory. This motivates us to study the application of BIBPA

to a variant of NMF, namely SymTriNMF; see (2.6) for the formulation of SymTriNMF as an optimization problem.

One popular application of SymTriNMF is community detection. Let  $X$  be the adjacency matrix of graph so that  $X_{ij} = 1$  if item  $i$  is connected to item  $j$ , and  $X_{ij} = 0$  otherwise. Let also  $X \approx UVU^T$  be a SymTriNMF decomposition of  $X$ . Each column of  $U$  corresponds to a community, that is, to a subset of items highly connected. In other words, the entry  $U_{jk}$  of  $U$  indicates the membership of item  $j$  within community  $k$ , and  $U_{jk} > 0$  if  $j$  belongs to community  $k$ . The  $r$ -by- $r$  matrix  $V$  indicates the relationship between communities, that is, whether the items within two communities are likely to interact:  $V_{kp}$  is the "strength" of the interaction between the  $k$ th and  $p$ th communities. We have  $X \approx \sum_{k=1}^r \sum_{p=1}^r U_{:k} V_{k,p} U_{:p}^T$ , so that  $X$  is decomposed via the sum of  $r^2$  rank-one factors corresponding to the  $r$  communities and their interactions; see [?,?] for more details. Note that SymTriNMF is closely related to the mixed membership stochastic blockmodel [?].

Given  $U^k$  and  $V^k$ , we next derive the closed-form solutions for  $U^{k+1}$  and  $V^{k+1}$ .

**Theorem 4.1** (closed-form solutions of the subproblem (3.5) for SymTriNMF). *Let  $h_1$  and  $h_2$  be the kernel functions given in (2.7) and (2.8) and  $U^k$  and  $V^k$  are given. Then,*

(i) *the iteration  $U^{k+1}$  of the subproblem (3.5) is given by*

$$U^{k+1} = \frac{1}{t_k} \mathbf{max} \left\{ \frac{1}{\gamma_1^k} (\nabla_U h_1(U^k, V^k) - \gamma_1^k \nabla_U f(U^k, V^k) + \alpha_1^k (U^k - U^{k-1})), 0 \right\} \quad (4.2)$$

with

$$\begin{aligned} \nabla_U f(U^k, V^k) &= -XU^k(V^k)^T - X^T U^k V^k + U^k V^k (U^k)^T U^k (V^k)^T + U^k (V^k)^T (U^k)^T U^k V^k, \\ \nabla_U h_1(U^k, V^k) &= (a_1 \|U^k\|_F^2 \|V^k\|_F^2 + b_1 (\|X\|_F \|V^k\|_F + \varepsilon_1)) U^k, \end{aligned}$$

and

$$t_k = \frac{\tau_1}{3} + \sqrt[3]{\frac{\tau_2 + \sqrt{A_1}}{2} + \frac{\tau_1^3}{27}} + \sqrt[3]{\frac{\tau_2 - \sqrt{A_1}}{2} + \frac{\tau_1^3}{27}}, \quad (4.3)$$

where  $\tau_1 = b_1 (\|X\|_F \|V^k\|_F + \varepsilon_1)$ ,  $\tau_2 = a_1 \|V^k\|_F^2 \mathbf{max}\{G^k, 0\} \|_F^2$ ,  $A_1 = \tau_2^2 + \frac{4}{27} \tau_2^2 \tau_1^3$  with  $G^k := \frac{1}{\gamma_1^k} (\nabla_U h_1(U^k, V^k) - \gamma_1^k \nabla_U f(U^k, V^k) + \alpha_1^k (U^k - U^{k-1}))$ .

(ii) *for  $\eta_k := a_2 \|U^{k+1}\|^4 + \varepsilon_2$ , the iteration  $V^{k+1}$  of the subproblem (3.5) is given by*

$$V^{k+1} = \mathbf{max} \left\{ V^k - \frac{1}{\eta_k} (\alpha_2^k (V^k - V^{k-1}) - \gamma_2^k \nabla_V f(U^{k+1}, V^k)), 0 \right\}, \quad (4.4)$$

with  $\nabla_V f(U^{k+1}, V^k) = (U^{k+1})^T X U^{k+1} + (U^{k+1})^T U^{k+1} V^k (U^{k+1})^T U^{k+1}$ .

*Proof.* Setting  $g_1 := \delta_{U \geq 0}$  and  $f(U, V) = \frac{1}{2} \|X - UVU^T\|_F^2$ , it follows from (3.5) that

$$\begin{aligned} U^{k+1} &= \mathbf{arg min}_{U \in \mathbb{R}^{m \times r}} \left\{ \langle \nabla_U f(U^k, V^k) - \frac{\alpha_1^k}{\gamma_1^k} (U^k - U^{k-1}), U - U^k \rangle \right. \\ &\quad \left. + \frac{1}{\gamma_1^k} \mathbf{D}_{h_1}((U, V^k), (U^k, V^k)) + g_1(U) \right\} \\ &= \mathbf{arg min}_{U \geq 0} \left\{ \frac{1}{\gamma_1^k} \langle \gamma_1^k \nabla_U f(U^k, V^k) - \nabla_U h_1(U^k, V^k) - \alpha_1^k (U^k - U^{k-1}), U \rangle + \frac{1}{\gamma_1^k} h_1(U, V^k) \right\}. \end{aligned} \quad (4.5)$$



By [?, Corollary 3.5], the normal cone of the nonnegativity constraint  $U \geq 0$  is  $\mathcal{N}_{U \geq 0}(U^k) = \{P \in \mathbb{R}^{m \times r} \mid U^k \odot P = 0, P \leq 0\}$  where  $U^k \odot P$  denotes the *Hadamard products* given pointwise by  $(U^k \odot P)_{ij} := U_{ij}^k P_{ij}$  for  $i \in 1, \dots, m$  and  $j \in 1, \dots, r$ . The first-order optimality conditions for the subproblem (4.5) yields that  $G^k - (a_1 \|U^{k+1}\|_F^2 \|V^k\|_F^2 + b_1 (\|X\|_F \|V^k\|_F + \varepsilon_1)) U^{k+1} \in \mathcal{N}_{U \geq 0}(U^{k+1})$ .

We now consider two cases: (i)  $G_{ij} \leq 0$ ; (ii)  $G_{ij} > 0$ . In Case (i), we have

$$P_{ij} = G_{ij}^k - (a_1 \|U^{k+1}\|_F^2 \|V^k\|_F^2 + b_1 (\|X\|_F \|V^k\|_F + \varepsilon_1)) U_{ij}^{k+1} \leq 0,$$

hence  $U_{ij}^{k+1} = 0$ . In Case (ii), if  $U_{ij}^{k+1} = 0$ , then  $P_{ij} = G_{ij}^k > 0$ , which contradicts  $P \leq 0$ ; hence  $G_{ij}^k - (a_1 \|U^{k+1}\|_F^2 \|V^k\|_F^2 + b_1 (\|X\|_F \|V^k\|_F + \varepsilon_1)) U_{ij}^{k+1} = 0$ . Combining both cases, we get  $(a_1 \|U^{k+1}\|_F^2 \|V^k\|_F^2 + b_1 (\|X\|_F \|V^k\|_F + \varepsilon_1)) U^{k+1} = \mathbf{Proj}_{G \geq 0}(G^k)$ . Denote  $t_k = a_1 \|U^{k+1}\|_F^2 \|V^k\|_F^2 + b_1 \|X\|_F \|V^k\|_F$ , then  $\|U^{k+1}\|_F^2 = (t_k - b_1 \|X\|_F \|V^k\|_F) / (a_1 \|V^k\|_F^2)$ . We have  $t_k^3 - b_1 \|X\|_F \|V^k\|_F t_k^2 - a_1 \|V^k\|_F^2 \mathbf{Proj}_{G \geq 0}(G^k) = 0$ . 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  $\Delta = c^2 + \frac{4}{27}ca^3$ . Then we get (4.3). Finally, the result follows from  $U^{k+1} = \frac{\mathbf{Proj}_{G \geq 0}(G^k)}{t_k}$ .

By setting  $g_2 := \delta_{V \geq 0}$  and invoking (3.5), we get

$$\begin{aligned} V^{k+1} &= \mathbf{arg\,min}_{V \in \mathbb{R}^{r \times r}} \left\{ \langle \nabla_V f(U^{k+1}, V^k) - \frac{\alpha_2^k}{\gamma_2^k} (V^k - V^{k-1}), V - V^k \rangle \right. \\ &\quad \left. + \frac{1}{\gamma_2^k} \mathbf{D}_{h_2}((U^{k+1}, V), (U^{k+1}, V^k)) + g_2(V) \right\} \\ &= \mathbf{arg\,min}_{V \geq 0} \left\{ \frac{1}{\gamma_2^k} (\gamma_2^k \nabla_V f(U^{k+1}, V^k) - \alpha_2^k (V^k - V^{k-1}) - \nabla h_2(U^{k+1}, V^k), V) + \frac{1}{\gamma_2^k} h_2(U^{k+1}, V) \right\} \\ &= \mathbf{arg\,min}_{V \geq 0} \left\{ \left\| V - \frac{1}{\alpha_2 \|U^{k+1}\|_F + \varepsilon_2} (\alpha_2^k (V^k - V^{k-1}) + \nabla h_2(U^{k+1}, V^k) - \gamma_2^k \nabla_V f(U^{k+1}, V^k)) \right\|_F^2 \right\} \\ &= \mathbf{Proj}_{V \geq 0} \left( V^k - \frac{1}{\eta_k} (\alpha_2^k (V^k - V^{k-1}) - \gamma_2^k \nabla_V f(U^{k+1}, V^k)) \right), \end{aligned}$$

which proves (4.4).  $\square$

## 5 Final remarks

The descent lemma is a key factor for analyzing the first-order methods in both Euclidean and non-Euclidean settings. Owing to the notion of block relative smoothness, it was shown that the descent lemma is still valid for each block of variables for structured nonsmooth nonconvex problems with non-Lipschitz gradients. Based on this development, **BIBPA** was introduced to deal with such problems, and it was shown to be globally convergent for KL functions and its convergence rate was also studied. Besides, it was shown that the objective of the symmetric nonnegative matrix tri-factorization (SymTriNMF) problem is block relatively smooth, and the corresponding subproblems can be solved in closed forms. To our knowledge, **BIBPA** is the first algorithm with rigorous theoretical guarantee of convergence for this problem. We emphasize that the main objective of this paper is to provide a theoretical and algorithmic framework that can handle block structured nonsmooth nonconvex problems under the block relative smoothness assumption. Hence, a comprehensive numerical experiments for such structured problems are postponed to a future work.

## 6 Acknowledgment

The authors are grateful of the associate editor and anonymous referees for their helpful comments and suggestions that improved the quality of the paper.