INERTIAL BLOCK MIRROR DESCENT METHOD FOR NON-CONVEX NON-SMOOTH OPTIMIZATION[*]

LE THI KHANH HIEN[†], NICOLAS GILLIS[†], AND PANAGIOTIS PATRINOS[‡]

Abstract. In this paper, we propose inertial versions of block coordinate descent methods for solving non-convex non-smooth composite optimization problems. We use the general framework of Bregman distance functions to compute the proximal maps. Our methods not only allow using two different extrapolation points to evaluate gradients and adding the inertial force, but also take advantage of randomly picking the block of variables to update. Moreover, our methods do not require a restarting step, and as such, it is not a monotonically decreasing method. To prove the convergence of the whole generated sequence to a critical point, we modify the convergence proof recipe of Bolte, Sabach and Teboulle (Proximal alternating linearized minimization for non-convex and non-smooth problems, Math. Prog. 146(1):459–494, 2014), and combine it with auxiliary functions. We deploy the proposed methods to solve non-negative matrix factorization (NMF) and show that they compete favourably with the state-of-the-art NMF algorithms.

Key words. Inertial acceleration, mirror descent, block coordinate descent method, randomization, non-convex optimization, nonnegative matrix factorization

1. Introduction.

In this paper, we consider the following non-smooth non-convex optimization problem

$$\text{minimize}_{x\in\mathbb{E}}\, F(x)\,, \quad \text{where} \;\; F(x) := f(x) + r(x), \tag{1.1}$$

and

- $\mathbb{E} = \mathbb{E}_1 \times \ldots \times \mathbb{E}_s$ with $\mathbb{E}_i$, $i = 1, \ldots, s$, being finite dimensional real linear spaces equipped with norm $\|\cdot\|_{(i)}$ and inner product $\langle\cdot,\cdot\rangle_{(i)}$,
- $f : \mathbb{E} \to \mathbb{R}$ is a continuous but possibly non-smooth non-convex function, and
- $r(x) = \sum_{i=1}^{s} r_i(x_i)$ with $r_i : \mathbb{E}_i \to \mathbb{R} \cup \{+\infty\}$ for $i = 1, \ldots, s$ being proper and lower semi-continuous functions.

Throughout the paper, we assume that F is bounded from below. Problem (1.1) covers many applications including compressed sensing with non-convex "norms" (see [4]), sparse dictionary learning (see [1, 49]), nonnegative matrix factorization (see [22]), and "l_p-norm" regularized sparse regression problems with $0 \le p < 1$ (see [11, 29]). The Gauss-Seidel iteration scheme, also known as block coordinate descent (BCD) method, is a standard approach to solve both convex and non-convex problems in the form of (1.1). Starting with a given initial point $x^{(0)}$, the method generates a sequence $\left\{x^{(k)}\right\}_{k\ge 0}$ by cyclically updating one block of variables at a time while fixing the values of the other blocks; and as such, it has a lower per-iteration cost than methods updating all blocks simultaneously. Based on how the blocks are updated, BCD methods can typically be classified into three categories:

1. Classical BCD methods update each block of variables as follows

$$x_i^{(k)} = \operatorname*{argmin}_{x_i\in\mathbb{E}_i} f_i^{(k)}(x_i) + r_i(x_i)\,,$$

where $f_i^{(k)}(x_i) := f\left(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k-1)}, \ldots, x_s^{(k-1)}\right)$; see for example [25, 26]. Under suitable convexity assumptions on the functions $f_i^{(k)}(x_i)$ for $i =$

[*] This work was supported by the Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO) under EOS Project no O005318F-RG47. NG acknowledges the support of the European Research Council (ERC starting grant no 679515). PP was supported by FWO projects: G086318N; G086518N, Research Council KUL: C14/18/068.

[†] Department of Mathematics and Operational Research, University of Mons, Belgium (thikhanhhien.le@umons.ac.be, nicolas.gillis@umons.ac.be).

[‡] Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Belgium (panos.patrinos@esat.kuleuven.be)

$1, \ldots, s$, the classical BCD method can be guaranteed to converge to a stationary point, see [45, 25, 48, 51]. However, it fails to converge for some non-convex problems; see for example [40].

2. Proximal BCD methods update each block of variables as follows

$$x_i^{(k)} = \operatorname*{argmin}_{x_i \in \mathbb{E}_i} f_i^{(k)}(x_i) + r_i(x_i) + \frac{1}{2\beta_i^{(k)}} \left\| x_i - x_i^{(k-1)} \right\|^2, \tag{1.2}$$

where $\beta_i^{(k)}$ is the stepsize; see for example [6, 25, 41, 48]. Coupling the classical BCD methods with a proximal term promotes stability and improves convergence properties, especially for non-smooth and non-convex problems. In [4], considering Problem (1.1) with $s = 2$, the authors established, for the first time in the non-convex and non-smooth setting, the convergence of the whole sequence $\left\{x^{(k)}\right\}_{k\geq 0}$ to a critical point of F. The Kurdyka-Łojasiewicz (KL) property of F (see [28, 13]) is the cornerstone assumption that is used in their analysis to obtain the convergence (see Section 3.2).

3. Proximal gradient BCD methods update each block of variables as follows

$$x_i^{(k)} = \operatorname*{argmin}_{x_i \in \mathbb{E}_i} \left\langle \nabla f_i^{(k)}\left(x_i^{(k-1)}\right), x_i - x_i^{(k-1)} \right\rangle + r_i(x_i) + \frac{1}{2\beta_i^{(k)}} \left\| x_i - x_i^{(k-1)} \right\|^2, \tag{1.3}$$

see for example [8, 14, 41, 47]. Proximal gradient BCD methods minimize a standard proximal linearization of the objective function, that is, they linearize f, which is assumed to be smooth, and take a proximal step on the non-smooth part r. Regarding the convergence of the whole sequence in the general non-convex non-smooth setting, [5] can be considered as the first work establishing the convergence for the proximal gradient method (which is also known as proximal-forward-backward algorithm) to solve (1.1) with $s = 1$. The authors in [14] later provided a self-contained convergence analysis framework for the proximal gradient BCD method applied to solve (1.1) with $s = 2$. Both works rely on the powerful KL property.

In the convex setting, incorporating inertial force is a popular and efficient method to accelerate the convergence of the gradient descent method whose rate is known to be suboptimal. The inertial term was first introduced by Polyak's heavy ball method (see [39]), which adds to the new direction a momentum computed by the difference of the two previous iterates. While calculating gradients used in Polyak's method is not affected by the momentum, the famous accelerated gradient method of Nesterov (see [30, 31, 32, 33]) evaluates the gradients at the points which are extrapolated by the momentum. In the convex setting, these methods are proved to achieve the optimal convergence rate, while the computational cost of each iteration is essentially unchanged.

In the non-convex setting, the heavy ball method was first considered in [52] to solve an unconstrained smooth minimization problem. Two inertial proximal gradient methods were proposed in [36] and [16] to solve (1.1) with $s = 1$. The method considered in [36], referred to as iPiano, makes use of the inertial force but does not use the extrapolated points to evaluate the gradients. The iPiano method was extended for $s > 1$ and analysed in [35]. The inertial forward-backward-forward method with a Bregman distance replacing the Euclidean distance in (1.3) is analyzed in [16]. In [38], the method called iPALM is proposed to solve (1.1) with $s = 2$. At each iteration, iPALM makes use of two different extrapolated points: one to compute the gradients and the other to add inertial force. In [48], the authors propose an inertial version for the proximal gradient BCD method (1.3) to solve (1.1) with the assumption that f is block-wise convex. The same authors, later, in [50], extend their method to solve general non-convex optimization problems. Furthermore, the proposed method in [50] allows the blocks to be chosen either deterministically or randomly as long as each block is updated at least once in every fixed number of iterations. They empirically show that randomly selecting the block of variables increases the chance to obtain better quality local solutions. However, their methods need restarting steps to guarantee the strict decrease of the objective function. This property is essential for their convergence analysis. As stated in [32], this relaxation property for some problem classes is too expensive and may not allow optimal convergence rates. It is important to note that using inertial terms will in general not guarantee the objective function to monotonically decrease in gradient descent methods.

In another line of works, it is worth mentioning the randomized block coordinate descent methods for solving convex problems; see for example [20, 34]. The methods randomly choose the update block according to some probability $p_i > 0$ with $\sum_{i=1}^s p_i = 1$. The analysis of this type of algorithms considers the convergence of the function values and iterates in expectation. This is out of the scope of this work.

1.1. Outline of the paper and contribution.

In this paper, we propose inertial versions for the proximal and proximal gradient BCD methods (1.2) and (1.3), for solving the non-convex non-smooth problem (1.1) with multiple blocks. Our method is put in the framework of the Bregman distance functions so that it is more general hence admits potentially more applications. For the inertial version of the proximal gradient BCD (1.3), two extrapolation points can be used to evaluate gradients and add the inertial force so that the corresponding scheme is more flexible and may lead to significantly better numerical performance compared with the inertial methods using a single extrapolation point; this will be confirmed in Section 5 with some numerical experiments. Moreover, our methods allow picking deterministically or randomly the block of variables to update and, as explained above, randomization may lead to better solutions and/or faster convergence. Finally, and this is a key aspect of our methods, they do not require restarting steps, and are not monotonically decreasing the objective function. To prove the convergence of the whole generated sequence to a critical point of F, we combine a modification of the convergence proof recipe established in [14] with the technique of using auxiliary functions in [36, 38].

The paper is organized as follows. In Section 2, we describe our proposed methods and explain some important notations. In Section 3, we give some preliminary results on non-convex non-smooth analysis, we also discuss Bregman distances and their proximal maps, which is a central tool in our methods. Section 4 presents the main convergence results: we first prove the subsequential convergence of the generated sequence to a critical point of F (Theorem 4.8) and then, with some additional assumptions, we prove the global convergence of the whole sequence (Theorem 4.13). Finally, we apply the methods to nonnegative matrix factorization (NMF) in Section 5, and show that they compete favourably with the state-of-the-art algorithms for NMF.

2. The proposed methods and notation.

Algorithm 2.1 describes the common framework for the two upcoming proposed methods, namely, an inertial block proximal method (IBP) and an inertial block proximal gradient method (IBPG). Algorithm 2.1 includes an outer loop which is indexed by k and an inner loop which is indexed by j. In each iteration of the inner loop, a single block of variables is updated using the update $\mathcal{A}$ (see below for more details). We use $x^{(k,j)}$ to denote the iterate at the jth iteration within the kth inner loop. At the jth iteration of an inner loop, we use y_i, $i = 1, \ldots, s$, to store the value of block i before it was updated to $x_i^{(k,j-1)}$; and block i is the only block to be updated from $x_i^{(k,j-1)}$ to $x_i^{(k,j)}$. We run T_k iterations within the kth inner loop.

We make the following assumption throughout this paper.

Assumption 2.1. *For all k, all blocks are updated after T_k iterations are performed within the kth inner loop, and there exists a positive constant $\bar{T}$ such that $s \le T_k \le \bar{T}$.*

Algorithm 2.1 Common framework of IBP and IBPG

```
**Initialize**: Choose initial points $\left(\tilde{x}_1^{(0)},\ldots,\tilde{x}_s^{(0)}\right)=\left(\tilde{x}_1^{(-1)},\ldots,\tilde{x}_s^{(-1)}\right)$.
**for** $k=1,\ldots$ **do**
    $x^{(k,0)}=\tilde{x}^{(k-1)}$.
    **for** $j=1,\ldots,T_k$ **do**
        Choose $i\in\{1,\ldots,s\}$ deterministically or randomly.
        Let $y_i$ be the value of the $i$th block before it was updated to $x_i^{(k,j-1)}$.
        Update $x_i^{(k,j)}=\mathcal{A}\left(x_i^{(k,j-1)},y_i\right)$, where the update $\mathcal{A}$ will be described in
        Algorithm 2.2 for IBP and in Algorithm 2.3 for IBPG.
        Let $x_{i'}^{(k,j)}=x_{i'}^{(k,j-1)}$ for $i'\neq i$.
    **end for**
    Update $\tilde{x}^{(k)}=x^{(k,T_k)}$.
**end for**
```

Assumption 2.1 is equivalent to the essentially cyclic block update used in [45, 50], where a constant $R\geq s$ is assumed to exist such that, for all r, every coordinate $i\in\{1,\ldots,s\}$ is updated at least once between the rth iteration and the $(r+R)$th iteration. In particular, our assumption satisfies the essentially cyclic block update with $R=2\bar{T}$; and conversely, we can choose the number of iterations T_k of the inner loop to be the value R in the assumption of the essentially cyclic block update. It is important to note that the use of an inner loop to describe our algorithms was chosen to significantly simplify the convergence analysis. In fact, as noted above, Algorithm 2.1 can be equivalently written with a single loop in which essentially cyclic rule is assumed as done in [45, 50].

To simplify the analysis, we introduce the notation $\bar{x}_i^{(k,m)}$ to denote the value of block i after it has been updated m times during the kth inner loop. For each inner loop k, we denote d_i^k the total number of times the ith block is updated. This means that there exists a subsequence $\{i_1,i_2,\ldots,i_{d_i^k}\}$ of $\{1,2,\ldots,T_k\}$ such that $\bar{x}_i^{(k,m)}=x_i^{(k,i_m)}$ for all $m=1,2,\ldots,d_i^k$. In particular, we have that $\bar{x}_i^{(k,0)}=\bar{x}_i^{(k-1,d_i^{k-1})}=\tilde{x}_i^{(k-1)}$ and $\bar{x}_i^{(k,d_i^k)}=\tilde{x}_i^{(k)}$. The previous value of block i before it is updated to $\bar{x}_i^{(k,m)}$ is $\bar{x}_i^{(k,m-1)}$, and, as for $\bar{x}_i^{(k,m)}$, we use the notation $\bar{x}_i^{(k,-1)}=\bar{x}_i^{(k-1,d_i^{k-1}-1)}$.

Since y_i is the value of block i before it was updated to $x_i^{(k,j-1)}$, we remark that y_i, $x_i^{(k,j-1)}$ and $x_i^{(k,j)}$ are three consecutive iterates of the sequence $\bar{x}_i^{(k,-1)},\ldots,\bar{x}_i^{(k,d_i^k)}$. For block i, we denote $\left\{\bar{x}_i^{(k,m)}\right\}_{k\geq 1}$ the sequence that contains the updates of the ith block, that is, $\left\{\bar{x}_i^{(k,m)}\right\}_{k\geq 1}=\left\{\bar{x}_i^{(1,1)},\ldots,\bar{x}_i^{(1,d_i^1)},\ldots,\bar{x}_i^{(k,1)},\ldots,\bar{x}_i^{(k,d_i^k)},\ldots\right\}$. We denote

$$f_i^{(k,j)}\left(x_i\right)=f\left(x_1^{(k,j-1)},\ldots,x_{i-1}^{(k,j-1)},x_i,x_{i+1}^{(k,j-1)},\ldots,x_s^{(k,j-1)}\right),$$

hence $f_i^{(k,j)}\left(x_i\right)$ is a function of the ith block while fixing the latest updated values of the other blocks. We let $F_i^{(k,j)}\left(x_i\right)=f_i^{(k,j)}\left(x_i\right)+r_i\left(x_i\right)$.

We now detail the update $\mathcal{A}$ of the blocks of variables for IBP and IBPG in Algorithm 2.2 and Algorithm 2.3 respectively. Algorithm 2.2 first computes an extrapolated point in (2.1), and then computes the next iterate as the proximal map (see (3.1)) at the extrapolated point. Algorithm 2.2 first computes two extrapolated points in (2.3), and then computes the next iterate as a proximal gradient map (see (3.2)) using the two extrapolated points. The proximal operators with respect to the generating functions H_i, $i=1,\ldots,s$, are presented in Section 3.3.1.

Algorithm 2.2 Update for IBP

At the j-th iteration of the k-th inner loop, extrapolate

$$\hat{x}_i = x_i^{(k,j-1)} + \alpha_i^{(k,j)} \left(x_i^{(k,j-1)} - y_i \right), \tag{2.1}$$

and compute

$$x_i^{(k,j)} \in \operatorname{prox}^{H_i}_{\beta_i^{(k,j)}, F_i^{(k,j)}} \left(\hat{x}_i \right). \tag{2.2}$$

The choice of the parameters $\alpha_i^{(k,j)}$ and $\beta_i^{(k,j)}$ are discussed in Section 4.1 while the Bregman proximal map with respect to the generating function H_i is defined in (3.1).

Algorithm 2.3 Update for IBPG

At the j-th iteration of the k-th inner loop, do the following extrapolation

$$\begin{aligned} \hat{x}_i &= x_i^{(k,j-1)} + \alpha_i^{(k,j)} \left(x_i^{(k,j-1)} - y_i \right), \\ \grave{x}_i &= x_i^{(k,j-1)} + \gamma_i^{(k,j)} \left(x_i^{(k,j-1)} - y_i \right), \end{aligned} \tag{2.3}$$

and compute

$$x_i^{(k,j)} \in \operatorname{Gprox}^{H_i}_{\beta_i^{(k,j)}, r_i, f_i} \left(\grave{x}_i, \hat{x}_i \right). \tag{2.4}$$

The choice of the parameters $\alpha_i^{(k,j)}$, $\beta_i^{(k,j)}$ and $\gamma_i^{(k,j)}$ are discussed in Section 4.1 while the Bregman proximal gradient map with respect to the generated function H_i is defined in (3.2).

As for $x_i^{(k,j)}$, we will also use the notation $\bar{\alpha}_i^{(k,m)}$, $\bar{\beta}_i^{(k,m)}$, and $\bar{\gamma}_i^{(k,m)}$ for the corresponding values of $\alpha_i^{(k,j)}$, $\beta_i^{(k,j)}$ and $\gamma_i^{(k,j)}$ that are used in (2.1), (2.2), (2.3) and (2.4) to update block i from $\bar{x}_i^{(k,m-1)}$ to $\bar{x}_i^{(k,m)}$.

3. Preliminaries.

In this section, we give important definitions and properties that will allow us to provide our convergence results.

3.1. Preliminaries of non-convex non-smooth optimization.

Let $g : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

Definition 3.1. *(i) For any $x \in \operatorname{dom} g$, and $d \in \mathbb{E}$, we denote the directional derivative of g at x in the direction d by*

$$g'(x; d) = \liminf_{\tau \downarrow 0} \frac{g(x + \tau d) - g(x)}{\tau}.$$

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

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

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

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

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

The following definition, see [45, Section 3], is necessary in our convergence analysis for the inertial version of (1.2) without the smoothness assumption on f.

Definition 3.2. *(i) We say that* $x^* \in \operatorname{dom F}$ *is a critical point type I of* F *if* $F'(x^*;d) \geq 0, \forall\, d$.

(ii) $x^* \in \operatorname{dom F}$ *is said to be a coordinatewise minimum of* F *if*

$$F\left(x^* + (0,\ldots,d_i,\ldots,0)\right) \geq F(x^*), \forall\, d_i \in \mathbb{E}_i, \forall\, i = 1,\ldots,s.$$

(iii) We say that F *is regular at* $x \in \operatorname{dom F}$ *if for all* $d = (d_1,\ldots,d_s)$ *such that*

$$F'\left(z;(0,\ldots,d_i,\ldots,0)\right) \geq 0, i = 1,\ldots,s,$$

then $F'(x;d) \geq 0$.

It is straightforward to see from the definition that if F is regular at x^* and x^* is a coordinate-wise minimum point of F then x^* is also a critical point type I of F. We refer the readers to Lemma 3.1 in [45] for the sufficient conditions that imply the regularity of F. When f is assumed to be smooth (for the analysis of inertial version of (1.3)), Definition 3.3 will be used.

Definition 3.3. *We call* $x^* \in \operatorname{dom F}$ *a critical point type II of* F *if* $0 \in \partial F\left(x^*\right).$

We note that if x^* is a minimizer of F then x^* is a critical point type I and type II of F.

3.2. Kurdyka-Łojasiewicz functions. Let us define Kurdyka-Łojasiewicz functions.

Definition 3.4. *A function* $\phi(x)$ *is said to have the Kurdyka-Łojasiewicz (KL) property at* $\bar{x} \in \operatorname{dom}\partial\,\phi$ *if there exists* $\eta \in (0,+\infty]$, *a neighborhood* U *of* $\bar{x}$ *and a concave function* $\xi : [0,\eta) \to \mathbb{R}_+$ *that is continuously differentiable on* $(0,\eta)$, *continuous at* 0, $\xi(0) = 0$, *and* $\xi'(s) > 0$ *for all* $s \in (0,\eta)$, *such that for all* $x \in U \cap [\phi(\bar{x}) < \phi(x) < \phi(\bar{x}) + \eta]$, *the following inequality holds*

$$\xi'\left(\phi(x) - \phi(\bar{x})\right) \operatorname{dist}\left(0, \partial\phi(x)\right) \geq 1.$$

Here $\operatorname{dist}\left(0,\partial\phi(x)\right) = \min\left\{\|y\| : y \in \partial\phi(x)\right\}$.

If $\phi(x)$ *satisfies the KL property at each point of* $\operatorname{dom}\partial\phi$ *then* ϕ *is a KL function.*

The class of functions that satisfy the KL property is rich enough to cover many nonconvex non-smooth functions found in practical applications. Some noticeable examples include real analytic functions, semi-algebraic functions, locally strongly convex functions; see [12, 48] and the appendix of [14].

The following lemma (see Lemma 6 of [14]), is the cornerstone to establish the global convergence of our proposed methods.

Lemma 3.5 (Uniformized KL property). *Let* ϕ *be a proper and lower semicontinuous function. Assuming that* ϕ *satisfies the KL property and is constant on a compact set* Ω. *Then there exist* $\varepsilon > 0$, $\eta > 0$ *and a function* ξ *satisfying the conditions in Definition 3.4 such that for all* $\bar{x} \in \Omega$ *and*

$$x \in \{x \in \mathbb{E} : \operatorname{dist}(x,\Omega) < \varepsilon\} \cap [\phi(\bar{x}) < \phi(x) < \phi(\bar{x}) + \eta]$$

we have

$$\xi'\left(\phi(x) - \phi(\bar{x})\right) \operatorname{dist}\left(0, \partial\phi(x)\right) \geq 1.$$

3.3. Bregman proximal maps. In this section, we discuss in details Bregman distances, their proximal maps and how they can be computed. They are key in our analysis of upcoming algorithms.

3.3.1. The Bregman distance.

Definition 3.6. *Let $H_i : \mathbb{E}_i \to \mathbb{R}$ be a strictly convex function that is continuously differentiable. The Bregman distance associated with H_i is defined as:*

$$D_i(u,v) = H_i(u) - H_i(v) - \langle \nabla H_i(v), u - v \rangle, \forall u, v \in \mathbb{E}_i.$$

The squared Euclidean distance $D_i(u,v) = \frac{1}{2}\|u-v\|_2^2$, which corresponds to $H_i(u) = \frac{1}{2}\|u\|_2^2$, is a simple example of Bregman distances.

We now recall some useful properties of Bregman distance in the following lemmas. Their proofs can be found in [17, 10].

Lemma 3.7. *(i) If H_i is strongly convex with constant σ_i, that is,*

$$H_i(u) \geq H_i(v) + \langle \nabla H_i(v), u - v \rangle + \frac{\sigma_i}{2}\|u-v\|^2, \forall u, v \in \mathbb{E}_i$$

then

$$D_i(u,v) \geq \frac{\sigma_i}{2}\|u-v\|^2, \forall u, v \in \mathbb{E}_i.$$

(ii) If ∇H_i is L_{H_i}-Lipschitz continuous, then

$$D_i(u,v) \leq \frac{L_{H_i}}{2}\|u-v\|^2, \forall u, v \in \mathbb{E}_i.$$

Lemma 3.8. *Let $D_i(u,v)$ be the Bregman distance associated with H_i.*

(i) (The three point identity) We have:

$$D_i(u,v) + D_i(v,w) = D_i(u,w) + \langle \nabla H_i(w) - \nabla H_i(v), u - v \rangle, \forall u, v, w \in \mathbb{E}_i.$$

(ii) [46, Property 1] Let $z^+ = \operatorname{argmin}_u \phi(u) + D_i(u,z)$, where ϕ is a proper convex function. Then for all $u \in \mathbb{E}_i$ we have

$$\phi(u) + D_i(u,z) \geq \phi(z^+) + D_i(z^+,z) + D_i(u,z^+).$$

3.3.2. The Bregman proximal maps. When we replace the squared Euclidean distance in the definition of the proximal point in (1.2) by the Bregman distance, we get the Bregman proximal map. Suppose the Bregman distance D_i is generated by H_i.

Definition 3.9. *For a given $v \in \mathbb{E}_i$, and a positive number β, the Bregman proximal map of a function ϕ is defined by*

$$\operatorname{prox}^{H_i}_{\beta,\phi}(v) := \operatorname{argmin}\left\{\phi(u) + \frac{1}{\beta}D_i(u,v) : u \in \mathbb{E}_i\right\}. \tag{3.1}$$

We need the definition of Bregman proximal gradient map for the analysis of the inertial version of (1.3).

Definition 3.10. *For given $u_1 \in \operatorname{int}\operatorname{dom} g, u_2 \in \mathbb{E}_i$ and $\beta > 0$, the Bregman proximal gradient map of a pair of non-convex function (ϕ, g) (g is continuously differentiable) is defined by*

$$\operatorname{Gprox}^{H_i}_{\beta,\phi,g}(u_1,u_2) := \operatorname{argmin}\left\{\phi(u) + \langle \nabla g(u_1), u \rangle + \frac{1}{\beta}D_i(u,u_2) : u \in \mathbb{E}_i\right\} \tag{3.2}$$

For notation succinctness, if the generating function H_i is clear in the context, we would omit the upper-script H_i in the notation of the corresponding Bregman proximal maps. As ϕ can be non-convex, $\operatorname{prox}_{\beta,\phi}(v)$ in (3.1) and $\operatorname{Gprox}_{\beta,\phi,g}(u_1,u_2)$ in (3.2) are set-valued maps in general. Various types of assumptions can be made to guarantee well-definedness of (3.1) and (3.2). See for example [19, 43, 44] for the well-posedness of (3.1). For the well-posedness of (3.2), we refer the readers to [15, Lemma 3.1], [9, Lemma 2] and [2, Section 4]. Note that the Bregman proximal gradient maps in [9, 15] use the same point for evaluating the gradient and the Bregman distance while ours allow using two different points u_1 and u_2. This modification is important for our analysis; however, it does not affect the proofs of the lemmas in [9] and [15].

Throughout this paper, we assume the following.

Assumption 3.11. *(A1) The function H_i, $i=1,\ldots,s$, is σ_i-strongly convex, continuously differentiable and ∇H_i is L_{H_i}-Lipschitz continuous.*
(A2) The proximal maps (3.1) and (3.2) are well-defined.

It is worth noting that Assumption (A1) holds if H_i satisfies $L_{H_i}\mathcal{I} \preceq \nabla^2 H_i \preceq \sigma_i \mathcal{I}$ [1]. The Euclidean distance (or, more generally, a quadratic entropy distance, see [42]) is a typical example of a Bregman distance that satisfies Assumption (A1).

The following inequality is crucial for our convergence analysis.

Lemma 3.12. *For a given $\hat{w}\in\mathbb{E}_i$, if $w^+\in \operatorname{prox}^{H_i}_{\beta,\phi}(\hat{w})$ then for all $w\in\mathbb{E}_i$ we have*

$$\phi(w^+)+\frac{1}{\beta}D_i\left(w^+,w\right)\le \phi(w)+\frac{1}{\beta}\left\langle \nabla H_i(\hat{w})-\nabla H_i(w), w^+-w\right\rangle .$$

Proof. It follows from the definition of w^+ that

$$\phi\left(w^+\right)+\frac{1}{\beta}D_i\left(w^+,\hat{w}\right)\le \phi(w)+\frac{1}{\beta}D_i\left(w,\hat{w}\right).$$

On the other hand, by Lemma 3.8 (i) we get

$$D_i\left(w^+,\hat{w}\right)-D_i\left(w,\hat{w}\right)=D_i\left(w^+,w\right)-\left\langle \nabla H_i\left(\hat{w}\right)-\nabla H_i\left(w\right), w^+-w\right\rangle .$$

The result follows. □

If ϕ is convex, applying Lemma 3.8 (ii), we get the following lemma. Lemma 3.13 will be used when the function r_i is convex.

Lemma 3.13. *For a given $\hat{w}\in\mathbb{E}_i$, if $w^+\in\operatorname{prox}^{H_i}_{\beta,\phi}(\hat{w})$ and ϕ is convex then for all $w\in\mathbb{E}_i$ we have*

$$\phi(w^+)+\frac{1}{\beta}D_i\left(w^+,\hat{w}\right)+\frac{1}{\beta}D_i\left(w,w^+\right)\le\phi(w)+\frac{1}{\beta}D_i\left(w,\hat{w}\right).$$

3.3.3. Evaluating Bregman proximal/proximal gradient maps.

It is crucial to be able to compute efficiently the Bregman proximal maps in (3.1) and (3.2). When D_i is the Euclidean distance, the maps reduce to the classical proximal/proximal gradient maps. We refer the readers to [37] for a comprehensive discussion on how to evaluate the classical maps.

In [9, Section 3.1], the authors present a splitting mechanism to evaluate (3.2) when u_1 and u_2 are identical. Following their methodology, we first define a Bregman gradient operator as follows:

$$\mathrm{p}_{\beta,g}(u_1,u_2):=\operatorname{argmin}\left\{\langle\nabla g(u_1),u\rangle+\frac{1}{\beta}D_i(u,u_2): u\in\mathbb{E}_i\right\}.$$

Writing the optimality conditions for (3.2) together with formal computations (see [9, Section 3.1] for the details), we can prove that

$$\operatorname{Gprox}_{\beta,\phi,g}(u_1,u_2)=\operatorname{prox}_{\beta,\phi}\left(\mathrm{p}_{\beta,g}(u_1,u_2)\right),$$

[1] A non-typical simple example of such function is $x\in\mathbb{R}\mapsto\log(x+\sqrt{1+x^2})+x^2$.

and

$$\mathrm{p}_{\beta,g}(u_1,u_2) = \nabla H_i^* \left(\nabla H_i(u_2) - \beta \nabla g(u_1)\right), \tag{3.3}$$

where H_i^* is the conjugate function of H_i. From (3.3), we see that the calculation of $\mathrm{p}_{\beta,g}(u_1,u_2)$ depends on the calculation of ∇H_i^*. Hence, once we can evaluate H_i^*, it is straighforward to evaluate $\mathrm{p}_{\beta,g}(u_1,u_2)$. A very simple example is the case $D_i\left(u,u_2\right) = \frac{1}{2}\left\|u-u_2\right\|_2^2$ for which $\mathrm{p}_{\beta,g}(u_1,u_2) = u_2 - \beta\nabla g(u_1)$; see [7, 9, 44] for more examples. Regarding to the evaluation of (3.1) in the general setting of Bregman distances, we note that the evaluation can be very difficult and refer the readers to [9, Section 5], [15, Section 5] and [44, Section 6] for some specific examples and discussions.

4. Convergence analysis.

We first discuss the choice of the parameters in Algorithms 2.2 and 2.3 (Section 4.1), then prove convergence of a subsequence of Algorithm 2.1 to a critical point of F (Section 4.2) and finally prove global convergence which will require some additional assumptions (Section 4.3).

4.1. Choosing parameters.

We first present methods to choose parameters for Algorithm 2.2 and Algorithm 2.3 such that their convergences are attainable.

Parameters for Algorithm 2.2. Let $0<\nu<1$. For $m=1,\ldots,d_i^k$ and $i=1,\ldots,s$, denote $\theta_i^{(k,m)} = \frac{\left(L_{H_i}\bar{\alpha}_i^{(k,m)}\right)^2}{2\nu\sigma_i\bar{\beta}_i^{(k,m)}}$. Let $\theta_i^{(k,d_i^k+1)} = \theta_i^{(k+1,1)}$. We choose $\bar{\alpha}_i^{(k,m)}$ and $\bar{\beta}_i^{(k,m)}$ satisfying

$$\frac{(1-\nu)\sigma_i}{2\bar{\beta}_i^{(k,m)}} \geq \delta\theta_i^{(k,m+1)}, \quad \text{for } m=1,\ldots,d_i^k, \text{ where } \delta>1. \tag{4.1}$$

Remark 4.1. After rearranging, Condition (4.1) can be expressed as

$$\frac{\bar{\beta}_i^{(k,m)}}{\bar{\beta}_i^{(k,m+1)}}(\bar{\alpha}_i^{(k,m+1)})^2 \leq \vartheta\left(\frac{\sigma_i}{L_{H_i}}\right)^2, \tag{4.2}$$

where $\vartheta = \frac{(1-\nu)\nu}{\delta} \in (0,1)$. We see that, for a given $\vartheta\in(0,1)$, there always exist $0<\nu<1$ and $\delta>1$ such that $\vartheta = \frac{(1-\nu)\nu}{\delta}$. Hence, we can replace (4.1) by its simplified form (4.2) (which involves one constant ϑ instead of two) and continue our analysis.

Parameters for Algorithm 2.3. Considering Algorithm 2.3, we need to assume that $\nabla f_i^{(k,j)}$ is $L_i^{(k,j)}$-Lipschitz continuous, with $L_i^{(k,j)}>0$. For notational clarity, we correspondingly use $\bar{L}_i^{(k,m)}$ for $L_i^{(k,j)}$. To simplify the upcoming analysis, we choose $\bar{\beta}_i^{(k,m)} = \frac{\sigma_i}{\kappa\bar{L}_i^{(k,m)}}$ with $\kappa>1$. Let $0<\nu<1$. Denote

$$\lambda_i^{(k,m)} = \frac{1}{2}\left(\bar{\gamma}_i^{(k,m)} + \frac{\kappa L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i}\right)^2 \frac{\bar{L}_i^{(k,m)}}{\nu(\kappa-1)}, \text{for } m=1,\ldots,d_i^k \text{ and } i=1,\ldots,s.$$

Let $\lambda_i^{(k,d_i^k+1)} = \lambda_i^{(k+1,1)}$. We choose $\bar{\alpha}_i^{(k,m)}$, $\bar{\beta}_i^{(k,m)}$ and $\bar{\gamma}_i^{(k,m)}$ satisfying

$$\frac{(1-\nu)(\kappa-1)\bar{L}_i^{(k,m)}}{2} \geq \delta\lambda_i^{(k,m+1)}, \text{for } m=1,\ldots,d_i^k, \text{ where } \delta>1 \tag{4.3}$$

Remark 4.2. The method iPALM in [38] is a special case of IBPG when the Bregman distance D is Euclidean distance, when $s=2$ and when the two blocks are cyclically updated; however, our chosen parameters are different. In particular, the stepsize $\bar{\beta}_i^{(k,m)}$ of iPALM depends on the inertial parameters (see [38, Formula 4.9]) while we choose $\bar{\beta}_i^{(k,m)}$

independent on $\bar{\alpha}_i^{(k,m)}$ and $\bar{\gamma}_i^{(k,m)}$. Our parameter choice results in a more flexible scheme for the inertial parameters, especially it allows using dynamic inertial parameters (see Section 5). As also experimentally tested in [38], choosing the inertial parameters dynamically leads to a significant improvement of the algorithm performance. The analysis in [38] does not support this choice of parameters, while ours, at least, guarantees a subsequential convergence under Condition (4.3).

Throughout the paper, we make the following assumption for the parameters.

Assumption 4.3.

- *For IBP, there exist positive numbers W_1, $\overline{\alpha}$ and $\underline{\beta}$ such that $\theta_i^{(k,m)} \geq W_1$, $\bar{\alpha}_i^{(k,m)} \leq \overline{\alpha}$ and $\underline{\beta} \leq \bar{\beta}_i^{(k,m)}$ for all $k \in \mathbb{N}$, $m = 1, \ldots, d_i^k$ and $i = 1, \ldots, s$.*
- *For IBPG, there exist positive numbers W_1, $\overline{L} > 0$, $\overline{\alpha}$, $\underline{\beta}$ and $\overline{\gamma}$ such that $\lambda_i^{(k,m)} \geq W_1$, $\bar{L}_i^{(k,m)} \leq \overline{L}$, $\bar{\alpha}_i^{(k,m)} \leq \overline{\alpha}$, $\underline{\beta} \leq \bar{\beta}_i^{(k,m)}$ and $\bar{\gamma}_i^{(k,m)} \leq \overline{\gamma}$ for all $k \in \mathbb{N}$, $m = 1, \ldots, d_i^k$ and $i = 1, \ldots, s$.*

4.2. Subsequential convergence. The following proposition serves as a cornerstone to prove the local convergence (that is, convergence of a subsequence to a critical point).

Proposition 4.4. *(i) We have*

$$\sum_{k=1}^{\infty} \left\| \tilde{x}^{(k)} - \tilde{x}^{(k)}_{\text{prev}} \right\|^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \sum_{i=1}^{s} \sum_{m=1}^{d_i^k} \left\| \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\|^2 < \infty.$$

(ii) If there exists a limit point x^* of the sequence $\left\{\tilde{x}^{(k)}\right\}_{k\in\mathbb{N}}$ (that is, there exists a subsequence $\left\{\tilde{x}^{(k_n)}\right\}$ converging to x^*), then we have $\lim_{n\to\infty} r_i\left(\bar{x}_i^{(k_n,m)}\right) = r_i\left(x_i^*\right)$.

Proof. See Appendix A.1

Remark 4.5 (Relaxing (4.1) for block-convex F). For IBP, if F is block-wise convex then we can choose $\bar{\alpha}_i^{(k,m)}$ and $\bar{\beta}_i^{(k,m)}$ satisfying

$$\frac{2(1-\nu)\sigma_i}{\bar{\beta}_i^{(k,m)}} \geq \delta\theta_i^{(k,m+1)}, \quad \text{for } m = 1, \ldots, d_i^k, \tag{4.4}$$

and Proposition 4.4 will still hold; see Appendix B.1 for the proof. Compared to the condition (4.1), condition (4.4) allows larger values of the extrapolation parameters $\bar{\alpha}_i^{(k,m)}$ when using the same stepsize $\bar{\beta}_i^{(k,m)}$.

Remark 4.6 (Relaxing (4.3) for convex r_i's). If the functions r_i's are convex (note that f is not necessary block-wise convex) then we can use a larger stepsize. Specifically, we can use

$$\bar{\beta}_i^{(k,m)} = \sigma_i/\bar{L}_i^{(k,m)}, \qquad \lambda_i^{(k,m)} = \frac{1}{2}\left(\bar{\gamma}_i^{(k,m)} + \frac{L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i}\right)^2 \frac{\bar{L}_i^{(k,m)}}{\nu},$$

and choose $\bar{\alpha}_i^{(k,m)}$ and $\bar{\gamma}_i^{(k,m)}$ satisfying

$$\frac{(1-\nu)\bar{L}_i^{(k,m)}}{2} \geq \delta\lambda_i^{(k,m+1)}, \text{for } m = 1, \ldots, d_i^k, \tag{4.5}$$

and Proposition 4.4 still holds; see Appendix B.2 for the proof.

Remark 4.7 (Relaxing (4.3) for block-convex f and convex r_i's). If the r_i's are convex and $f(x)$ is block-wise convex, then we can use larger extrapolation parameters. Specifically, we choose $H_i(x_i) = \frac{1}{2}\left\|x_i\right\|^2$ and let $\bar{\beta}_i^{(k,m)} = 1/\bar{L}_i^{(k,m)}$ and

$$\lambda_i^{(k,m)} = \left(\left(\bar{\gamma}_i^{(k,m)}\right)^2 + \frac{\left(\bar{\gamma}_i^{(k,m)} - \bar{\alpha}_i^{(k,m)}\right)^2}{\nu}\right)\frac{\bar{L}_i^{(k,m)}}{2},$$

where $0 < \nu < 1$, and choose $\bar{\alpha}_i^{(k,m)}$ and $\bar{\gamma}_i^{(k,m)}$ satisfying

$$\frac{1-\nu}{2}\bar{L}_i^{(k,m)} \geq \delta\lambda_i^{(k,m+1)}, \text{for } m = 1,\dots,d_i^k.$$

For these values, Proposition 4.4 still holds; see Appendix B.3 for the proof. In Section 5 we numerically show that choosing $\bar{\gamma}_i^{(k,m)} \neq \bar{\alpha}_i^{(k,m)}$ can significantly improve the performance of the algorithm.

We are now ready to state the local convergence result.

THEOREM 4.8. *(i) For IBP, if F is regular then every limit point of $\left\{\tilde{x}^{(k)}\right\}_{k\in\mathbb{N}}$ is a critical point type I of F. If f is continuously differentiable then every limit point of $\left\{\tilde{x}^{(k)}\right\}_{k\in\mathbb{N}}$ is a critical point type II of F.*

(ii) For IBPG, every limit point of $\left\{\tilde{x}^{(k)}\right\}_{k\in\mathbb{N}}$ is a critical point type II of F.

Proof. (i) Let n in (A.7) go to ∞. We have that

$$F\left(x^*\right) \leq F\left(x_1^*,\dots,x_i,\dots,x_s^*\right) + \frac{1}{\underline{\beta}}D_i\left(x_i,x_i^*\right), \forall x_i \in \mathbb{E}_i. \tag{4.6}$$

Inequality (4.6) shows that x_i^* is a minimum point of $x_i \mapsto F\left(x_1^*,\dots,x_i,\dots,x_s^*\right) + \frac{1}{\underline{\beta}}D_i\left(x_i,x_i^*\right)$. Note that $\nabla D_i\left(x_i,x_i^*\right) = \nabla H_i(x_i) - \nabla H_i(x_i^*)$, hence the directional derivative of $x_i \mapsto D_i\left(x_i,x_i^*\right)$ at x_i^* along d_i equals 0 for all $d_i \in \mathbb{E}_i$. Hence, from (4.6) we deduce that $F'\left(x^*;(0,\dots,d_i,\dots,0)\right) \geq 0, \forall d_i \in \mathbb{E}_i$. Together with the regularity assumption gives the result.

When f is continuously differentiable, Proposition 2.1 in [4] shows that $\partial F(x^*) = \left\{\partial_{x_1}F(x^*)\right\} \times \ldots \times \left\{\partial_{x_s}F(x^*)\right\}$. Together with (4.6) (which implies $0 \in \partial_{x_i}F(x^*)$), we obtain $0 \in \partial F(x^*)$.

(ii) As $\nabla f_i^{(k,j)}$ is $\bar{L}_i^{(k,m)}$-Lipschitz continuous, we have

$$f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right) \leq f_i^{(k,j)}\left(x_i\right) + \left\langle \nabla f_i^{(k,j)}\left(x_i\right), \bar{x}_i^{(k,m)} - x_i\right\rangle + \frac{\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m)} - x_i\right\|^2.$$

Together with (A.11) we obtain

$$\begin{aligned}
&f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right) + r_i\left(\bar{x}_i^{(k,m)}\right) \\
&\leq f_i^{(k,j)}\left(x_i\right) + \left\langle \nabla f_i^{(k,j)}\left(x_i\right) - \nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)} - x_i\right\rangle + r_i(x_i) \\
&\quad + \frac{1}{\bar{\beta}_i^{k,m}}D_i\left(x_i,\hat{x}_i\right) - \frac{1}{\bar{\beta}_i^{k,m}}D_i\left(\bar{x}_i^{(k,m)},\hat{x}_i\right) + \frac{\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m)} - x_i\right\|^2 \\
&\leq f_i^{(k,j)}\left(x_i\right) + r_i(x_i) + \left\langle \nabla f_i^{(k,j)}\left(x_i\right) - \nabla f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right), \bar{x}_i^{(k,m)} - x_i\right\rangle \\
&+ \left\langle \nabla f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right) - \nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)} - x_i\right\rangle + \frac{1}{\underline{\beta}}D_i(x_i,\hat{x}_i) + \frac{\overline{L}}{2\sigma_i}D_i\left(x_i,\bar{x}_i^{(k,m)}\right).
\end{aligned} \tag{4.7}$$

Let $k=k_n$ in (4.7). Similarly to the proof of Theorem 4.8(i), we first take $n\to\infty$ to obtain

$$F(x^*)\le f(x_1^*,\dots,x_i,\dots,x_s^*)+r_i(x_i)+\left(\frac{1}{\breve{\beta}}+\frac{\overline{L}}{2\sigma_i}\right)D_i\left(x_i,x_i^*\right),\forall\, x_i\in\mathbb{E}_i.$$

This inequality yields that $0\in\partial_{x_i}F(x^*)$. We then use [4, Proposition 2.1] to complete the proof. □

4.3. Global convergence. A key tool of the upcoming global convergence analysis is the use of the auxiliary function Ψ defined as follows

$$\Psi\left(\acute{y},\breve{y}\right):=F\left(\acute{y}\right)+\rho D\left(\acute{y},\breve{y}\right),$$

where ρ is a positive constant and $D\left(\acute{y},\breve{y}\right)=\sum_{i=1}^s D_i\left(\acute{y}_i,\breve{y}_i\right)$. Let us consider the sequence $\left\{Y^{(k)}\right\}_{k\in\mathbb{N}}$ with $Y^{(k)}=\left(\acute{y}^{(k)},\breve{y}^{(k)}\right)=\left(\tilde{x}^{(k)},\tilde{x}^{(k)}_{\rm prev}\right)$, where $\left\{\tilde{x}^{(k)}_{\rm prev}\right\}_{k\in\mathbb{N}}$ with $\left(\tilde{x}^{(k)}_{\rm prev}\right)_i=\bar{x}_i^{(k,d_i^k-1)}$ and $\left\{\tilde{x}^{(k)}\right\}_{k\in\mathbb{N}}$ are the sequences generated by Algorithm 2.1. We have

$$\Psi\left(Y^{(k)}\right)=F\left(\tilde{x}^{(k)}\right)+\rho D\left(\tilde{x}^{(k)},\tilde{x}^{(k)}_{\rm prev}\right), \tag{4.8}$$

and

$$\left\|Y^{(k)}-Y^{(k+1)}\right\|^2=\left\|\tilde{x}^{(k)}-\tilde{x}^{(k+1)}\right\|^2+\left\|\tilde{x}^{(k)}_{\rm prev}-\tilde{x}^{(k+1)}_{\rm prev}\right\|^2.$$

We define

$$\begin{aligned}\varphi_k^2&:=\sum_{i=1}^s\sum_{m=0}^{d_i^{(k+1)}}\left\|\bar{x}_i^{(k+1,m)}-\bar{x}_i^{(k+1,m-1)}\right\|^2\\&=\sum_{i=1}^s\sum_{m=1}^{d_i^{(k+1)}}\left\|\bar{x}_i^{(k+1,m)}-\bar{x}_i^{(k+1,m-1)}\right\|^2+\sum_{i=1}^s\left\|\bar{x}_i^{(k+1,0)}-\bar{x}_i^{(k+1,-1)}\right\|^2\\&=\sum_{i=1}^s\sum_{m=1}^{d_i^{(k+1)}}\left\|\bar{x}_i^{(k+1,m)}-\bar{x}_i^{(k+1,m-1)}\right\|^2+\left\|\tilde{x}^{(k)}-\tilde{x}^{(k)}_{\rm prev}\right\|^2.\end{aligned}$$

We make the following additional assumption.

ASSUMPTION 4.9. *The sequences $\left\{\tilde{x}^{(k)}\right\}_{k\in\mathbb{N}}$ generated by Algorithm 2.2 and 2.3 are bounded.*

In Proposition 4.12 we will prove that $\Psi\left(Y^{(k)}\right)$ is non-increasing; thus, $\Psi\left(Y^{(k)}\right)$ is upper bounded by $\Psi\left(Y^{(-1)}\right)$. Moreover, note that $D\left(\tilde{x}^{(k)},\tilde{x}^{(k)}_{\rm prev}\right)\ge 0$. Hence, from (4.8) this implies that $F\left(\tilde{x}^{(k)}\right)$ is also upper bounded by $\Psi\left(Y^{(-1)}\right)$. Therefore, we can say that Assumption 4.9 is satisfied when F has bounded level sets.

Denote $\sigma=\min\{\sigma_1,\dots,\sigma_s\}$ and $L_H=\max\{L_{H_1},\dots,L_{H_s}\}$.

PROPOSITION 4.10. *We have*

(i) $\varphi_k^2\ge\frac{1}{2(\bar{T}-s+1)}\left\|Y^{(k)}-Y^{(k+1)}\right\|^2$.

(ii) Denote

$$\nabla H=\left(\nabla H_1,\dots,\nabla H_s\right),\text{ and }\nabla^2H=\left(\nabla^2H_1,\dots,\nabla^2H_s\right).$$

Let $q^{(k)}\in\partial F\left(\tilde{x}^{(k)}\right)$. *Denote*

$$\hat{q}^{(k)}=\left(q^{(k)}+\rho\nabla H\left(\tilde{x}^{(k)}\right)-\nabla H\left(\tilde{x}^{(k)}_{\rm prev}\right),\rho\nabla^2H\left(\tilde{x}^{(k)}_{\rm prev}\right)\left[\tilde{x}^{(k)}_{\rm prev}-\tilde{x}^{(k)}\right]\right).$$

If f is smooth, then we have $\hat{q}^{(k)} \in \partial\Psi\left(Y^{(k)}\right)$, and

$$\left\|\hat{q}^{(k)}\right\|^2 \le 2\left\|q^{(k)}\right\|^2 + O\left(\left\|\tilde{x}^{(k)} - \tilde{x}^{(k)}_{\text{prev}}\right\|^2\right). \tag{4.9}$$

Proof. (i) We use the inequality $(a_1+\ldots+a_n)^2 \le n(a_1^2+\ldots+a_n^2)$ and $d_i^{(k)} \le \bar{T}-s+1$ to obtain

$$\begin{aligned} \varphi_k^2 &\ge \textstyle\sum_{i=1}^s \sum_{m=1}^{d_i^{(k+1)}} \left\|\bar{x}_i^{(k+1,m)} - \bar{x}_i^{(k+1,m-1)}\right\|^2 \\ &\ge \textstyle\sum_{i=1}^s \left(1/d_i^{(k+1)}\right) \left\|\bar{x}_i^{\left(k+1,d_i^{(k+1)}\right)} - \bar{x}_i^{(k+1,0)}\right\|^2 \ge \frac{1}{\bar{T}-s+1}\left\|\tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right\|^2. \end{aligned} \tag{4.10}$$

Note that $\left(\tilde{x}^{(k+1)}_{\text{prev}}\right)_i = \bar{x}_i^{\left(k+1,d_i^{(k+1)}-1\right)}$. Similarly to (4.10), we have

$$\varphi_k^2 = \sum_{i=1}^s \sum_{m=0}^{d_i^{(k+1)}-1} \left\|\bar{x}_i^{(k+1,m)} - \bar{x}_i^{(k+1,m-1)}\right\|^2 \ge \frac{1}{(\bar{T}-s+1)}\left\|\tilde{x}^{(k+1)}_{\text{prev}} - \tilde{x}^{(k)}_{\text{prev}}\right\|^2. \tag{4.11}$$

Summing up (4.10) and (4.11) we get the result.

(ii) Similarly to [4, Proposition 2.1], we can prove that

$$\partial\Psi\left(\acute{y},\breve{y}\right) = \left\{\partial F(\acute{y}) + \rho\nabla H(\acute{y}) - \nabla H(\breve{y})\right\} \times \left\{\rho\nabla^2 H\left(\breve{y}\right)\left[\breve{y}-\acute{y}\right]\right\}. \tag{4.12}$$

Therefore, $\hat{q}^{(k)} \in \partial\Psi\left(Y^{(k)}\right)$. We have

$$\begin{aligned} \left\|\hat{q}^{(k)}\right\|^2 &= \left\|q^{(k)} + \rho\nabla H\left(\tilde{x}^{(k)}\right) - \rho\nabla H\left(\tilde{x}^{(k)}_{\text{prev}}\right)\right\|^2 + \rho^2\left\|\nabla^2 H\left(\tilde{x}^{(k)}_{\text{prev}}\right)\left[\tilde{x}^{(k)}_{\text{prev}} - \tilde{x}^{(k)}\right]\right\|^2 \\ &\le 2\left\|q^{(k)}\right\|^2 + 2\rho^2 L_H^2\left\|\tilde{x}^{(k)} - \tilde{x}^{(k)}_{\text{prev}}\right\|^2 + \rho^2\left\|\nabla^2 H\left(\tilde{x}^{(k)}_{\text{prev}}\right)\left[\tilde{x}^{(k)}_{\text{prev}} - \tilde{x}^{(k)}\right]\right\|^2, \end{aligned}$$

where we use the inequality $(a+b)^2 \le 2a^2+2b^2$ and the Lipschitz continuity of ∇H_i. Finally, note that $\nabla^2 H_i \preceq L_H\mathbf{I}$, where $\mathbf{I}$ is the identity operator. Then (4.9) follows. □

4.3.1. Global convergence recipe.

As mentioned in the introduction, we know that [4, 5] and [14] are the first works proving the global convergence (that is, the convergence of the whole sequence to a critical point) of proximal point algorithms for solving non-convex non-smooth problems. They propose a general proof recipe in which two important conditions – sufficient decrease property and relative error condition (or a subgradient lower bound for the iterates gap) are required for the generated sequence. We note that a direct deployment of the methodology to our proposed algorithms is not possible since the relaxation property does not hold (that is, the objective functions are not monotonically decreasing) and our methods allow for a randomized strategy. In the following theorem, we modify the proof recipe proposed in [14] so that it is applicable to our proposed methods.

THEOREM 4.11. *Let $\Phi : \mathbb{R}^N \to (-\infty,+\infty]$ be a proper and lower semicontinuous function which is bounded from below. Let $\mathcal{A}$ be a generic algorithm which generates a bounded sequence $\left\{z^{(k)}\right\}_{k\in\mathbb{N}}$ by*

$$z^{(0)} \in \mathbb{R}^N, z^{(k+1)} \in \mathcal{A}\left(z^{(k)}\right), \quad k = 0,1,\ldots.$$

Assume that there exist positive constants ρ_1,ρ_2 and ρ_3 and a nonnegative sequence $\{\zeta_k\}_{k\in\mathbb{N}}$ such that the following conditions are satisfied

(B1) ***Sufficient decrease property:***

$$\rho_1 \left\| z^{(k)} - z^{(k+1)} \right\|^2 \le \rho_2 \zeta_k^2 \le \Phi\left(z^{(k)}\right) - \Phi\left(z^{(k+1)}\right), \quad \forall k = 0, 1, \ldots$$

(B2) ***Boundedness of subgradient:***

$$\left\| w^{(k+1)} \right\| \le \rho_3 \zeta_k, \quad w^{(k)} \in \partial\Phi\left(z^{(k)}\right), \quad \forall k = 0, 1, \ldots$$

Furthermore, assume that

(B3) ***KL property:*** *Φ is a KL function.*

(B4) ***A continuity condition:*** *If a subsequence $\left\{z^{(k_n)}\right\}_{n\in\mathbb{N}}$ of $\left\{z^{(k)}\right\}$ converges to $\bar{z}$ then $\Phi\left(z^{(k_n)}\right) \to \Phi\left(\bar{z}\right)$ as $n \to \infty$.*

Then we have $\sum_{k=1}^{\infty} \zeta_k < \infty$, and $\left\{z^{(k)}\right\}$ converges to a critical point of Φ.

To prove Theorem 4.11, we use the same methodology established in [14] (see the proof of [14, Theorem 1 (i)]). It is worth noting that the same techniques were used in the recent paper [35] to prove an abstract inexact convergence theorem, see Section 3 of [35]. To make our paper self contained, we give the proof of Theorem 4.11 in Appendix A.2.

4.3.2. Global convergence of IBP and IBPG.

The following proposition gives an upper bound for the subgradients and a sufficient decrease property for $\left\{\Psi\left(Y^{(k)}\right)\right\}$.

PROPOSITION 4.12. *(i) We assume that f is continuously differentiable and ∇f is Lipschitz continuous on bounded subsets of $\mathbb{E}$. We then have $\left\|\hat{q}^{(k+1)}\right\| = O\left(\varphi_k\right)$, for some $\hat{q}^{(k)} \in \partial\Psi\left(Y^{(k)}\right)$.*

(ii) Together with the condition in Proposition 4.12 (i), let us assume that there exists a constant W_2 such that, for all $k \in \mathbb{N}$, $m = 1, \ldots, d_i^k$ and $i = 1, \ldots, s$, we have $\theta_i^{(k,m)} \le W_2$ for IBP, $\lambda_i^{(k,m)} \le W_2$ for IBPG and $\delta > (L_H W_2)/(\sigma W_1)$. Let $\rho = \frac{\delta W_1}{L_H} + \frac{W_2}{\sigma}$ in (4.8) and let $\rho_2 = \frac{\delta\sigma W_1}{2L_H} - \frac{W_2}{2}$. Then we have

$$\Psi\left(Y^{(k)}\right) - \Psi\left(Y^{(k+1)}\right) \ge \rho_2 \varphi_k^2.$$

Proof. See Appendix A.3 □

We are now ready to state our global convergence result.

THEOREM 4.13. *Assume F is a KL-function and the conditions of Proposition 4.12 are satisfied. Then the whole sequence $\left\{\tilde{x}^{(k)}\right\}_{k\in\mathbb{N}}$ generated by IBP or IBPG converges to critical point type II of F.*

Proof. We now use Theorem 4.11 to prove the global convergence for both IBP and IBPG. We verify the Conditions (B1)-(B4) in Theorem 4.11 for the auxiliary function Ψ and the sequence $\left\{Y^{(k)}\right\}_{k\in\mathbb{N}}$. Proposition 4.10(i) and Proposition 4.12 show that the Conditions (B1) and (B2) are satisfied. Since F is a KL-function, Ψ is also a KL function. Hence Condition (B3) is satisfied.

Suppose $Y^* \in w\left(Y^{(0)}\right)$ is a limit point of $\left\{Y^{(k)}\right\}$, then there exists a subsequence $\{k_n\}$ such that $\left\{Y^{(k_n)}\right\} = \left\{\left(\tilde{x}^{(k_n)}, \tilde{x}_{\text{prev}}^{(k_n)}\right)\right\}$ converges to Y^*. We remind that if $\left\{\tilde{x}^{(k_n)}\right\}$ converges to x^* then $\left\{\tilde{x}_{\text{prev}}^{(k_n)}\right\}$ also converges to x^*. Hence $Y^* = (x^*, x^*)$. Moreover, from Theorem 4.8, we have x^* is a critical point of F, that is, $0 \in \partial F(x^*)$. Hence, we derive from (4.12) that $0 \in \partial\Psi(Y^*)$, that is, Y^* is a critical point of Ψ. On the other hand, from Proposition 4.4(ii) we have $F\left(\tilde{x}^{(k_n)}\right) \to F(x^*)$ (choose $m = d_i^{k_n}$). Therefore, (4.8) implies that $\Psi\left(Y^{(k_n)}\right)$ converges to $\Psi(Y^*)$. And consequently, Condition (B4) is satisfied. Applying Theorem 4.11, we have that the sequence $\left\{Y^{(k)}\right\}$ converges to (x^*, x^*). Hence the sequence $\left\{\tilde{x}^{(k)}\right\}$ converges to x^*. □

Remark 4.14. Note that we need the additional condition $\delta > \frac{L_H W_2}{\sigma W_1}$ in order to obtain the global convergence in Theorem 4.13. Therefore, it makes sense to show that there exists such δ that Condition (4.3) for IBPG (or Condition (4.1) for IBP) is also satisfied. Let us prove it for IBPG, it would be similar for IBP. Indeed, such δ would exist if we have $\frac{(1-\nu)(\kappa-1)\bar{L}_i^{(k,m)}}{2\lambda_i^{(k,m+1)}} > \frac{L_H W_2}{\sigma W_1}$, which would be satisfied if $\frac{\nu(1-\nu)(\kappa-1)^2\bar{L}_i^{(k,m)}}{\bar{L}_i^{(k,m+1)}\xi_i^{(k,m+1)}} > \frac{L_H W_2}{\sigma W_1}$, where $\xi_i^{(k,m)} = \big(\bar{\gamma}_i^{(k,m)} + \frac{\kappa L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i}\big)^2$. In other words, δ would exist if we have

$$\frac{\sigma\nu(1-\nu)(\kappa-1)^2\bar{L}_i^{(k,m)}}{L_H\bar{L}_i^{(k,m+1)}}\frac{W_1}{W_2} > \xi_i^{(k,m+1)}. \tag{4.13}$$

Suppose $\xi_1 \le \xi_i^{k,m} \le \xi_2$ and $0 < L_1 \le \bar{L}^{(k,m)} \le L_2$. We then have $\frac{W_1}{W_2} = \frac{\xi_1 L_1}{\xi_2 L_2}$, and (4.13) holds if $\frac{\sigma\nu(1-\nu)(\kappa-1)^2\bar{L}_i^{(k,m)}L_1}{L_H\bar{L}_i^{(k,m+1)}L_2} > \xi_2$. Therefore, if we choose in advance two constants ξ_1 and ξ_2 such that $\xi_2 < \frac{\sigma\nu(1-\nu)(\kappa-1)^2L_1^2}{L_H L_2^2}$ and $0 < \xi_1 < \xi_2$, then there always exists $\xi_i^{(k,m)}$ accordingly such that Condition (4.13) is satisfied.

5. Application to nonnegative matrix factorization (NMF).

Let us consider the following well-known NMF problem: Given $X \in \mathbb{R}_+^{\mathbf{m}\times\mathbf{n}}$ and the integer $\mathbf{r} < \min(\mathbf{m},\mathbf{n})$, solve

$$\min_{U\ge 0, V\ge 0} \|X - UV\|_F^2 \text{ such that}\, U \in \mathbb{R}_+^{\mathbf{m}\times\mathbf{r}} \text{ and}\, V \in \mathbb{R}_+^{\mathbf{r}\times\mathbf{n}}. \tag{5.1}$$

NMF is a key problem in data analysis and machine learning with applications in image processing, document classification, hyperspecral unmixing and audio source separation, to cite a few; see [18, 22, 21] and the references therein for more details.

NMF can be written as a problem of the form (1.1) with $s = 2$, letting $f(U,V) = \|X - UV\|_F^2$, and r_1 and r_2 being indicator functions of the nonnegative orthants containing U and V, that is, $r_1(U) = I_{\mathbb{R}_+^{\mathbf{m}\times\mathbf{r}}}(U)$, and $r_2(V) = I_{\mathbb{R}_+^{\mathbf{r}\times\mathbf{n}}}(V)$. We now apply IBPG and choose the parameters following Remark 4.7. We simply choose the Bregman distance to be the Euclidean distance. Applying IBP would not be as efficient because solving the subproblems in U and V exactly using a nonnegative least squares solver would be rather expensive; see, e.g., the discussion in [23]. We have

$$\nabla_U f = UVV^T - XV^T \;\text{ and }\; \nabla_V f = U^TUV - U^TX.$$

In each inner loop, our algorithm allows updating a block matrix U or V several times before updating the other one. As explained in [23], this repeating update would accelerate the convergence of the algorithm compared to the pure cyclic update rule, because the terms VV^T and XV^T (resp. U^TU and U^TX) for the update of U (resp. V) do not need to be recomputed hence the second evaluation of the gradient is much cheaper; namely, $O(\mathbf{mr}^2)$ (resp. $O(\mathbf{nr}^2)$) vs. $O(\mathbf{mnr})$ operations since in practice $\mathbf{r} \ll \min(\mathbf{m},\mathbf{n})$. We have $\bar{L}_1^{(k,m)} = \tilde{L}_1^{(k)} = \left\|\left(\tilde{V}^{(k-1)}\right)^T\tilde{V}^{(k-1)}\right\|$ and $\bar{L}_2^{(k,m)} = \tilde{L}_2^{(k)} = \left\|\left(\tilde{U}^{(k)}\right)^T\tilde{U}^{(k)}\right\|$ for $m \ge 1$.

In our experiment, we choose $\bar{\beta}_i^{(k,m)} = 1/\tilde{L}_i^{(k)}$, and

$$\bar{\gamma}_i^{(k,m)} = \min\left\{\frac{\tau_k - 1}{\tau_k}, \tilde{\gamma}\sqrt{\frac{\tilde{L}_i^{(k-1)}}{\tilde{L}_i^{(k)}}}\right\} \;\text{ and }\; \bar{\alpha}_i^{(k,m)} = \breve{\alpha}\bar{\gamma}_i^{(k,m)},$$

where $\tau_0 = 1$, $\tau_k = \frac{1}{2}(1 + \sqrt{1 + 4\tau_{k-1}^2})$, $\tilde{\gamma} = 0.99$ and $\breve{\alpha} = 1.01$. It is easy to verify that there exists $\delta > 1$ such that $\breve{\gamma}^2\left((\breve{\alpha}-1)^2/\nu + 1\right) < (1-\nu)/\delta$ with $\nu = 0.0099$. Hence, our choice of parameters satisfy the conditions of Remark 4.7.

Interestingly, we can decompose the NMF problem in more than $s = 2$ blocks of variables. For example, noting that $UV = \sum_{i=1}^{r} U_{:i}V_{i:}$, we can also write NMF as a function of $2 \times \mathbf{r}$ variables $U_{:i}$, $i = 1, \ldots, \mathbf{r}$ (the columns of U) and $V_{i:}$, $i = 1, \ldots, \mathbf{r}$ (the rows of V). In that case, we can apply IBP efficiently for NMF as the updates will have a closed-form solutions, see [23]. In each inner loop, we cyclically update the columns of U and the rows of V several times before doing so for the other one. After some simple computations, the explicit formulas of proximal points can be derived as follows. Let us consider the ith column of U, fixing the other columns of U and V. We have

$$\begin{aligned}
&\operatorname*{argmin}_{U_{:i}\geq 0} \sum \frac{1}{2}\left\| X - \sum_{q=1}^{i-1} U_{:q}V_{q:} - \sum_{q=i+1}^{r} U_{:q}V_{q:} - U_{:i}V_{i:}\right\|^2 + \frac{1}{2\beta_i}\left\|U_{:i} - \hat{U}_{:i}\right\|^2 \\
&\quad = \max\left(\frac{X(V_{i:})^T - (UV)(V_{i:})^T + U_{:i}V_{i:}(V_{i:})^T + 1/\beta_i \hat{U}_{:i}}{V_{i:}V_{i:}^T + 1/\beta_i}, 0\right).
\end{aligned}$$

A similar update for the rows of V can be derived by symmetry since $\|M - UV\|_F^2 = \|M^T - V^TU^T\|_F^2$.

In the upcoming experiments, we compare the following algorithms:

- A-HALS: the accelerated hierarchical alternating least squares algorithm in [23]. This is a block coordinate descent method on the columns of U and rows of V. A-HALS is a state-of-the-art NMF algorithm and outperforms standard projected gradient, the popular multiplicative updates and alternating nonnegative least squares (two-block coordinate descent optimizing U and V alternatively), see [27, 22].
- E-A-HALS: the acceleration version of A-HALS using extrapolation points proposed in [3]. We used the default values of the parameters. This is, as far as we know, one of the most efficient NMF algorithms. Note that E-A-HALS is a heuristic with no convergence guarantees. This is what initially motivated us to study the schemes IBP and IBPG.
- IBPG: the inertial proximal gradient method with $\tilde{\gamma} = 0.99$, $\breve{\alpha} = 1.01$ and $m = 1$ (that is, Algorithm 2.3 with $s = 2$ and cyclic updates of U and V).
- APGC: the accelerated proximal gradient coordinate descent method proposed in [48] which corresponds exactly to iMPG with $\tilde{\gamma} = \breve{\alpha} = 0.9999$.
- IBPG-A: the accelerated version of IBPG using the strategy of updating U several times before updating V, and vice versa (that is, Algorithm 2.3 with $s = 2$ and $m > 1$). For the value of m, we use exactly the same strategy as for A-HALS in [23] that is based on the computational cost of the first update of U (resp. V) compared to the next ones.
- IBP: the inertial coordinate-wise proximal point algorithm (that is, Algorithm 2.2 with $s = 2 \times \mathbf{r}$) using the strategy of alternatively updating the columns of U and rows V several times, as done in A-HALS. We choose $1/\beta_i^{(k,m)} = 0.001$ and $\alpha_i^{(k,m)} = \tilde{\alpha}^{(k)} = \min(\bar{\beta}, \gamma\tilde{\alpha}^{(k-1)})$, with $\bar{\beta} = 1$, $\gamma = 1.01$ and $\tilde{\alpha}^{(1)} = 0.6$. We can verify that this choice of parameters satisfies the conditions for the global convergence in Theorem 4.13.

The relative errors are defined by $\text{relerror}_k = \frac{\|X - \tilde{U}^{(k)}\tilde{V}^{(k)}\|_F}{\|X\|_F}$. We define $e_{\min} = 0$ for the experiments with low-rank synthetic data sets, and in the other experiments $e_{\min}$ is the lowest relative error obtained by any algorithms with any initializations. We define $E(k) = \text{relerror}_k - e_{\min}$. These are the same settings as in [23, 3].

All tests are preformed using Matlab R2015a on a laptop Intel CORE i7-8550U CPU @1.8GHz 16GB RAM. The code is available from https://doi.org/10.24433/CO.6813991.v1.

5.1. Experiments with synthetic data sets. We first perform experiments on synthetic data sets.

Low-rank synthetic data sets. Two low-rank matrices of size 200 × 200 and 200 × 500 are generated by letting $X = UV$, where U and V are generated by MATLAB commands $rand(\mathbf{m}, \mathbf{r})$ and $rand(\mathbf{r}, \mathbf{n})$ respectively, with $\mathbf{r} = 20$. For each matrix X, we run all algorithms with the same 50 random initializations $W_0 = rand(\mathbf{m}, \mathbf{r})$ and $V_0 = rand(\mathbf{r}, \mathbf{n})$, and for each initialization we run each algorithm for 20 seconds. Figure 1 illustrates the evolution of the average of $E(k)$ over 50 initializations with respect to time.

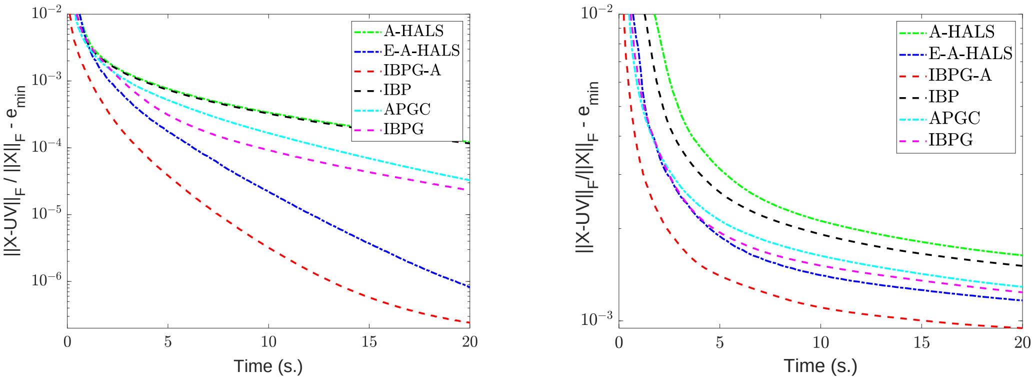

FIG. 1. *Average value of* $E(k)$ *with respect to time on 2 random low-rank matrices:* 200 × 200 *(left) and* 200 × 500 *(right).*

To compare the accuracy of the solutions, we generate 50 random matrices $\mathbf{m} \times \mathbf{n}$ with $\mathbf{m}$ and $\mathbf{n}$ being random integer numbers in the interval [200,500]. For each matrix X we run the algorithms for 20 seconds with 1 random initialization. Table 1 reports the average and standard deviation (std) of the errors. It also provides a ranking between the different algorithms: the ith entry of the ranking vector indicates how many times the corresponding algorithm obtained the ith best solution.

TABLE 1
Average, standard deviation and ranking of the value of $E(k)$ *at the last iteration among the different runs on the low-rank synthetic data sets. The best performance is highlighed in bold.*

Algorithm	mean ± std	ranking
A-HALS	$1.990\,10^{-3} \pm 7.910\ 10^{-4}$	(0, 0, 1, 3, 6, 40)
E-A-HALS	$1.486\,10^{-3} \pm 7.233\ 10^{-4}$	(13, 22, 6, 6, 3, 0)
IBPG-A	$\mathbf{1.081\,10^{-3}} \pm 6.012\ 10^{-4}$	(**34**, 14, 1, 1, 0, 0)
IBP	$1.916\,10^{-3} \pm 7.762\ 10^{-4}$	(0, 1, 5, 9, 34, 1)
APGC	$1.729\,10^{-3} \pm 7.452\ 10^{-4}$	(0, 4, 20, 16, 1, 9)
IBPG	$1.672\,10^{-3} \pm 7.313\ 10^{-4}$	(3, 9, 17, 15, 6, 0)

The two main observations are the following

- In terms of convergence speed and the final errors obtained, A-iMPG outperforms the other algorithms.
- Interestingly, APGC converges slower than iMPG and produces worse solutions. This illustrates the fact that using two extrapolated points allows a faster convergence.

Full-rank synthetic data sets. Two full-rank matrices of size 200×200 and 200×500 are generated by MATLAB command $X = rand(m, n)$. We take $\mathbf{r} = 20$. For each matrix

X, we run all algorithms with the same 50 random initializations $W_0 = rand(\mathbf{m}, \mathbf{r})$ and $V_0 = rand(\mathbf{r}, \mathbf{n})$, and for each initialization we run each algorithm for 20 seconds. Figure 2 illustrates the evolution of the average of $E(k)$ over 50 initializations with respect to time.

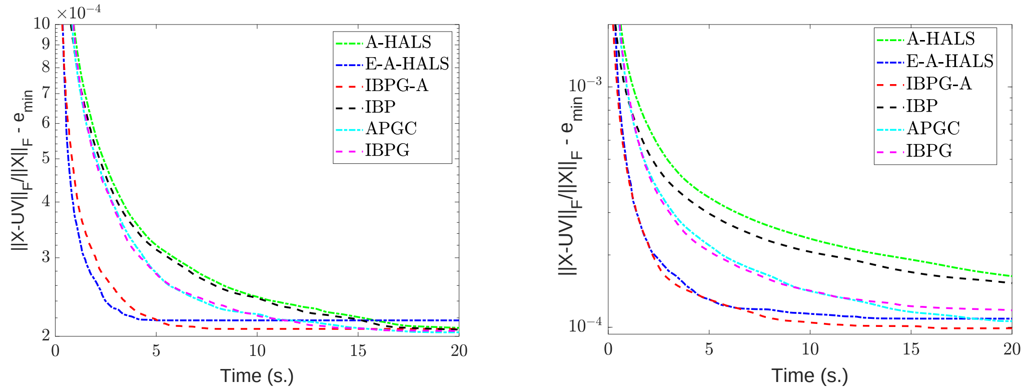

FIG. 2. *Average value of $E(k)$ with respect to time on 2 random full-rank matrices: 200×200 (left) and 200×500 (right).*

We then generate 50 full-rank matrices $X = rand(m, n)$, with $\mathbf{m}$ and $\mathbf{n}$ being random integer numbers in the interval [200,500]. For each matrix X, we run the algorithms for 20 seconds with a single random initialization. Table 2 reports the average, standard deviation (std) and ranking of the relative errors.

TABLE 2
Average, standard deviation and ranking of the value of $E(k)$ at the last iteration among the different runs on full-rank synthetic data sets. The best performance is highlighted in bold.

Algorithm	mean ± std	ranking
A-HALS	$0.450174 \pm 9.048\,10^{-3}$	(0, 2, 7, 6, 12, 23)
E-A-HALS	$0.450127 \pm 9.028\,10^{-3}$	(18, 8, 8, 9, 1, 6)
IBPG-A	$\mathbf{0.450115} \pm 9.050\,10^{-3}$	(**19**, 12, 7, 9, 2, 1)
IBP	$0.450168 \pm 9.048\,10^{-3}$	(1, 9, 7, 9, 23, 1)
APGC	$0.450146 \pm 9.056\,10^{-3}$	(4, 10, 9, 9, 3, 15)
IBPG	$0.450139 \pm 9.055\,10^{-3}$	(8, 9, 12, 8, 9, 4)

We observe the following:

- In both cases, A-iMPG and E-A-HALS have similar convergence rate, but A-iMPG converges to better solution than E-A-HALS more often. These algorithms outperform the others.
- iMPG performs better than APGC in terms of final error obtained, while the convergence speeds are similar.

5.2. Experiments with real data sets. In the experiment with real data sets, we will only keep the best performing algorithms, namely A-iMPG and E-A-HALS (as this algorithm is experimentally shown to outperform A-HALS and many other algorithms, see [3]), along with APGC for our observation purpose. For each data set, we generate 35 random initializations[2], and for each initialization we run each algorithm for 200 seconds.

Sparse document data sets. We test the algorithms on the same six sparse document data sets with $r = 10$ as in [3]. Figure 3 reports the evolution of the average of

[2] This allows us to run all experiments over one night on the considered laptop.

$E(k)$ over 35 initializations, and Table 3 reports the average error, standard deviation and ranking of the final value of $E(k)$ among the 210 runs (6 data sets with 35 initializations for each data set).

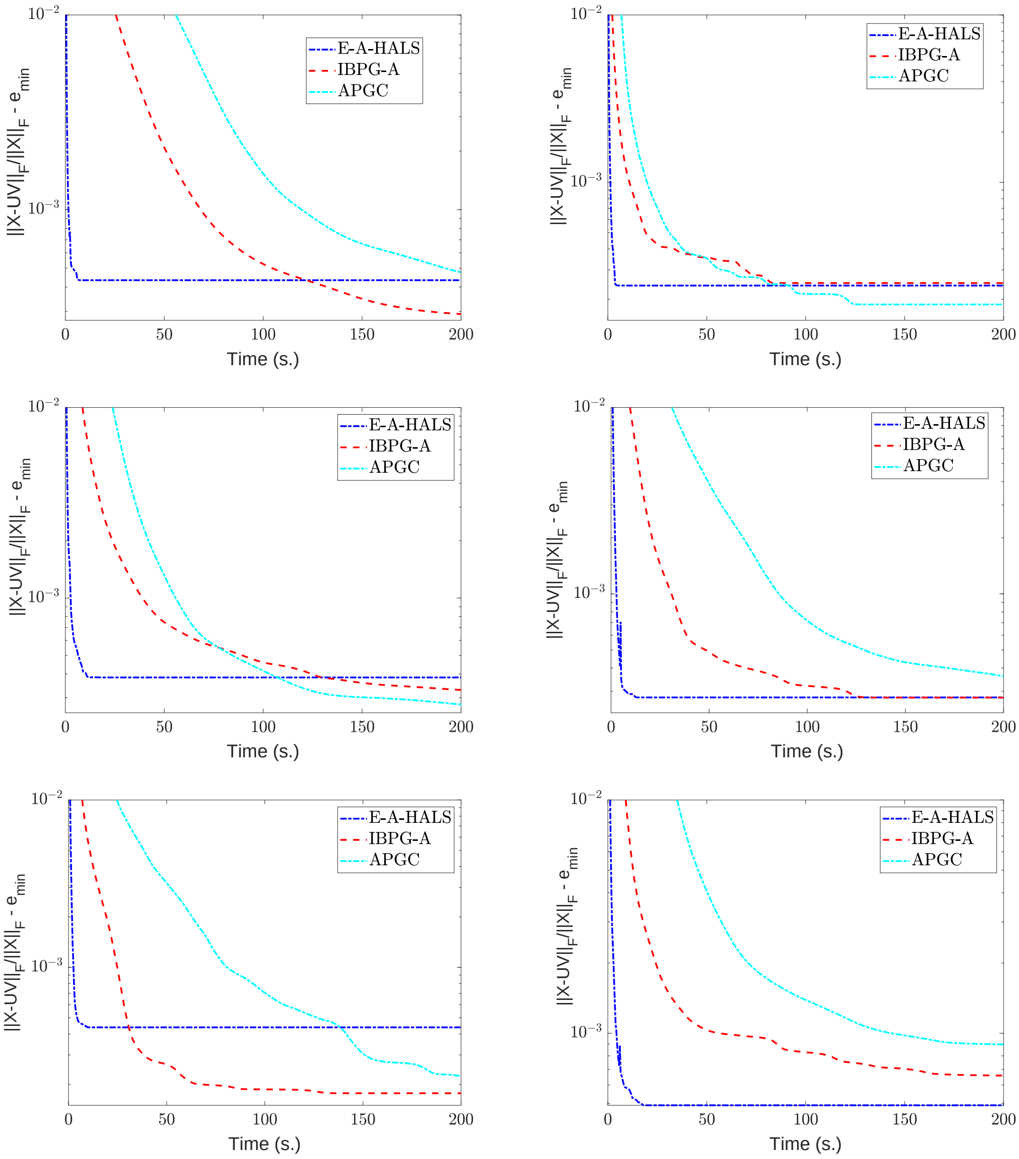

FIG. 3. *Average value of* $E(k)$ *with respect to time on 6 document data sets: Classic (top left), Hitech (top right), La1 (middle left), Ohscal (middle right), Reviews (bottom left) and Sports (bottom right).*

For these sparse datasets, E-A-HALS converges with the fastest rate, followed by A-iMPG. However, A-iMPG generates in average the best final solutions.

Dense hyperspectral images. In this experiment, we test the algorithms on two widely used hyperspectral images, namely the Urban and San Diego data sets; see, e.g., [24]. We choose the rank $\mathbf{r} = 10$. Figure 4 reports the evolution of the average value of $E(k)$, and Table 4 reports the average error, standard deviation and ranking of the final value of $E(k)$ among the 70 runs (2 data sets with 35 initializations for each data set).

Table 3

Average, standard deviation and ranking of the value of $E(k)$ at the last iteration among the different runs on the document data sets. The best performance is highlighted in bold.

Algorithm	mean ± std	ranking
E-A-HALS	$0.881969 \pm 3.021\ 10^{-2}$	(73, 55, 82)
IBPG-A	$\mathbf{0.881921} \pm 3.021\ 10^{-2}$	(**87**, 68, 55)
APGC	$0.881992 \pm 3.019\ 10^{-2}$	(51, 86, 73)

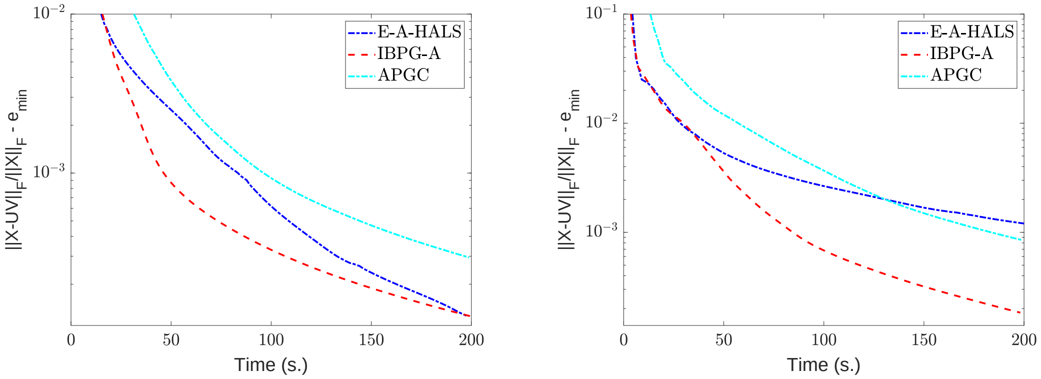

Fig. 4. *Average value of $E(k)$ with respect to time on 2 hyperspectral images: urban (the left) and SanDiego (the right).*

Table 4

Average error, standard deviation and ranking among the different runs for urban and SanDiego data sets.

Algorithm	mean ± std	ranking
E-A-HALS	$0.018823 \pm 6.739\ 10^{-4}$	(17, 28, 25)
IBPG-A	$\mathbf{0.018316} \pm 9.745\ 10^{-4}$	(**53**, 15, 2)
APGC	$0.018728 \pm 7.779\ 10^{-4}$	(0, 27, 43)

We see that A-iMPG outperforms E-A-HALS both in terms of convergence speed and accuracy.

6. Conclusions. We have analysed inertial versions of proximal BCD and proximal gradient BCD methods for solving a class of non-convex non-smooth composite optimization problems in the context of general Bregman distance. Our methods do not require restarting steps, and allow the use of randomized strategies and of two extrapolation points. We first proved convergence of a subsequence of the iterates to a critical point of F (Theorem 4.8) and then, under some additional assumptions, convergence of the whole sequence (Theorem 4.13). We showed that the proposed methods compared favourably with state-of-the-art algorithms for nonnegative matrix factorization.

Appendix A. Proofs of Propositions and Theorems.

A.1. Proof of Proposition 4.4.

Proof for IBP. (i) Applying Lemma 3.12 for (2.2) with $\beta = \beta_i^{(k,j)}$, $w = x_i^{(k,j-1)}$, $w^+ = x_i^{(k,j)}$, $\hat{w} = \hat{x}_i$ we have

$$\begin{aligned}
&F_i^{(k,j)}\left(x_i^{(k,j)}\right) + \tfrac{1}{\beta_i^{(k,j)}} D_i\left(x_i^{(k,j)}, x_i^{(k,j-1)}\right) \\
&\le F_i^{(k,j)}\left(x_i^{(k,j-1)}\right) + \tfrac{1}{\beta_i^{(k,j)}} \left\langle \nabla H_i\left(\hat{x}_i\right) - \nabla H_i\left(x_i^{(k,j-1)}\right), x_i^{(k,j)} - x_i^{(k,j-1)} \right\rangle \\
&\overset{(a)}{\le} F_i^{(k,j)}\left(x_i^{(k,j-1)}\right) + \tfrac{L_{H_i}}{\beta_i^{(k,j)}} \left\|\hat{x}_i - x_i^{(k,j-1)}\right\| \left\|x_i^{(k,j)} - x_i^{(k,j-1)}\right\| \\
&\overset{(b)}{\le} F_i^{(k,j)}\left(x_i^{(k,j-1)}\right) + \frac{\left(L_{H_i}\alpha_i^{(k,j)}\right)^2}{2\nu\sigma_i\beta_i^{(k,j)}} \left\|x_i^{(k,j-1)} - y_i\right\|^2 + \tfrac{\sigma_i\nu}{2\beta_i^{(k,j)}} \left\|x_i^{(k,j)} - x_i^{(k,j-1)}\right\|^2,
\end{aligned}$$

where we use the Lipschitz continuity of ∇H_i in (a), use (2.1) and the inequality $ab \le a^2/(2s) + sb^2/2$ in (b). Together with the inequality $D_i\left(w_1, w_2\right) \ge \frac{\sigma_i}{2}\left\|w_1 - w_2\right\|^2$ (see Lemma 3.7) and noting that $F\left(x^{(k,j)}\right) = F_i^{(k,j)}\left(x_i^{(k,j)}\right)$, we get

(A.1)
$$F\left(x^{(k,j)}\right) + \tfrac{\sigma_i(1-\nu)}{2\beta_i^{(k,j)}} \left\|x_i^{(k,j)} - x_i^{(k,j-1)}\right\|^2 \le F\left(x^{(k,j-1)}\right) + \frac{\left(L_{H_i}\alpha_i^{(k,j)}\right)^2}{2\nu\sigma_i\beta_i^{(k,j)}} \left\|x_i^{(k,j-1)} - y_i\right\|^2.$$

Note that y_i, $x_i^{(k,j-1)}$ and $x_i^{(k,j)}$ are 3 consecutive iterates of $\bar{x}_i^{(k,-1)}, \ldots, \bar{x}_i^{(k,d_i^k)}$. Summing up Inequality (A.1) for $j = 1$ to T_k, and combining with (4.1) we obtain

$$\begin{aligned}
&F\left(x^{(k,T_k)}\right) + \textstyle\sum_{i=1}^s \sum_{m=1}^{d_i^k} \delta\theta_i^{(k,m+1)} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\
&\le F\left(x^{(k,0)}\right) + \textstyle\sum_{i=1}^s \sum_{m=1}^{d_i^k} \theta_i^{(k,m)} \left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)}\right\|^2,
\end{aligned} \tag{A.2}$$

which implies

$$\begin{aligned}
&F\left(\tilde{x}^{(k)}\right) + \textstyle\sum_{i=1}^s \delta\theta_i^{(k,d_i^k+1)} \left\|\bar{x}_i^{(k,d_i^k)} - \bar{x}_i^{(k,d_i^k-1)}\right\|^2 \\
&\quad + \textstyle\sum_{i=1}^s \sum_{m=1}^{d_i^k-1} (\delta-1)\theta_i^{(k,m+1)} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\
&\le F\left(\tilde{x}^{(k-1)}\right) + \textstyle\sum_{i=1}^s \theta_i^{(k,1)} \left\|\tilde{x}_i^{(k-1)} - \left(\tilde{x}_{\text{prev}}^{(k-1)}\right)_i\right\|^2,
\end{aligned} \tag{A.3}$$

where $\sum_{i=a}^b (.)_i = 0$ if $a > b$. Note that $\left\|\bar{x}_i^{(k,d_i^k)} - \bar{x}_i^{(k,d_i^k-1)}\right\|^2 = \left\|\tilde{x}_i^{(k)} - \left(\tilde{x}_{\text{prev}}^{(k)}\right)_i\right\|^2$. Hence from (A.3) we get

$$\begin{aligned}
&F\left(\tilde{x}^{(k)}\right) + \delta \textstyle\sum_{i=1}^s \theta_i^{(k+1,1)} \left\|\tilde{x}_i^{(k)} - \left(\tilde{x}_{\text{prev}}^{(k)}\right)_i\right\|^2 \\
&\quad + \textstyle\sum_{i=1}^s \sum_{m=1}^{d_i^k-1} (\delta-1)\theta_i^{(k,m+1)} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\
&\le F\left(\tilde{x}^{(k-1)}\right) + \textstyle\sum_{i=1}^s \theta_i^{(k,1)} \left\|\tilde{x}_i^{(k-1)} - \left(\tilde{x}_{\text{prev}}^{(k-1)}\right)_i\right\|^2.
\end{aligned} \tag{A.4}$$

Summing up Inequality (A.4) from $k = 1$ to $k = K$ we obtain

$$\begin{aligned}
&F\left(\tilde{x}^{(K)}\right) + (\delta-1) \sum_{k=1}^K \sum_{i=1}^s \theta_i^{(k+1,1)} \left\|\tilde{x}_i^{(k)} - \left(\tilde{x}_{\text{prev}}^{(k)}\right)_i\right\|^2 \\
&\quad + \sum_{k=1}^K \sum_{i=1}^s \sum_{m=1}^{d_i^k-1} (\delta-1)\theta_i^{(k,m+1)} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\
&\le F\left(\tilde{x}^{(0)}\right) + \sum_{i=1}^s \theta_i^{(1,1)} \left\|\tilde{x}_i^{(0)} - \tilde{x}_i^{(-1)}\right\|^2.
\end{aligned} \tag{A.5}$$

Note that F is lower bounded and $\theta_i^{(k,m)} \geq W_1 > 0$. We deduce the result from (A.5).

(ii) We derive from Proposition 4.4 (i) that

$$\left\{ \left\| \tilde{x}^{(k)} - \tilde{x}^{(k)}_{\text{prev}} \right\| \right\}_{k \geq 1} \text{ and } \left\{ \sum_{m=1}^{d_i^k} \left\| \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\| \right\}_{k \geq 1} \text{ converge to } 0. \tag{A.6}$$

By Assumption 2.1, $d_i^k \geq 1$ and d_i^k is finite. We also note that $\bar{x}_i^{(k_n, d_i^{k_n})} = \tilde{x}_i^{(k_n)}$. Therefore, as $\sum_{m=1}^{d_i^{k_n}} \left\| \bar{x}_i^{(k_n,m)} - \bar{x}_i^{(k_n,m-1)} \right\| \to 0$, we deduce that $\left\{ \bar{x}_i^{(k_n,m)} \right\}_{m=0,\ldots,d_i^{k_n}}$ also converges to x_i^*. Then, let k in (A.6) be $k_n - 1$ and note that $\bar{x}_i^{(k_n,0)} = \tilde{x}_i^{(k_n-1)}$, $\bar{x}_i^{(k_n,-1)} = \left(\tilde{x}^{(k_n-1)}_{\text{prev}} \right)_i$. We thus have $\bar{x}_i^{(k_n,-1)} \to x_i^*$. At the k_n-th inner loop, we recall that $\hat{x}_i = \bar{x}_i^{(k_n,m-1)} + \bar{\alpha}_i^{(k_n,m)} \left(\bar{x}_i^{(k_n,m-1)} - \bar{x}_i^{(k_n,m-2)} \right)$. Hence $\hat{x}_i$ also converges to x_i^*. From (2.2), for all $x_i \in \mathbb{E}_i$ we have

$$\begin{aligned} & f\left(x^{(k_n,j)}\right) + r_i\left(\bar{x}_i^{(k_n,m)}\right) + \tfrac{1}{\bar{\beta}_i^{(k_n,m)}} D_i\left(\bar{x}_i^{(k_n,m)}, \hat{x}_i\right) \\ & \quad \leq f_i^{(k_n,j)}\left(x_i\right) + r_i\left(x_i\right) + \tfrac{1}{\bar{\beta}_i^{(k_n,m)}} D_i\left(x_i, \hat{x}_i\right). \end{aligned} \tag{A.7}$$

In (A.7), let $x_i = x_i^*$ and let $n \to \infty$ to get $\limsup_{n\to\infty} r_i\left(\bar{x}_i^{(k_n,m)}\right) \leq r_i\left(x_i^*\right)$. Furthermore, as r_i is lower semicontinuous, we have $\liminf_{n\to\infty} r_i\left(\bar{x}_i^{(k_n,m)}\right) \geq r_i\left(x_i^*\right)$. This completes the proof.

Proof for IBPG. (i) From the assumption that $\nabla f_i^{(k,j)}$ is $\bar{L}_i^{(k,m)}$-Lipschitz continuous, we have

$$\begin{aligned} f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right) \leq f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right) + \left\langle \nabla f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\rangle \\ + \frac{\bar{L}_i^{(k,m)}}{2} \left\| \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\|^2. \end{aligned} \tag{A.8}$$

Applying Lemma 3.12 with $\phi(w) = \left\langle \nabla f_i^{(k,j)}\left(\grave{x}_i\right), w - \bar{x}_i^{(k,m-1)} \right\rangle + r_i(w)$, $w^+ = \bar{x}_i^{(k,m)}$, $\hat{w} = \hat{x}_i$, and $w = \bar{x}_i^{(k,m-1)}$ we get

$$\begin{aligned} & \left\langle \nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\rangle + r_i\left(\bar{x}_i^{(k,m)}\right) + \frac{1}{\bar{\beta}_i^{(m)}} D_i\left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)}\right) \\ & \leq r_i\left(\bar{x}_i^{(k,m-1)}\right) + \frac{1}{\bar{\beta}_i^{(m)}} \left\langle \nabla H_i\left(\hat{x}_i\right) - \nabla H_i\left(\bar{x}_i^{(k,m-1)}\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\rangle. \end{aligned} \tag{A.9}$$

Note that $\frac{\sigma_i}{2} \left\| \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\|^2 \leq D_i\left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)}\right)$. From (A.8) and (A.9), we get

$$\begin{aligned} & f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right) + r_i\left(\bar{x}_i^{(k,m)}\right) \\ & \leq f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right) + \left\langle \nabla f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\rangle + r_i\left(\bar{x}_i^{(k,m)}\right) \\ & \qquad + \frac{\bar{L}_i^{(k,m)}}{\sigma_i} D_i\left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)}\right) \\ & \leq f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right) + \left\langle \nabla f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right) - \nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\rangle \\ & \qquad + \left(\frac{\bar{L}_i^{(k,m)}}{\sigma_i} - \frac{1}{\bar{\beta}_i^{(k,m)}} \right) D_i\left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)}\right) + r_i\left(\bar{x}_i^{(k,m-1)}\right) \end{aligned}$$

$$+\frac{1}{\bar{\beta}_i^{(k,m)}}\left\langle \nabla H_i\left(\hat{x}_i\right)-\nabla H_i\left(\bar{x}_i^{(k,m-1)}\right), \bar{x}_i^{(k,m)}-\bar{x}_i^{(k,m-1)}\right\rangle .$$

This implies

$$\begin{aligned}
& f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right)+r_i\left(\bar{x}_i^{(k,m)}\right)+\frac{(\kappa-1)\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m)}-\bar{x}_i^{(k,m-1)}\right\|^2 \\
& \leq f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right)+r_i\left(\bar{x}_i^{(k,m-1)}\right)+\bar{L}_i^{(k,m)}\left\|\grave{x}_i-\bar{x}_i^{(k,m-1)}\right\|\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\| \\
& \quad+\frac{\kappa L_{H_i}\bar{L}_i^{(k,m)}}{\sigma_i}\left\|\hat{x}_i-\bar{x}_i^{(k,m-1)}\right\|\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\| \\
& = f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right)+r_i\left(\bar{x}_i^{(k,m-1)}\right) \\
& \quad+\left(\bar{\gamma}_i^{(k,m)}+\frac{\kappa L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i}\right)\bar{L}_i^{(k,m)}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m-2)}\right\|\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|
\end{aligned}$$

Note that $x_i^{(k,j)}=\bar{x}_i^{(k,m)}$, $x_i^{(k,j-1)}=\bar{x}_i^{(k,m-1)}$. We apply the Young inequality to get

$$\begin{aligned}
& F\left(x^{(k,j)}\right)+\frac{(\kappa-1)\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2 \\
& \leq F\left(x^{(k,j-1)}\right)+\frac{\nu(\kappa-1)\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2 \\
& \quad+\frac{1}{2}\left(\bar{\gamma}_i^{(k,m)}+\frac{\kappa L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i}\right)^2\frac{\bar{L}_i^{(k,m)}}{\nu(\kappa-1)}\left\|\bar{x}_i^{(k,m-2)}-\bar{x}_i^{(k,m-1)}\right\|^2,
\end{aligned}$$

where $0<\nu<1$. We then have

$$\begin{aligned}
& F\left(x^{(k,j)}\right)+\frac{(1-\nu)(\kappa-1)\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2 \\
& \leq F\left(x^{(k,j-1)}\right)+\frac{1}{2}\left(\bar{\gamma}_i^{(k,m)}+\frac{\kappa L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i}\right)^2\frac{\bar{L}_i^{(k,m)}}{\nu(\kappa-1)}\left\|\bar{x}_i^{(k,m-2)}-\bar{x}_i^{(k,m-1)}\right\|^2 .
\end{aligned} \tag{A.10}$$

Summing up Inequality (A.10) from $j=1$ to T_k we obtain

$$\begin{aligned}
& F\left(x^{(k,T_k)}\right)+\sum_{i=1}^{s}\sum_{m=1}^{d_i^k}\frac{(1-\nu)(\kappa-1)\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2 \\
& \leq F\left(x^{(k,0)}\right)+\sum_{i=1}^{s}\sum_{m=1}^{d_i^k}\lambda_i^{(k,m)}\left\|\bar{x}_i^{(k,m-2)}-\bar{x}_i^{(k,m-1)}\right\|^2 .
\end{aligned}$$

Together with Condition (4.3), we see that this inequality is similar to (A.2). Hence, we can use the same technique as in the proof for IBP to obtain the result.

(ii) For all $x_i\in\mathbb{E}_i$, from (2.4) have

$$\begin{aligned}
& \left\langle\nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)}\right\rangle+r_i\left(\bar{x}_i^{(k,m)}\right)+\frac{1}{\bar{\beta}_i^{k,m}}D_i\left(\bar{x}_i^{(k,m)}, \hat{x}_i\right) \\
& \leq\left\langle\nabla f_i^{(k,j)}\left(\grave{x}_i\right), x_i\right\rangle+r_i\left(x_i\right)+\frac{1}{\bar{\beta}_i^{k,m}}D_i\left(x_i, \hat{x}_i\right).
\end{aligned} \tag{A.11}$$

Similarly to the proof for IBP, we can prove $\grave{x}_i\to x_i^*$, $\hat{x}_i\to x_i^*$; and consequently, by choosing $x_i=x_i^*$ in (A.11) we have $r_i\left(\bar{x}_i^{(k_n,m)}\right)\to r_i(x_i^*)$ as $n\to\infty$.

A.2. Proof of Theorem 4.11. We first prove that Φ is constant on the set $w\left(z^{(0)}\right)$ of all limit points of $\left\{z^{(k)}\right\}$. Indeed, from Condition (B1), we derive that $\Phi\left(z^{(k)}\right)$ is non-increasing. Together with the fact that it is bounded from below, we deduce that $\Phi\left(z^{(k)}\right)$ converges to some value $\bar{\Phi}$. Therefore, Condition (B4) shows that if $\bar{z} \in w\left(z^{(0)}\right)$ then $\Phi\left(\bar{z}\right) = \bar{\Phi}$.

Condition (B1) and the fact that Φ is bounded from below imply $\left\|z^{(k)} - z^{(k+1)}\right\| \to 0$. As proved in [14, Lemma 5], we then have $w\left(z^{(0)}\right)$ is connected and compact.

If there exists an integer $\bar{k}$ such that $\Phi\left(z^{(\bar{k})}\right) = \bar{\Phi}$ is trivial due to Condition (B1). Otherwise $\Phi\left(\bar{z}\right) < \bar{\Phi}$ for all $k > 0$. As $\Phi\left(z^{(k)}\right) \to \bar{\Phi}$, we derive that for any $\eta > 0$, there exists a positive integer k_0 such that $\Phi\left(z^{(k)}\right) < \Phi\left(\bar{z}\right) + \eta$ for all $k > k_0$. On the other hand, there exists a positive integer k_1 such that $\operatorname{dist}\left(z^{(k)}, w\left(z^{(0)}\right)\right) < \varepsilon$ for all $k > k_1$. Applying Lemma 3.5 we have

$$\xi'\left(\Phi\left(z^{(k)}\right) - \Phi\left(\bar{z}\right)\right) \operatorname{dist}\left(0, \partial\Phi\left(z^{(k)}\right)\right) \geq 1, \text{for any } k > l := \max\{k_0, k_1\}. \tag{A.12}$$

From Condition (B2) we get

$$\xi'\left(\Phi\left(z^{(k)}\right) - \Phi\left(\bar{z}\right)\right) \geq \frac{1}{\rho_3 \zeta_{k-1}}. \tag{A.13}$$

Denote $A_{i,j} = \xi\left(\Phi(z^{(i)}) - \Phi(\bar{z})\right) - \xi\left(\Phi(z^{(j)}) - \Phi(\bar{z})\right)$. From the concavity of ξ, Condition (B1) and Inequality (A.13) we obtain

$$A_{k,k+1} \geq \xi'\left(\Phi\left(z^{(k)}\right) - \Phi\left(\bar{z}\right)\right)\left[\Phi\left(z^{(k)}\right) - \Phi\left(z^{(k+1)}\right)\right] \geq \frac{\rho_2 \zeta_k^2}{\rho_3 \zeta_{k-1}}.$$

Hence we get $2\zeta_k \leq 2\sqrt{\frac{\rho_3}{\rho_2} A_{k,k+1} \zeta_{k-1}} \leq \zeta_{k-1} + \frac{\rho_3}{\rho_2} A_{k,k+1}$. Summing these inequalities from $k = l+1, \ldots, K$ we obtain

$$2\sum_{k=l+1}^{K} \zeta_k \leq \sum_{k=l}^{K-1} \zeta_k + \frac{\rho_3}{\rho_2} \sum_{k=l+1}^{K} A_{k,k+1} \leq \sum_{k=l+1}^{K} \zeta_k + \zeta_l + \frac{\rho_3}{\rho_2} A_{l+1,K+1}.$$

This implies that for all $K > l$ we have $\sum_{k=l+1}^{K} \zeta_k \leq \zeta_l + \frac{\rho_3}{\rho_2} \xi\left(\Phi\left(z^{(l+1)}\right) - \bar{\Phi}\right)$. Hence, $\sum_{k=1}^{\infty} \zeta_k < +\infty$. Condition (B1) then gives us $\sum_{k=1}^{\infty} \left\|z^{(k+1)} - z^{(k)}\right\| < \infty$. The whole sequence $\left\{z^{(k)}\right\}$ thus converges to some $\bar{z}$. Together with Condition (B2) and the closedness property of $\partial\Phi$, we have $0 \in \partial\Phi(\bar{z})$, that is, $\bar{z}$ is a critical point of Φ.

A.3. Proof of Proposition 4.12.

Proof for IBP. For all $k \geq 0$, we have $\left\|\tilde{x}^{(k)}\right\| \leq C_1$ (see Assumption 4.9) and $\left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\| \leq C_2$ (see Proposition 4.4). Furthermore, $\left\|\bar{x}_i^{(k,m)}\right\| \leq \left\|\bar{x}_i^{(k,0)}\right\| + \sum_{j=1}^{m} \left\|\bar{x}_i^{(k,j)} - \bar{x}_i^{(k,j-1)}\right\|$. Hence, $\left\|\bar{x}_i^{(k,m)}\right\| \leq C_1 + mC_2 \leq C_1 + (\bar{T} - s + 1)C_2$. In other words, the sequence $\left\{\bar{x}_i^{(k,m)}\right\}_{k\geq 0, m=1,\ldots,d_i^k}$ is bounded. Consequently, the sequence $\left\{x^{(k,j)}\right\}_{k\geq 0, j=1,\ldots,T_k}$ is bounded.

(i) We denote $\nabla_i f(x) = \nabla_{x_i} f(x)$ and let

$$\bar{q}_i^{(k,m)} = \frac{1}{\bar{\beta}_i^{(k,m)}}\left(\nabla H_i\left(\hat{x}_i\right) - \nabla H_i\left(\bar{x}_i^{(k,m)}\right)\right) + \nabla_i f\left(x^{(k,T_k)}\right) - \nabla f_i^{(k,j)}\left(x_i^{(k,j)}\right).$$

Let L_G is the Lipschitz constant of ∇f on the bounded set containing the sequence $\left\{x^{(k,j)}\right\}_{k\geq 0, j=1,\ldots,T_k}$. From (2.2) we get $\bar{q}_i^{(k,m)} \in \nabla_i f\left(\tilde{x}^{(k)}\right) + \partial r_i\left(\bar{x}_i^{(k,m)}\right)$. Also note that $\nabla f_i^{(k,j)}\left(x_i^{(k,j)}\right) = \nabla_i f\left(x^{(k,j)}\right)$. Hence,

$$\begin{aligned} \left\|\bar{q}_i^{(k,m)}\right\|^2 &\leq \frac{2L_H\left\|\hat{x}_i - \bar{x}_i^{(k,m)}\right\|^2}{\underline{\beta}} + 2L_G\left\|x^{(k,j)} - x^{(k,T)}\right\|^2 \\ &\leq \frac{4L_H}{\underline{\beta}}\left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m)}\right\|^2 + \frac{4L_H\overline{\alpha}^2}{\underline{\beta}}\left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)}\right\|^2 \\ &\qquad +2L_G\left\|x^{(k,j)} - x^{(k,T)}\right\|^2. \end{aligned} \tag{A.14}$$

We also note that

$$\begin{aligned} \left\|x^{(k,j)} - x^{(k,T)}\right\|^2 &= O\left(\textstyle\sum_{i=j}^{T-1}\left\|x^{(k,i)} - x^{(k,i+1)}\right\|^2\right) \\ &= O\left(\textstyle\sum_{i=1}^{s}\sum_{m=1}^{d_i^{(k)}}\left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2\right). \end{aligned}$$

Therefore, from (A.14) we deduce that $\left\|\bar{q}_i^{(k,m)}\right\|^2 = O\left(\varphi_{k-1}^2\right)$.

We now let $\bar{q}^{(k)} = \left(\bar{q}_1^{(k,d_1^k)}, \ldots, \bar{q}_s^{(k,d_s^k)}\right)$. Since $\bar{q}_i^{(k,d_i^k)} \in \nabla_i f\left(\tilde{x}^{(k)}\right) + \partial r_i\left(\tilde{x}_i^{(k)}\right)$, we have $\bar{q}^{(k)} \in \partial F\left(\tilde{x}^{(k)}\right)$ by Proposition 4.10. From $\left\|\bar{q}_i^{(k,d_i^k)}\right\|^2 = O\left(\varphi_{k-1}^2\right)$, we can easily obtain $\left\|\bar{q}^{(k)}\right\| = O\left(\varphi_{k-1}\right)$. Hence, there exists a positive number ρ_3 such that

$$\left\|\bar{q}^{(k+1)}\right\| \leq \rho_3\varphi_k. \tag{A.15}$$

Combined with Proposition 4.10(iii), we get the result.

(ii) From Inequality (A.2), we have

$$\begin{aligned} &F\left(\tilde{x}^{(k-1)}\right) + \tfrac{2W_2}{\sigma}\textstyle\sum_{i=1}^{s}\sum_{m=1}^{d_i^k} D_i\left(\bar{x}_i^{(k,m-1)}, \bar{x}_i^{(k,m-2)}\right) \\ &\geq F\left(\tilde{x}^{(k)}\right) + W_2\textstyle\sum_{i=1}^{s}\sum_{m=1}^{d_i^k}\left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)}\right\|^2 \\ &\geq F\left(\tilde{x}^{(k)}\right) + \delta W_1\textstyle\sum_{i=1}^{s}\sum_{m=1}^{d_i^k}\left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\ &\geq F\left(\tilde{x}^{(k)}\right) + \tfrac{2\delta W_1}{L_H}\textstyle\sum_{i=1}^{s}\sum_{m=1}^{d_i^k} D_i\left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)}\right). \end{aligned} \tag{A.16}$$

Denote

$$a_k = \sum_{i=1}^{s}\sum_{m=1}^{d_i^k} D_i\left(\bar{x}_i^{(k,m-1)}, \bar{x}_i^{(k,m-2)}\right) \text{ and } b_k = \sum_{i=1}^{s}\sum_{m=1}^{d_i^k} D_i\left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)}\right).$$

From (A.16) we get $F\left(\tilde{x}^{(k-1)}\right) + \frac{2W_2}{\sigma}a_k \geq F\left(\tilde{x}^{(k)}\right) + \frac{2\delta W_1}{L_H}b_k$. We thus obtain

$$\begin{aligned} F\left(\tilde{x}^{(k-1)}\right) + \rho a_k - F\left(\tilde{x}^{(k)}\right) - \rho b_k &\geq \left(\frac{\delta W_1}{L_H} - \frac{W_2}{\sigma}\right)(a_k + b_k) \\ &\geq \left(\frac{\delta W_1}{L_H} - \frac{W_2}{\sigma}\right)\frac{\sigma}{2}\varphi_k^2 = \rho_2\varphi_k^2. \end{aligned} \tag{A.17}$$

Note that $a_k - b_k = D\left(\tilde{x}^{(k-1)}, \tilde{x}_{\text{prev}}^{(k-1)}\right) - D\left(\tilde{x}^{(k)}, \tilde{x}_{\text{prev}}^{(k)}\right)$ and $a_k + b_k \geq \frac{\sigma}{2}\varphi_{k-1}^2$. Hence, from (A.17) we deduce that

$$F\left(\tilde{x}^{(k)}\right) + \rho D\left(\tilde{x}^{(k)}, \tilde{x}_{\text{prev}}^{(k)}\right) - F\left(\tilde{x}^{(k+1)}\right) - \rho D\left(\tilde{x}^{(k+1)}, \tilde{x}_{\text{prev}}^{(k+1)}\right) \geq \rho_2\varphi_k^2.$$

Together with (4.8) we obtain the result.

Proof for IBPG. (i) Let

$$\bar{q}_i^{(k,m)} = \frac{1}{\bar{\beta}_i^{(k,m)}} \left(\nabla H_i \left(\hat{x}_i \right) - \nabla H_i \left(\bar{x}_i^{(k,m)} \right) \right) + \nabla_i f \left(x^{(k,T_k)} \right) - \nabla f_i^{(k,j)} \left(\grave{x}_i \right).$$

From (2.4), we get $\bar{q}_i^{(k,m)} \in \nabla_i f \left(\tilde{x}^{(k)} \right) + \partial r_i \left(\bar{x}_i^{(k,m)} \right)$. Let us recall that the sequences $\left\{ \bar{x}_i^{(k,m)} \right\}_{k\ge 0, m=1,\ldots,d_i^k}$ and $\left\{ x^{(k,j)} \right\}_{k \ge 0, j=1,\ldots,T_k}$ are bounded. Furthermore, we have

$$\begin{aligned} \|\grave{x}_i\| &= \left\| \bar{x}_i^{(k,m-1)} + \bar{\gamma}_i^{(k,m)} \left(\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)} \right) \right\| \\ &\le \left\| \bar{x}_i^{(k,m-1)} \right\| + \bar{\gamma} \left\| \bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)} \right\|. \end{aligned}$$

Hence, $\grave{x}_i$ is also bounded. As a consequence, the value of $\grave{\mathbf{x}}$, which is formed by replacing the i-th block of $x^{(k,j-1)}$ by $\grave{x}_i = \bar{x}_i^{(k,m-1)} + \bar{\gamma}_i^{(k,m)} \left(\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)} \right)$, is also bounded. Let L_G be the Lipschitz constant of ∇f on the bounded set containing $x^{(k,j)}$ and $\grave{x}$. Note that $\nabla_i f \left(\grave{\mathbf{x}} \right) = \nabla f_i^{(k,j)} \left(\grave{x}_i \right)$. We have

$$\begin{aligned} &\left\| \nabla_i f \left(x^{(k,T_k)} \right) - \nabla f_i^{(k,j)} \left(\grave{x}_i \right) \right\|^2 \\ &= \left\| \nabla_i f \left(x^{(k,T_k)} \right) - \nabla_i f \left(x^{(k,j)} \right) + \nabla_i f \left(x^{(k,j)} \right) - \nabla_i f \left(\grave{\mathbf{x}} \right) \right\|^2 \\ &\le 2L_G \left\| x^{(k,T_k)} - x^{(k,j)} \right\|^2 + 2L_G \left\| x^{(k,j)} - \grave{\mathbf{x}} \right\|^2 \\ &= 2L_G \left\| x^{(k,T_k)} - x^{(k,j)} \right\|^2 + 2L_G \left\| \bar{x}_i^{(k,m)} - \grave{x}_i \right\|^2. \end{aligned}$$

We then continue with the same technique as in the proof for IBP to get the bound in (A.15).

(ii) The proof is follows exactly the same steps as for IBP.

Appendix B. Proof of Remarks.

B.1. Proof of Remark 4.5.

Applying Lemma 3.13 for (2.2) we have

$$\begin{aligned} &F_i^{(i,j)} \left(\bar{x}_i^{(k,m)} \right) + \tfrac{1}{\bar{\beta}_i^{(k,m)}} D_i \left(\bar{x}_i^{(k,m)}, \hat{x} \right) + \tfrac{1}{\bar{\beta}_i^{(k,m)}} D_i \left(\bar{x}_i^{(k,m-1)}, \bar{x}_i^{(k,m)} \right) \\ &\le F_i^{(i,j)} \left(\bar{x}_i^{(k,m)} \right) + \tfrac{1}{\bar{\beta}_i^{(k,m)}} D_i \left(\bar{x}_i^{(k,m-1)}, \hat{x} \right). \end{aligned}$$

Applying Lemma 3.8, we get

$$\begin{aligned} &D_i \left(\bar{x}_i^{(k,m)}, \hat{x}_i \right) - D_i \left(\bar{x}_i^{(k,m-1)}, \hat{x}_i \right) \\ &= D_i \left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)} \right) - \left\langle \nabla H_i \left(\hat{x}_i \right) - \nabla H_i \left(\bar{x}_i^{(k,m-1)} \right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\rangle. \end{aligned} \tag{B.1}$$

Therefore, we have

$$\begin{aligned} &F_i^{(k,j)} \left(\bar{x}_i^{(k,m)} \right) + \tfrac{2}{\bar{\beta}_i^{(k,m)}} D_i \left(\bar{x}_i^{(k,m-1)}, \bar{x}_i^{(k,m)} \right) \\ &\le F_i^{(k,j)} \left(\bar{x}_i^{(k,m)} \right) + \tfrac{1}{\bar{\beta}_i^{(k,m)}} \left\langle \nabla H_i \left(\hat{x}_i \right) - \nabla H_i \left(\bar{x}_i^{(k,m-1)} \right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\rangle \\ &\le F_i^{(k,j)} \left(\bar{x}_i^{(k,m)} \right) + L_{H_i} \tfrac{\bar{\alpha}_i^{(k,m)}}{\bar{\beta}_i^{(k,m)}} \left\| \bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)} \right\| \left\| \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\| \\ &\le F_i^{(k,j)} \left(\bar{x}_i^{(k,m)} \right) + \tfrac{\nu \sigma_i}{\bar{\beta}_i^{(k,m)}} \left\| \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)} \right\|^2 + \tfrac{\left(L_{H_i} \bar{\alpha}_i^{(k,m)} \right)^2}{4 \nu \sigma_i \bar{\beta}_i^{(k,m)}} \left\| \bar{x}_i^{(k,m-2)} - \bar{x}_i^{(k,m-1)} \right\|^2. \end{aligned}$$

We then obtain

$$\begin{aligned} & F\left(x^{(k,j)}\right) + \tfrac{(1-\nu)\sigma_i}{\bar{\beta}_i^{(k,m)}} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\ & \leq F\left(x^{(k,j-1)}\right) + \frac{\left(L_{H_i}\bar{\alpha}_i^{(k,m)}\right)^2}{4\nu\sigma_i\bar{\beta}_i^{(k,m)}} \left\|\bar{x}_i^{(k,m-2)} - \bar{x}_i^{(k,m-1)}\right\|^2 . \end{aligned} \tag{B.2}$$

We have obtained an inequality which is similar to (A.1). We therefore continue with the same technique as in the proof of Proposition 4.4 to get the result.

B.2. Proof of Remark 4.6. If r_i is convex then $\left\langle \nabla f_i^{(k,j)}\left(\grave{x}_i\right), w - \bar{x}_i^{(k,m-1)}\right\rangle + r_i(w)$ is also convex. Applying Lemma 3.13 for (2.4) we have

$$\begin{aligned} & \left\langle \nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\rangle + r_i(\bar{x}_i^{(k,m)}) + \tfrac{\bar{L}_i^{(k,m)}}{\sigma_i} D_i\left(\bar{x}_i^{(k,m)}, \hat{x}_i\right) \\ & \quad + \tfrac{\bar{L}_i^{(k,m)}}{\sigma_i} D_i\left(\bar{x}_i^{(k,m-1)}, \bar{x}_i^{(k,m)}\right) \\ & \leq r_i(\bar{x}_i^{(k,m-1)}) + \tfrac{\bar{L}_i^{(k,m)}}{\sigma_i} D_i\left(\bar{x}_i^{(k,m-1)}, \hat{x}_i\right) . \end{aligned} \tag{B.3}$$

Together with (A.8) we have

$$\begin{aligned} & f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right) + r_i\left(\bar{x}_i^{(k,m)}\right) + \tfrac{\bar{L}_i^{(k,m)}}{\sigma_i} D_i\left(\bar{x}_i^{(k,m)}, \hat{x}_i\right) \\ & \leq f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right) + r_i\left(\bar{x}_i^{(k,m-1)}\right) + \tfrac{\bar{L}_i^{(k,m)}}{\sigma_i} D_i\left(\bar{x}_i^{(k,m-1)}, \hat{x}_i\right) \\ & \quad + \left\langle \nabla f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right) - \nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\rangle . \end{aligned} \tag{B.4}$$

Together with (B.1) we obtain

$$\begin{aligned} & F\left(x^{(k,j)}\right) + \tfrac{\bar{L}_i^{(k,m)}}{\sigma_i} D\left(\bar{x}_i^{(k,m)}, \bar{x}_i^{(k,m-1)}\right) \\ & \leq F\left(x^{(k,j-1)}\right) + \tfrac{\bar{L}_i^{(k,m)}}{\sigma_i} \left\langle \nabla H_i\left(\hat{x}_i\right) - \nabla H_i\left(\bar{x}_i^{(k,m-1)}\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\rangle \\ & \quad + \left\langle \nabla f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right) - \nabla f_i^{(k,j)}\left(\grave{x}_i\right), \bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\rangle , \end{aligned}$$

from which we have

$$\begin{aligned} & F\left(x^{(k,j)}\right) + \tfrac{\bar{L}_i^{(k,m)}}{2} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\ & \leq F\left(x^{(k,j-1)}\right) + \tfrac{\bar{L}_i^{(k,m)} L_{H_i}}{\sigma_i} \bar{\alpha}_i^{(k,m)} \left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)}\right\| \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\| \\ & \quad + \bar{L}_i^{(k,m)} \bar{\gamma}_i^{(k,m)} \left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)}\right\| \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\| \\ & \leq F\left(x^{(k,j-1)}\right) + \tfrac{\nu\bar{L}_i^{(k,m)}}{2} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\ & \quad + \tfrac{1}{2} \left(\tfrac{L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i} + \bar{\gamma}_i^{(k,m)}\right)^2 \tfrac{\bar{L}_i^{(k,m)}}{\nu} \left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)}\right\|^2 , \end{aligned}$$

where $0 < \nu < 1$. Therefore, we have

$$\begin{aligned} & F\left(x^{(k,j)}\right) + \tfrac{(1-\nu)\bar{L}_i^{(k,m)}}{2} \left\|\bar{x}_i^{(k,m)} - \bar{x}_i^{(k,m-1)}\right\|^2 \\ & \leq F\left(x^{(k,j-1)}\right) + \tfrac{1}{2} \left(\tfrac{L_{H_i}\bar{\alpha}_i^{(k,m)}}{\sigma_i} + \bar{\gamma}_i^{(k,m)}\right)^2 \tfrac{\bar{L}_i^{(k,m)}}{\nu} \left\|\bar{x}_i^{(k,m-1)} - \bar{x}_i^{(k,m-2)}\right\|^2 . \end{aligned} \tag{B.5}$$

We get a similar inequality with (A.10). Summing up Inequality (B.5) from $j = 1$ to T_k and continuing with the same techniques as in the proof of Proposition 4.4, we get the result.

B.3. Proof of Remark 4.7. Using the technique in [48, Lemma 2.1], we first prove that

(B.6)
$$\begin{aligned}&F\left(x^{(k,j)}\right)+\tfrac{(1-\nu)\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2\\&\leq F\left(x^{(k,j-1)}\right)+\left(\left(\bar{\gamma}_i^{(k,m)}\right)^2+\tfrac{\left(\bar{\gamma}_i^{(k,m)}-\bar{\alpha}_i^{(k,m)}\right)^2}{\nu}\right)\tfrac{\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m-2)}\right\|^2.\end{aligned}$$

Indeed, we derive from (2.4) that

$$\left\langle\nabla f_i^{(k,j)}\left(\grave{x}_i\right)+\bar{g}_i^{(k,m)}+\frac{\nabla H_i\left(\bar{x}_i^{(k,m)}\right)-\nabla H_i\left(\hat{x}_i\right)}{\bar{\beta}_i^{(k,m)}},\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\rangle\geq 0, \tag{B.7}$$

where $\bar{g}_i^{(k,m)}\in\partial r_i\left(\bar{x}_i^{(k,m)}\right)$. On the other hand, since r_i is convex and $f_i^{(k,j)}$ is $\bar{L}_i^{(k,m)}$-smooth , we have $r_i\left(\bar{x}_i^{(k,m-1)}\right)-r_i\left(\bar{x}_i^{(k,m)}\right)\geq\left\langle\bar{g}_i^{(k,m)},\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\rangle$, and

$$f_i^{(k,j)}\left(\grave{x}_i\right)-f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right)+\frac{\bar{L}_i^{(k,m)}}{2}\left\|\grave{x}_i-\bar{x}_i^{(k,m)}\right\|^2\geq\left\langle\nabla f_i^{(k,j)}\left(\grave{x}_i\right),\grave{x}_i-\bar{x}_i^{(k,m)}\right\rangle.$$

Together with (B.7) we have

$$\begin{aligned}&r_i\left(\bar{x}_i^{(k,m-1)}\right)-r_i\left(\bar{x}_i^{(k,m)}\right)-f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right)+f_i^{(k,j)}\left(\grave{x}_i\right)+\frac{\bar{L}_i^{(k,m)}}{2}\left\|\grave{x}_i-\bar{x}_i^{(k,m)}\right\|^2\\&\geq\frac{1}{\bar{\beta}_i^{(k,m)}}\left\langle\nabla H_i\left(\hat{x}_i\right)-\nabla H_i\left(\bar{x}_i^{(k,m)}\right),\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\rangle-\left\langle\nabla f_i^{(k,j)}\left(\grave{x}_i\right),\bar{x}_i^{(k,m-1)}-\grave{x}_i\right\rangle.\end{aligned}$$

We then apply the convexity property of $f_i^{(k,j)}$ to obtain

$$\begin{aligned}&r_i\left(\bar{x}_i^{(k,m)}\right)+f_i^{(k,j)}\left(\bar{x}_i^{(k,m)}\right)+\tfrac{1}{\bar{\beta}_i^{(k,m)}}\left\langle\nabla H_i\left(\hat{x}_i\right)-\nabla H_i\left(\bar{x}_i^{(k,m)}\right),\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\rangle\\&\leq r_i\left(\bar{x}_i^{(k,m-1)}\right)+f_i^{(k,j)}\left(\bar{x}_i^{(k,m-1)}\right)+\tfrac{\bar{L}_i^{(k,m)}}{2}\left\|\grave{x}_i-\bar{x}_i^{(k,m)}\right\|^2.\end{aligned}$$

Note that $H_i(x_i)=\|x_i\|^2/2$. We then have

$$\begin{aligned}&F\left(x^{(k,j)}\right)+\tfrac{1}{\bar{\beta}_i^{(k,m)}}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2\\&\leq F\left(x^{(k,j-1)}\right)+\tfrac{\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2+\tfrac{\bar{L}_i^{(k,m)}\left(\bar{\gamma}_i^{(k,m)}\right)^2}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m-2)}\right\|^2\\&\quad+\bar{L}_i^{(k,m)}\left(\bar{\gamma}_i^{(k,m)}-\bar{\alpha}_i^{(k,m)}\right)\left\langle\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m-2)},\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\rangle,\end{aligned}$$

which implies that

$$\begin{aligned}&F\left(x^{(k,j)}\right)+\tfrac{\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2\\&\leq F\left(x^{(k,j-1)}\right)+\tfrac{\bar{L}_i^{(k,m)}}{2}\left(\bar{\gamma}_i^{(k,m)}\right)^2\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m-2)}\right\|^2\\&+\nu\tfrac{\bar{L}_i^{(k,m)}}{2}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m)}\right\|^2+\tfrac{\bar{L}_i^{(k,m)}\left(\bar{\gamma}_i^{(k,m)}-\bar{\alpha}_i^{(k,m)}\right)^2}{2\nu}\left\|\bar{x}_i^{(k,m-1)}-\bar{x}_i^{(k,m-2)}\right\|^2.\end{aligned}$$

Hence we get Inequality (B.6). In other words, we have a similar inequality with (A.10). We then continue with the same techniques as in the proof of Proposition 4.4 to get the result.

REFERENCES

[1] M. Aharon, M. Elad, A. Bruckstein, et al. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311, 2006.

[2] M. Ahookhosh, A. Themelis, and P. Patrinos. Bregman forward-backward splitting for nonconvex composite optimization: superlinear convergence to nonisolated critical points. *arXiv preprint arXiv:1905.11904*, 2019.

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

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

[5] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, Feb 2013.

[6] A. Auslender. Asymptotic properties of the Fenchel dual functional and applications to decomposition problems. *Journal of Optimization Theory and Applications*, 73(3):427–449, Jun 1992.

[7] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[8] A. B. and L. T. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23:2037–2060, 2013.

[9] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[10] H. H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Space*. Springer, 2011.

[11] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265 – 274, 2009.

[12] J. Bochnak, M. Coste, and M.-F. Roy. *Real Algebraic Geometry*. Springer, 1998.

[13] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

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

[15] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[16] R. I. Boţ and E. R. Csetnek. An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems. *Journal of Optimization Theory and Applications*, 171(2):600–616, Nov 2016.

[17] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[18] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[19] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[20] O. Fercoq and P. Richtarik. Accelerated, parallel, and proximal coordinate descent. *SIAM J. Optim.*, 25(4):1997–2023, 2015.

[21] X. Fu, K. Huang, N. D. Sidiropoulos, and W. Ma. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Processing Magazine*, 2018. to appear.

[22] N. Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257):257–291, 2014.

[23] N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4):1085–1105, 2012.

[24] N. Gillis, D. Kuang, and H. Park. Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):2066–2078, 2015.

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

[26] C. Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85, 1957.

[27] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[28] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48(3):769–783, 1998.

[29] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[30] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2), 1983.

[31] Y. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonom. i. Mat. Metody*, 24:509–517, 1998.

[32] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publ., 2004.

[33] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Prog.*, 103(1):127–152, 2005.

[34] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

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

[37] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

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

[39] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1 – 17, 1964.

[40] M. J. D. Powell. On search directions for minimization algorithms. *Mathematical Programming*, 4(1):193–201, Dec 1973.

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

[42] D. Reem, S. Reich, and A. D. Pierro. Re-examination of Bregman functions and new properties of their divergences. *Optimization*, 68(1):279–348, 2019.

[43] M. Teboulle. Convergence of proximal-like algorithms. *SIAM J. Optim.*, 7(4):1069–1083, 1997.

[44] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, Jul 2018.

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

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

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

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

[49] Y. Xu and W. Yin. A fast patch-dictionary method for whole image recovery. *Inverse Problems & Imaging*, 10:563, 2016.

[50] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, Aug 2017.

[51] W. Zangwill. *Nonlinear programming: a unified approach*. Prentice-Hall, 1969.

[52] S. Zavriev and F. Kostyuk. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 1993.