AN INERTIAL BLOCK MAJORIZATION MINIMIZATION FRAMEWORK FOR NONSMOOTH NONCONVEX OPTIMIZATION*

LE THI KHANH HIEN†, DUY NHAT PHAN‡, AND NICOLAS GILLIS†

Abstract. In this paper, we introduce TITAN, a novel inerTial block majorIzation minimization framework for non-smooth non-convex opTimizAtioN problems. TITAN is a block coordinate method (BCM) that embeds inertial force to each majorization-minimization step of the block updates. The inertial force is obtained via an extrapolation operator that subsumes heavy-ball and Nesterov-type accelerations for block proximal gradient methods as special cases. By choosing various surrogate functions, such as proximal, Lipschitz gradient, Bregman, quadratic, and composite surrogate functions, and by varying the extrapolation operator, TITAN produces a rich set of inertial BCMs. We study sub-sequential convergence as well as global convergence for the generated sequence of TITAN. We illustrate the effectiveness of TITAN on two important machine learning problems, namely sparse non-negative matrix factorization and matrix completion.

Key words. inertial method, block coordinate method, majorization minimization, surrogate functions, sparse non-negative matrix factorization, matrix completion

1. Introduction.

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

$$\begin{aligned} \min_{x} \quad & F(x) := f(x_1, \ldots, x_m) + \sum_{i=1}^{m} g_i(x_i) \\ \text{such that} \quad & x_i \in \mathcal{X}_i \text{ for } i = 1, \ldots, m, \end{aligned} \tag{1}$$

where $\mathcal{X}_i \subseteq \mathbb{E}_i$ is a closed convex set of a finite dimensional real linear space $\mathbb{E}_i$, x can be decomposed into m blocks $x = (x_1, \ldots, x_m)$ with $x_i \in \mathcal{X}_i$, $f(\cdot)$ is a non-smooth non-convex function, and $g_i(\cdot)$ is a proper and lower semi-continuous function (possibly with extended values). We will denote by $\mathcal{X} := \prod_{i=1}^{m} \mathcal{X}_i$ the feasible set. Throughout this paper, we assume that F is bounded from below and[1]

$$\partial F(x) = \{\partial_{x_1} F(x)\} \times \ldots \times \{\partial_{x_m} F(x)\} \quad \text{for all} \quad x \in \mathcal{X}, \tag{2}$$

where $\partial F(x)$ denotes the limiting subdifferential of F at x (see Appendix A).

1.1. Related works.

Block Coordinate Descent Methods and BSUM. Block coordinate descent (BCD) method is a standard approach to solve the non-smooth non-convex problem (1). Starting with a given initial point, BCD cyclically updates one block of variables at a time while fixing the values of the other blocks. Typically, there are three main types of BCD methods: classical BCD [17, 20, 39, 42], proximal BCD [17, 40, 44], and proximal gradient BCD [8, 12, 40, 43]. Fixing x_j for $j \in \{1, \ldots, m\} \setminus \{i\}$, let us call the function $x_i \mapsto f(x)$ a block i function of f. The classical BCD methods

* **Funding:** L. T. K. Hien and N. Gillis are supported by the Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO) under EOS project no 30468160 (SeLMA), and by the European Research Council (ERC starting grant 679515).

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

‡ Institute of Research and Development, Duy Tan University, Danang 550000, Vietnam (phanduynhat@duytan.edu.vn).

[1] When f is a sum of a continuously differentiable function and a block separable function that is subdifferential then [5, Proposition 2.1] shows that $\partial F(x) = \{\partial_{x_1} F(x)\} \times \ldots \times \{\partial_{x_m} F(x)\}$.

alternatively minimize the block i functions of the objective. These methods fail to converge for some non-convex problems, see for example [39]. The proximal BCD methods improve the classical BCD methods by coupling the block i objective functions with a proximal term. Considering Problem (1) with $m = 2$, the authors in [5] proved the global convergence of the generated sequence of the proximal BCD methods to a critical point of F, which is assumed to satisfy the Kurdyka-Łojasiewicz (KL) property [25, 11]. The 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 (which can involve Bregman divergences) on the non-smooth part g. Using the KL property of F, the authors in [12] proved the global convergence of the proximal gradient BCD for solving Problem (1) when each block function of f is assumed to be Lipschitz smooth, and the authors in [2, 18, 41] prove the global convergence when the block functions are relative smooth [7, 27].

These BCD methods belong to a more general framework that was proposed by Razaviyayn, Hong and Luo [40], and named the block successive upper-bound minimization algorithm (BSUM). BSUM for one block problem is closely related to the majorization-minimization algorithm. BSUM updates one block i of x by minimizing an upper-bound approximation function (also known as a majorized or a surrogate function) of the corresponding block i objective function. BSUM recovers proximal BCD when the proximal surrogate functions are chosen, and it recovers proximal gradient BCD when the Lipschitz gradient surrogate or Bregman surrogate functions are chosen, see Section 4 and [28] for examples of surrogate functions. Considering the non-smooth non-convex Problem (1) with $g = 0$, the authors in [40] established sub-sequential convergence for the generated sequence of BSUM under some suitable assumptions. When f and g are convex functions, the iteration complexity of BSUM with respect to the optimality gap $F(x^k) - F(x^*)$, where x^* is the optimal solution of (1), was studied in [21]. We note that global convergence for the generated sequence of BSUM for solving non-smooth non-convex Problem (1) was not studied in [40].

Inertial methods. In the convex setting, the gradient descent (GD) method is known to have suboptimal convergence rate. To accelerate the convergence of the GD method, Polyak, for the first time, proposed the heavy ball method (see [38]), which adds an inertial force to the gradient direction using $\alpha^k(x^k - x^{k-1})$, where x^k is the current iterate, x^{k-1} is the previous iterate, and α^k is an extrapolation parameter. Later, in a series of works [29, 30, 31, 32], Nesterov proposed the well-known accelerated fast gradient methods. While extrapolation is not used to calculate the gradients in the heavy ball method, Nesterov acceleration uses it to evaluate the gradients as well as adding the inertial force. The spirit of using inertial terms to accelerate first-order methods has been brought to non-convex problems. In the non-convex setting, the heavy ball acceleration type was used in [46, 34, 33], the Nesterov acceleration type was used in [44, 45]. Interestingly, using two different extrapolation points (one is for evaluating gradients and another one is for adding the inertial force) was also considered in [37, 19]. Sub-sequential and global convergence of some specific inertial BCD methods for non-convex problems have been established when F is assumed to have the KL property, see e.g., [3, 19, 33, 44, 45]. To the best of our knowledge, applying acceleration strategies to the general framework BSUM has not been studied in the literature.

1.2. Contribution. First, we propose TITAN, a novel inertial block majorization minimization framework for solving the non-smooth non-convex problem (1). TITAN updates one block of x at a time by choosing a surrogate function (see Def-

inition 2.1 and Section 4) for the corresponding block objective function, embedding inertial force to this surrogate function and then minimizing the obtained inertial surrogate function. The novelty of TITAN lies in how we control the inertial force. Specifically, we use an extrapolation operator that can be wisely chosen depending on specific assumptions considered for Problem (1) to produce various types of acceleration; see Section 4 for examples. For instances, considering Problem (1) when each block function of f is Lipschitz smooth, our extrapolation operator recovers heavy ball acceleration type, Nesterov acceleration type, the acceleration type in [37, 19], and the type of inertial gradient algorithm with Hessian damping in [1].

Then, we study sub-sequential convergence as well as global convergence for TITAN, which in turns unifies the convergence analysis of many acceleration algorithms that TITAN subsumes. TITAN can be thought of as BSUM with extrapolation. However, it is important noting that Problem (1) is more general than the problem considered for BSUM in [40]. In particular, the objective function of Problem (1) includes a separable non-smooth function g that is very important to model the regularizers of many practical optimization problems, while the objective function in [40] excludes g. Hence, [40, Assumption 2 (B4)] on the continuity of the surrogate functions on the joint variables is violated for Problem (1); and as such the analysis in [40] is not applicable to Problem (1). Furthermore, when no extrapolation is applied and $g = 0$, TITAN becomes BSUM. Hence, the global convergence established for TITAN with suitable assumptions can be applied to derive the global convergence for BSUM, which was not studied in [40].

Finally, we illustrate the effectiveness of TITAN on two applications, namely sparse non-negative matrix factorization (sparse NMF) and the matrix completion problem (MCP). Applying TITAN to sparse NMF illustrates the benefit of using inertial terms in BCD methods. The deployment of TITAN in solving MCP illustrates the advantages of using suitable surrogate functions. Specifically, we will use a composite surrogate function for the MCP. Compared to the typical proximal gradient BCD method, each minimization step of TITAN has a closed-form solution while each proximal gradient step does not. In our experiments, TITAN outperforms the proximal gradient BCD method.

1.3. Organization of the paper. In the next section, we present TITAN with cyclic block update rule. In Section 3, we establish the subsequential and global convergence for TITAN. In Section 4, we employ various surrogate functions and wisely choose the extrapolation operators to derive specific accelerated BCMs. We extend TITAN to allow essentially cyclic rule in choosing the block to update in Section 5. In Section 6, we report the numerical results of TITAN applied on the sparse NMF and the MCP. We conclude the paper in Section 7.

2. Inertial Block Alternating Majorization Minimization. In this section, we introduce TITAN. Let us describe TITAN, an inertial block alternating majorization-minimization framework; see Algorithm 1. At the k-th iteration, we cyclically update one block i at a time while fixing the values of the other blocks. In Algorithm 1 and throughout the paper, we use the notation

$$x^{k,i} = (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k) \text{ for } i = 1, \ldots, m, \quad \text{and } x^{k+1} = x^{k,m}.$$

To update block i at the k-th iteration, we first need to choose a block i surrogate function u_i of f, which is defined in the following. Examples of such functions, and a discussion on their use in the context of TITAN are provided in Section 4.

Algorithm 1 TITAN with cyclic update to solve Problem (1)

Input: Choose $x^{-1}, x^0 \in \mathcal{X}$ (x^{-1} can be chosen equal to x^0).
Output: x^k that approximately solves (1).

1: **for** $k = 0, 1, \ldots$ **do**
2: Set $x^{k,0} = x^k$
3: **for** $i = 1, ..., m$ **do**
4: Choose a block i surrogate function u_i of f and an extrapolation $\mathcal{G}_i^k(x_i^k, x_i^{k-1})$.
5: Update block i by

$$x_i^{k,i} \in \operatorname*{argmin}_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i), \tag{3}$$

and set $x_j^{k,i} = x_j^{k,i-1}$ for all $j \neq i$.
6: **end for**
7: Set $x^{k+1} = x^{k,m}$.
8: **end for**

DEFINITION 2.1 (Block surrogate function). *A function $u_i : \mathcal{X}_i \times \mathcal{X} \to \mathbb{R}$ is called a block i surrogate function of f if the following conditions are satisfied:*
(a) $u_i(y_i, y) = f(y)$ for all $y \in \mathcal{X}$,
(b) $u_i(x_i, y) \geq f(x_i, y_{\neq i})$ for all $x_i \in \mathcal{X}_i$ and $y \in \mathcal{X}$, where

$$f(x_i, y_{\neq i}) := f(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_m).$$

The block approximation error is defined as $h_i(x_i, y) := u_i(x_i, y) - f(x_i, y_{\neq i})$.

Then, we solve the sub-problem (3) in which the block surrogate function is equipped with an inertial force operated by an extrapolation operator $\mathcal{G}_i^k$. Examples of such operators are discussed in Section 4.

Conditions for $\mathcal{G}_i^k$ and u_i. TITAN must choose u_i and $\mathcal{G}_i^k$ such that Step 5 generates x_i^{k+1} (note that $x_i^{k+1} = x_i^{k,i}$) satisfying the following nearly sufficiently decreasing property (NSDP) (see the discussion below for more details)

$$F(x^{k,i-1}) + \frac{\gamma_i^{(x^{k,i-1})}}{2}\|x_i^k - x_i^{k-1}\|^2 \geq F(x^{k,i}) + \frac{\eta_i^{(x^{k,i-1})}}{2}\|x_i^{k+1} - x_i^k\|^2, k = 0, 1, \ldots \tag{4}$$

where[2] $\gamma_i^{(x^{k,i-1})}$ and $\eta_i^{(x^{k,i-1})}$ depend on the extrapolation parameters used in $\mathcal{G}_i^k$ and the parameters used in u_i. For notation succinctness, we denote $\gamma_i^k = \gamma_i^{(x^{k,i-1})}$ and $\eta_i^k = \eta_i^{(x^{k,i-1})}$. The parameters of TITAN must also satisfy the following assumptions.

ASSUMPTION 2.2. *(A) For $i \in [m]$, where $[m] := \{1, \ldots, m\}$, the block i surrogate function $u_i(x_i, y)$ is continuous in y and lower semi-continuous in x_i.*
(B) For $i \in [m]$, given $y \in \mathcal{X}$, there exists a function $x_i \mapsto \bar{h}_i(x_i, y)$ such that $\bar{h}_i(\cdot, y)$ is continuously differentiable at y_i and $\nabla_{x_i}\bar{h}_i(y_i, y) = 0$, and the block approximation error $x_i \mapsto h_i(x_i, y)$ satisfies

$$h_i(x_i, y) \leq \bar{h}_i(x_i, y) \ \text{ for all } \ x_i \in \mathcal{X}_i. \tag{5}$$

[2]We use the upperscript to mean that the parameters depend on the current point $x^{k,i-1}$.

Assumption 2.2(A) is weaker than [40, Assumption 2(B4)] of BSUM, which imposes the continuity of u_i on the joint variable (x_i, y). The following lemma provides some sufficient conditions for Assumption 2.2(B). Lemma 2.3 will be used to verify Assumption 2.2 for the block surrogate functions that will be given in Section 4.

LEMMA 2.3. *Assumption 2.2 is satisfied when one of the following two conditions holds:*

- *the block error $h_i(\cdot, y)$ is continuously differentiable at y_i and $\nabla_{x_i} h_i(y_i, y) = 0$,*
- *$h_i(x_i, y) \leq \upsilon_i \|x_i - y_i\|^{1+\epsilon_i}$ for some $\epsilon_i > 0$ and $\upsilon_i > 0$.*

Proof. In the first case, we take $\bar{h}_i(x_i, y) = h_i(x_i, y)$, and in the second case, we take $\bar{h}_i(x_i, y) = \upsilon_i \|x_i - y_i\|^{1+\epsilon_i}$. ∎

Let us discuss the parameters γ_i^k and η_i^k in (4). In Section 4, we provide their explicit formulas in some specific examples of TITAN which correspond to specific choices of u_i and $\mathcal{G}_i^k$. In the following, we characterize general choices of u_i and $\mathcal{G}_i^k$ such that the condition (4) is satisfied. For $\mathcal{G}_i^k$, we will use the following assumption.

ASSUMPTION 2.4. *There exist constants A_i^k such that the extrapolation operator $\mathcal{G}_i^k$ satisfies $\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \leq A_i^k \|x_i^k - x_i^{k-1}\|$ for $i \in [m]$ and $k \geq 0$.*

For u_i, we will use one of the two following assumptions.

ASSUMPTION 2.5. *Given $y \in \mathcal{X}$, there exists a positive constant $\rho_i^{(y)}$ such that the block i approximation error satisfies the inequality*

$$h_i(x_i, y) \geq \tfrac{\rho_i^{(y)}}{2} \|x_i - y_i\|^2 \quad \text{for all} \quad x_i \in \mathcal{X}_i.$$

ASSUMPTION 2.6. *Given $y \in \mathcal{X}$, the function $x_i \mapsto u_i(x_i, y) + g_i(x_i)$ is $\rho_i^{(y)}$-strongly convex.*

Assumption 2.4 is satisfied taking $A_i^k = \frac{\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\|}{\|x_i^k - x_i^{k-1}\|}$. Assumption 2.5 is always satisfied for the regularized block i surrogate function $u_i(x_i, y) + \frac{\rho_i^{(y)}}{2}\|x_i - y_i\|^2$, where $u_i(x_i, y)$ is any block i surrogate function of f. The following theorem is a cornerstone to characterize general choices of u_i and $\mathcal{G}_i^k$ that satisfy the NSDP condition (4). The two important parameters in Theorem 2.7 to compute γ_i^k and η_i^k of Condition (4) are $\rho_i^{(y)}$ of Assumption 2.5 (or $\rho_i^{(y)}$ of Assumption 2.6) and A_i^k of Assumption 2.4.

THEOREM 2.7. *Suppose $\mathcal{G}_i^k$ satisfies Assumption 2.4, and u_i satisfies Assumption 2.5 or Assumption 2.6 for $y = x^{k,i-1}$. Then the Condition (4) holds with*

$$\gamma_i^k = \frac{(A_i^k)^2}{2\nu\rho_i^{(x^{k,i-1})}}, \qquad \eta_i^k = \frac{(1-\nu)\rho_i^{(x^{k,i-1})}}{2},$$

where $0 < \nu < 1$ is a constant. For notation succinctness, we denote $\rho_i^k = \rho_i^{(x^{k,i-1})}$.

Proof. In this proof, we denote $y = x^{k,i-1}$. Let us consider the first case: Assumptions 2.4 and Assumptions 2.5 hold. We have

$$u_i(x_i^{k+1}, y) = f(x_i^{k+1}, y_{\neq i}) + h_i(x_i^{k+1}, y) \geq f(x_i^{k+1}, y_{\neq i}) + \tfrac{\rho_i^k}{2}\|x_i^{k+1} - x_i^k\|^2. \tag{6}$$

On the other hand, it follows from (3) that, for all $x_i \in \mathcal{X}_i$, we have

$$u_i(x_i^{k+1}, y) + g_i(x_i^{k+1}) \leq u_i(x_i, y) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i - x_i^{k+1}\rangle + g_i(x_i). \tag{7}$$

Choosing $x_i = x_i^k$ in (7), we get the following inequality from (7) and (6):

$$\begin{aligned} u_i(x_i^k, y) + g_i(x_i^k) - \langle \mathcal{G}_i^k(x_i^k, y), x_i^k - x_i^{k+1}\rangle \\ \geq f(x_i^{k+1}, y_{\neq i}) + g_i(x_i^{k+1}) + \tfrac{\rho_i^k}{2}\|x_i^k - x_i^{k+1}\|^2. \end{aligned} \tag{8}$$

Since $u_i(x_i^k, y) = f(y)$, we derive from (8) that

$$F(x^{k,i-1}) - \left\langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i^k - x_i^{k+1}\right\rangle \geq F(x^{k,i}) + \tfrac{\rho_i^k}{2}\|x_i^{k+1} - x_i^k\|^2. \tag{9}$$

From Young's inequality, we have

$$A_i^k\|x_i^k - x_i^{k-1}\|\|x_i^{k+1} - x_i^k\| \leq \tfrac{\nu\rho_i^k}{2}\|x_i^{k+1} - x_i^k\|^2 + \tfrac{(A_i^k)^2}{2\nu\rho_i^k}\|x_i^k - x_i^{k-1}\|^2.$$

Hence, from (9) and Assumption 2.4, we obtain

$$F(x^{k,i}) + \tfrac{(1-\nu)\rho_i^k}{2}\|x_i^{k+1} - x_i^k\|^2 \leq F(x^{k,i-1}) + \tfrac{(A_i^k)^2}{2\nu\rho_i^k}\|x_i^k - x_i^{k-1}\|,$$

which gives the result.

Consider the second case when Assumptions 2.4 and 2.6 hold. Let $\tilde{u}_i(x_i, y) = u_i(x_i, y) + g_i(x_i)$. It follows from the optimality conditions of (3) that

$$\left\langle \mathbf{s}_i(x_i^{k+1}) - \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i^k - x_i^{k+1}\right\rangle \geq 0, \tag{10}$$

where $\mathbf{s}_i(x_i^{k+1})$ is a subgradient of $\tilde{u}_i(\cdot, y)$ at x_i^{k+1}. Since $\tilde{u}_i(\cdot, y)$ is strongly convex, we have $\tilde{u}_i(x_i^k, y) \geq \tilde{u}_i(x_i^{k+1}, y) + \left\langle \mathbf{s}_i(x_i^{k+1}), x_i^k - x_i^{k+1}\right\rangle + \frac{\rho_i^k}{2}\|x_i^k - x_i^{k+1}\|^2$. Together with (10) and noting that $u_i(x_i^{k+1}, y) \geq f(x_i^{k+1}, y_{\neq i})$, we get (8). The result follows using the same proof as in the first case. ∎

Let us provide a sufficient condition for Assumption 2.5.

LEMMA 2.8. *If $h_i(\cdot, y)$ is $\rho_i^{(y)}$-strongly convex and is differentiable at y_i, and $\nabla_i h_i(y_i, y) = 0$, then we have $h_i(x_i, y) \geq \frac{\rho_i^{(y)}}{2}\|x_i - y_i\|^2$.*

Proof. The result follows from the definition of $\rho_i^{(y)}$-strong convexity, that is,

$$h_i(x_i, y) \geq h_i(y_i, y) + \langle \nabla_i h_i(y_i, y), x_i - y_i\rangle + \frac{\rho_i^{(y)}}{2}\|x_i - y_i\|^2,$$

the assumption $\nabla_i h_i(y_i, y) = 0$, and the property $h_i(y_i, y) = 0$ from Definition 2.1. ∎

3. Convergence analysis. In this section, we study sub-sequential convergence as well as global convergence of TITAN. For our upcoming analysis, we need the following first-order optimality condition of (1):

$$\langle p(x^*), x - x^*\rangle \geq 0 \text{ for all } x \in \mathcal{X}, \text{ for some } p(x^*) \in \partial F(x^*). \tag{11}$$

As we assume $\partial F(x^*) = \{\partial_{x_1} F(x^*)\} \times \ldots \times \{\partial_{x_m} F(x^*)\}$, (11) is equivalent to

$$\langle p_i(x^*), x_i - x_i^*\rangle \geq 0 \text{ for all } x_i \in \mathcal{X}_i, \text{for some } p_i(x^*) \in \partial_{x_i} F(x^*) \text{ for } i \in [m]. \tag{12}$$

If x^* is in the interior of $\mathcal{X}$ or $\mathcal{X}_i = \mathbb{E}_i$ then (11) reduces to the condition $0 \in \partial F(x^*)$, that is, x^* is a critical point of F.

Throughout this section, we assume that the parameters are chosen such that, for $i = 1, \ldots, m$ and $k \geq 0$, the condition in (4) and Assumption 2.2 are satisfied. The following proposition provides a requirement for γ_i^k and η_i^k in the NSDP condition (4) such that a sub-sequential convergence can be achieved.

PROPOSITION 3.1. *Let $\{x^k\}$ be the sequence generated by Algorithm 1. Suppose that, for $k = 0, 1, \ldots$,*

$$\gamma_i^{k+1} \leq C\eta_i^k \tag{13}$$

for some constant $0 < C < 1$. Let $\eta_i^{-1} = \gamma_i^0/C$. The following statements hold.

(A) For any $K > 1$, we have

$$F(x^K) + (1-C)\sum_{k=0}^{K-1}\sum_{i=1}^{m}\tfrac{\eta_i^k}{2}\|x_i^{k+1} - x_i^k\|^2 \leq F(x^0) + C\sum_{i=1}^{m}\tfrac{\eta_i^{-1}}{2}\|x_i^0 - x_i^{-1}\|^2. \tag{14}$$

(B) If there exists a positive number $\underline{l}$ such that $\min_{i,k}\left\{\frac{\eta_i^k}{2}\right\} \geq \underline{l}$, then we have $\sum_{k=0}^{+\infty}\sum_{i=1}^{m}\|x_i^{k+1} - x_i^k\|^2 < +\infty$.

Proof. (A) It follows from (4) and (13) that, for $k = 0, 1, \ldots$, we have

$$F(x^{k,i}) + \tfrac{\eta_i^k}{2}\|x_i^{k+1} - x_i^k\|^2 \leq F(x^{k,i-1}) + C\tfrac{\eta_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|^2.$$

Summing this inequality over $i = 1, ..., m$ gives

$$F(x^{k+1}) + \sum_{i=1}^{m}\tfrac{\eta_i^k}{2}\|x_i^{k+1} - x_i^k\|^2 \leq F(x^k) + C\sum_{i=1}^{m}\tfrac{\eta_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|^2. \tag{15}$$

Summing up Inequality (15) from $k = 0$ to $K - 1$, we obtain

$$\begin{aligned} &F(x^0) + \sum_{i=1}^{m} C\tfrac{\eta_i^{-1}}{2}\|x_i^0 - x_i^{-1}\|^2 \\ &\quad\geq F(x^K) + C\sum_{i=1}^{m}\tfrac{\eta_i^{K-1}}{2}\|x_i^K - x_i^{K-1}\|^2 + (1-C)\sum_{k=0}^{K-1}\sum_{i=1}^{m}\tfrac{\eta_i^k}{2}\|x_i^{k+1} - x_i^k\|^2, \end{aligned}$$

which gives the result.

(B) The result is a direct consequence of the inequality (14). ∎

3.1. Sub-sequential Convergence. Let us now prove sub-sequential convergence of TITAN.

THEOREM 3.2 (Sub-sequential convergence). *Suppose the conditions in Proposition 3.1 are satisfied. We further assume that the generated sequence $\{x^k\}$ by Algorithm 1 is bounded[3] and $\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\|$ goes to[4] 0 when k goes to ∞. Then every limit point x^* of $\{x^k\}$ satisfies the optimality condition in (11).*

Proof. Suppose a subsequence $\{x^{k_n}\}$ of $\{x^k\}$ converges to $x^* \in \mathcal{X}$. Proposition 3.1(B) implies that $x^{k_n-1} \to x^*$ and $x^{k_n+1} \to x^*$. Choosing $x_i = x_i^*$ and $k = k_n$ in (7), we obtain

$$\begin{aligned} &u_i(x_i^{k_n+1}, x^{k_n,i-1}) + g_i(x_i^{k_n+1}) \\ &\quad\leq u_i(x_i^*, x^{k_n,i-1}) - \langle \mathcal{G}_i^{k_n}(x_i^{k_n}, x_i^{k_n-1}), x_i^* - x_i^{k_n+1}\rangle + g_i(x_i^*). \end{aligned} \tag{16}$$

[3]Proposition 3.1 proves that $F(x^k) \leq F(x^0) + C\sum_{i=1}^{m}\frac{\eta_i^{-1}}{2}\|x_i^0 - x_i^{-1}\|^2$. Therefore, the boundedness of $\{x^k\}$ is satisfied for the class of bounded-level set functions F.

[4]From Proposition 3.1(B), we have $\|x_i^k - x_i^{k-1}\|$ converges to 0 when k goes to ∞. On the other hand, we assume $\{x^k\}$ is bounded. Hence, the condition that $\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\|$ goes to 0 is satisfied by $\mathcal{G}_i^k$ satisfying Assumption 2.4, in which A_i^k is bounded for the bounded sequence $\{x^k\}$.

Note that $x^{k_n,i-1} \to x^*$ and $u_i(x_i, y)$ is continuous in y by Assumption 2.2(A). Hence, we derive from (16) that

$$\limsup_{n\to\infty} u_i(x_i^{k_n+1}, x^{k_n,i-1}) + g_i(x_i^{k_n+1}) \leq u_i(x_i^*, x^*) + g_i(x^*).$$

Furthermore, $u_i(x_i, y) + g_i(x_i)$ is lower semi-continuous. Hence, $u_i(x_i^{k_n+1}, x^{k_n,i-1}) + g_i(x_i^{k_n+1}) \to u_i(x_i^*, x^*) + g_i(x_i^*)$. We choose $k = k_n$ in (7) and let $n \to \infty$ to obtain

$$u_i(x_i^*, x^*) + g_i(x_i^*) \leq u_i(x_i, x^*) + g_i(x_i) \quad \text{for all} \quad x_i \in \mathcal{X}_i.$$

Note that $u_i(x_i^*, x^*) + g_i(x_i^*) = F(x^*)$ and $u_i(x_i, x^*) = f(x_i, x^*_{\neq i}) + h_i(x_i, x^*)$. Therefore, for all $x_i \in \mathcal{X}_i$,

$$\begin{aligned} F(x^*) &\leq F(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_m^*) + h_i(x_i, x^*) \\ &\leq F(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_m^*) + \bar{h}_i(x_i, x^*), \end{aligned} \tag{17}$$

where we have used Assumption 2.2(B). Inequality (17) shows that, for $i = 1, \ldots, m$, x_i^* is a minimizer of the problem

$$\min_{x_i \in \mathcal{X}_i} F(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_m^*) + \bar{h}_i(x_i, x^*). \tag{18}$$

The result follows from the optimality condition of (18) and $\nabla_i \bar{h}_i(x_i^*, x^*) = 0$. ∎

Remark 3.3. Consider the case $\mathcal{X} = \mathbb{E} := \mathbb{E}_1 \times \ldots \times \mathbb{E}_m$. For this case, Inequality (5) for $x_i \in \mathbb{E}_i$ can be relaxed to x_i on bounded subsets of $\mathbb{E}_i$. By noting that if x_i^* is a local minimizer of Problem (18) then $0 \in \partial_{x_i} F(x^*)$, we can repeat the analysis to prove the subsequential convergence to a critical point of F for the relaxed condition.

3.2. Global Convergence. A global convergence recipe was proposed in [5, 6, 12] to prove the global convergence of alternating proximal (that is, when the proximal surrogate function in Section 4.1 is used) and alternating proximal gradient methods (that is, when the Lipschitz gradient surrogate function in Section 4.2 is used) for solving non-smooth non-convex problems. The recipe was extended in [33] and [19, Theorem 2] to deal with the accelerated algorithms, which may produce non-monotone sequences of objective function values. For completeness, we provide [19, Theorem 2], which will be used to prove the global convergence of TITAN, in Appendix B. One typical assumption to prove the global convergence is that the gradient of f in Problem (1) is Lipschitz continuous on bounded subsets of $\mathbb{E}$. This assumption is naturally satisfied when f is continuously differentiable over $\mathbb{E}$. It is important to note that the Lipschitz constant of this assumption does not influence how to choose parameters for the algorithms, its existence is just for the purpose of proving the global convergence of the generated sequence, which is usually assumed to be bounded.

Problem (1) is equivalent to the following unconstrained optimization problem

$$\min_{x \in \mathbb{E}} \Phi(x) := F(x) + \sum_{i=1}^{m} \mathcal{I}_{\mathcal{X}_i}(x_i), \tag{19}$$

where $\mathcal{I}_{\mathcal{X}_i}(\cdot)$, for $i \in [m]$, is the indicator function of $\mathcal{X}_i$. Hence, it makes sense to consider the optimality condition $0 \in \partial\Phi(x^*)$ for Problem (1), that is x^* is a critical point of Φ. In the upcoming analysis, we will prove the global convergence of TITAN to a critical point of Φ. Note that $\Phi(x) = F(x)$ when $\mathcal{X}_i = \mathbb{E}_i$. We make the following additional assumption.

ASSUMPTION 3.4. *For $i = 1, \ldots, m$, the block surrogate functions $u_i(x_i, y)$ is continuous on the joint variable (x_i, y). Moreover, the limiting subgradient $\partial_{x_i} u_i(\cdot, \cdot)$ is Lipschitz continuous on bounded subsets of $\mathcal{X}_i \times \mathcal{X}$ in the sense that, for any (x_i, w) and (y_i, v) in a bounded subset of $\mathcal{X}_i \times \mathcal{X}$, if $\mathbf{s}_i \in \partial_{x_i}\big(u_i(x_i, w) - u_i(y_i, v)\big)$ then $\|\mathbf{s}_i\| \leq B_i \|(x_i, w) - (y_i, v)\|$ for some constant B_i.*

Assumption 3.4 is naturally satisfied when the surrogate functions $u_i(\cdot, \cdot)$ is continuously differentiable. We will see that all the surrogate functions given in Sections 4.1–4.4 satisfy Assumption 3.4 when f is continuously differentiable.

THEOREM 3.5 (Global convergence). *Suppose the conditions in Proposition 3.1 is satisfied. We further assume that f is continuously differentiable, Assumption 2.2(B) is satisfied[5] with $\bar{h}(x_i, y) = h(x_i, y)$, Assumption 3.4 holds, Assumption 2.4 holds with bounded A_i^k, F is a KL function (see Appendix A), and together with the existence of $\underline{l}$ in Proposition 3.1, we also assume there exists $\bar{l} > 0$ such that $\max_{i,k} \left\{\frac{\eta_i^k}{2}\right\} \leq \bar{l}$. Suppose one of the following two conditions hold.*

1. *Condition (13) is satisfied with some C satisfying $C < \underline{l}/\bar{l}$.*
2. *We use a restarting regime for TITAN, that is, if $F(x^{k+1}) \geq F(x^k)$ then we re-do the k-iteration with $\mathcal{G}_i^k = 0$ (that is, no extrapolation is used). In this situation, we assume that Condition (4) is satisfied with[6] $\gamma_i^k = 0$, for $i \in [m]$.*

Then the whole generated sequence $\{x^k\}$ of Algorithm 1, which is assumed to be bounded, converges to a critical point of Φ.

Proof. See Appendix C.1. ∎

We make some remarks to end this section.

Remark 3.6. As long as a global convergence is guaranteed, we can derive a convergence rate for the generated sequence by using the same technique as in the proof of [4, Theorem 2]. We refer the reader to [19, Theorem 3] and [44, Theorem 2.9] for some examples of using the technique of [4, Theorem 2] to derive the convergence rate and omit the details of the convergence rate for TITAN.

Remark 3.7. In general, it is not easy to estimate the bounds $\underline{l}$ and $\bar{l}$ for the situations when η_i^k vary along with the block updates. Hence, if we target a global convergence guarantee, TITAN without restarting step is recommended when these bounds are easy to estimate and the objective function is expensive to evaluated. For surrogate functions with varying η_i^k, TITAN with a restarting regime is recommended to guarantee a global convergence. It is important noting that TITAN always guarantees a sub-sequential convergence with or without restarting steps.

4. Some TITAN Accelerated Block Coordinate Methods. In order to guarantee some convergence, TITAN must choose the parameters that satisfy the conditions in Theorem 3.2, which include Assumption 2.2, the NSDP condition (4), the condition (13) and the condition $\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \to 0$. As noted in the footnote 4, the condition $\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \to 0$ is satisfied by the extrapolation satisfying Assumption 2.4 with bounded A_i^k. Theorem 2.7 characterizes some general properties of u_i and $\mathcal{G}_i^k$ that make the NSDP condition (4) hold, and it determines the corresponding values of η_i^k and γ_i^k when Assumption 2.4 and Assumption 2.5 (or Assumption 2.6) is satisfied. In the following, we consider some important block surrogate functions from the literature (more examples can be found in [28]), and derive several specific

[5] The block error function of the surrogate functions in Sections 4.1 – 4.4 satisfies this condition.

[6] If u_i satisfies Assumption 2.5 or Assumption 2.6 then we repeat the proof of Theorem 2.7 to derive Inequality (9) which leads to Condition (4) being satisfied with $\gamma_i^k = 0$ and $\eta_i^k = \rho_i^k/2$.

instances of TITAN. We verify Assumption 2.2(B) (by using Lemma 2.3) and provide the formulas of η_i^k and γ_i^k (by using Theorem 2.7).

4.1. TITAN with proximal surrogate function. The proximal surrogate function, which has been used for example in [4, 6, 19], has the following form

$$u_i(x_i, y) = f(x_i, y_{\neq i}) + \tfrac{\rho_i^{(y)}}{2}\|x_i - y_i\|^2,$$

where f is a lower semi-continuous function and $\rho_i^{(y)} > 0$ is a scalar.

We have $h_i(x_i, y) = \frac{\rho_i^{(y)}}{2}\|x_i - y_i\|^2$. Hence, Assumption 2.2(B) and Assumption 2.5 are satisfied. Let us choose $\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \rho_i^k \beta_i^k (x_i^k - x_i^{k-1})$, where β_i^k are some extrapolation parameters and $\rho_i^k = \rho_i^{(x^{k,i-1})}$. In this case, we have $A_i^k = \rho_i^k \beta_i^k$. The minimization problem in the update (3) becomes

$$\min_{x_i \in \mathcal{X}_i} f(x_i, x_{\neq i}^{k,i-1}) + \tfrac{\rho_i^k}{2}\|x_i - (x_i^k + \beta_i^k (x_i^k - x_i^{k-1}))\|^2 + g_i(x_i).$$

The formulas of η_i^k and γ_i^k are determined as in Theorem 2.7. This TITAN scheme recovers the inertial block proximal algorithm in [19].

4.2. TITAN with Lipschitz gradient surrogates. The Lipschitz gradient surrogate function, which has been used for example in [44, 45, 19], has the form

$$u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \tfrac{\kappa_i L_i^{(y)}}{2}\|x_i - y_i\|^2,$$

where $\kappa_i \geq 1$, the block function $x_i \mapsto f(x_i, y_{\neq i})$ is differentiable and $\nabla_i f(x_i, y_{\neq i})$ is $L_i^{(y)}$-Lipschitz continuous. Note that $L_i^{(y)}$ may depend on y. We have

$$\nabla_{x_i} h_i(x_i, y) = \kappa_i L_i^{(y)}(x_i - y_i) + \nabla_i f(y) - \nabla_i f(x_i, y_{\neq i}).$$

So, $\nabla_{x_i} h_i(y_i, y) = 0$. Hence, Assumption 2.2(B) is satisfied with $\bar{h}_i(x_i, y) = h_i(x_i, y)$.

On the other hand, in general when $g_i(x_i)$ is a non-convex function, we always have $x_i \mapsto h_i(x_i, y)$ is a $(\kappa_i - 1)L_i^{(y)}$-strongly convex function. In this case we need to choose $\kappa_i > 1$, and then Assumption 2.5 is satisfied with $\rho_i^{(y)} = (\kappa_i - 1)L_i^{(y)}$. If $g_i(x_i)$ is convex then we have $x_i \mapsto u_i(x_i, y) + g_i(x_i)$ is a $\kappa_i L_i^{(y)}$-strongly convex function; as such, in this case we can choose $\kappa_i = 1$ and Assumption 2.6 is satisfied with $\rho_i^{(y)} = L_i^{(y)}$. Taking $y = x^{k,i-1}$, the formulas of η_i^k and γ_i^k are determined as in Theorem 2.7. In the following, we consider specific choices for $\mathcal{G}_i^k$ and determine the corresponding values for A_i^k.

4.2.1. Deriving inertial block proximal gradient methods. Let us consider the case $\mathcal{X}_i = \mathbb{E}_i$ and choose

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \nabla_i f(x^{k,i-1}) - \nabla_i f(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1}), \tag{20}$$

where $\bar{x}_i^k = x_i^k + \tau_i^k (x_i^k - x_i^{k-1})$, τ_i^k, β_i^k are some extrapolation parameters and $L_i^k = L_i^{(x^{k,i-1})}$. The update in (3) becomes

$$\begin{aligned}
&\operatorname*{argmin}_{x_i} f(x^{k,i-1}) + \left\langle \nabla_i f(x^{k,i-1}), x_i - x_i^k \right\rangle + \tfrac{\kappa_i L_i^k}{2}\|x_i - x_i^k\|^2 \\
&\quad - \left\langle \nabla_i f(x^{k,i-1}) - \nabla_i f(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1}), x_i \right\rangle + g_i(x_i) \\
&= \operatorname*{argmin}_{x_i} \left\langle \nabla_i f(\bar{x}_i^k, x_{\neq i}^{k,i-1}), x_i \right\rangle + g_i(x_i) + \tfrac{\kappa_i L_i^k}{2}\left\|x_i - (x_i^k + \beta_i^k (x_i^k - x_i^{k-1}))\right\|^2.
\end{aligned}$$

Let us now determine the values of A_i^k in Assumption 2.4. In general, we have

$$\|\mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \leq L_i^k(\tau_i^k + \kappa_i\beta_i^k)\|x_i^k - x_i^{k-1}\|.$$

Hence, in a general case, we can take $A_i^k = L_i^k(\tau_i^k + \kappa_i\beta_i^k)$.

If we additionally assume that the block function $f(\cdot, x_{\neq i}^{k,i-1})$ is convex then we can get a tighter bound for A_i^k. Specifically, if we choose $\beta_i^k \geq \tau_i^k$, then the function $x_i \mapsto \xi(x_i) = \frac{1}{2}\kappa_i L_i^k \frac{\beta_i^k}{\tau_i^k}(x_i)^2 - f(x_i, x_{\neq i}^{k,i-1})$ is convex and it has $\big(\kappa_i L_i^k \frac{\beta_i^k}{\tau_i^k}\big)$-Lipschitz gradient. Therefore, we get

$$\|\nabla\xi(\bar{x}_i^k) - \nabla\xi(x_i^k)\| \leq \kappa_i L_i^k \tfrac{\beta_i^k}{\tau_i^k}\|\bar{x}_i^k - x_i^k\| = \kappa_i L_i^k \beta_i^k \|x_i^k - x_i^{k-1}\|.$$

On the other hand, we see that

$$\begin{aligned}\nabla\xi(\bar{x}_i^k) - \nabla\xi(x_i^k) &= \kappa_i L_i^k \tfrac{\beta_i^k}{\tau_i^k}\bar{x}_i^k - \nabla_i f(\bar{x}_i^k, x_{\neq i}^{k,i-1}) - \kappa_i L_i^k \tfrac{\beta_i^k}{\tau_i^k} x_i^k + \nabla_i f(x^{k,i-1}) \\ &= \mathcal{G}_i^k(x_i^k, x_i^{k-1}).\end{aligned}$$

Hence, in this case we can take $A_i^k = \kappa_i L_i^k \beta_i^k$.

We can recover the accelerated methods in the literature as follows.

- If we use $\mathcal{G}_i^k$ in (20) and choose $\beta_i^k = \tau_i^k$ then we recover the Nesterov type acceleration as in [44, 45].
- If we use $\mathcal{G}_i^k$ in (20) and let $\beta_i^k \neq \tau_i^k$ and $\beta_i^k \geq \tau_i^k$ then the update in (3) uses two different extrapolation points as in [19].

It is important noting that we can also recover the heavy-ball type acceleration by choosing $\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1})$, and, for this case, we can take $\mathcal{X}_i$ to be any closed convex subset of $\mathbb{E}_i$.

4.2.2. Inertial block proximal gradient algorithm with Hessian damping. Let us choose

$$\mathcal{G}_i^k = \alpha_i^k\big(\nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1}) - \nabla_i f(x^{k,i-1})\big) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1}), \tag{21}$$

where α_i^k and β_i^k are some extrapolation parameters. The problem in (3) becomes

$$\begin{aligned}&\operatorname*{argmin}_{x_i} f(x^{k,i-1}) + \big\langle \nabla_i f(x^{k,i-1}), x_i - x_i^k\big\rangle + \tfrac{\kappa_i L_i^k}{2}\|x_i - x_i^k\|^2 \\ &\quad - \Big\langle \alpha_i^k\big(\nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1}) - \nabla_i f(x^{k,i-1})\big) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1}), x_i\Big\rangle + g_i(x_i) \\ &= \operatorname*{argmin}_{x_i} \Big\langle \nabla_i f(x^{k,i-1}) + \alpha_i^k\big(\nabla_i f(x^{k,i-1}) - \nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1})\big), x_i\Big\rangle + g_i(x_i) \\ &\qquad + \tfrac{\kappa_i L_i^k}{2}\big\|x_i - (x_i^k + \beta_i^k(x_i^k - x_i^{k-1}))\big\|^2.\end{aligned}$$

We have derived an inertial block proximal gradient algorithm with the corrective term $\nabla_i f(x^{k,i-1}) - \nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1})$ (this term is related to the discretization of the Hessian-driven damping term, see [1]). When $g_i(x_i) = 0$ the update in (3) becomes

$$x_i^{k+1} = x_i^k + \beta_i^k(x_i^k - x_i^{k-1}) - \frac{1}{\kappa_i L_i^k}\Big(\nabla_i f(x^{k,i-1}) + \alpha_i^k\big(\nabla_i f(x^{k,i-1}) - \nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1})\big)\Big),$$

which has the form of the inertial gradient algorithm with Hessian damping as in [1].

Let us now determine the values of A_i^k. We can take $A_i^k = L_i^k(\alpha_i^k + \kappa_i\beta_i^k)$ in a general case. If we additionally assume that the block function $f(\cdot, x_{\neq i}^{k,i-1})$ is convex, we then choose $\alpha_i^k \leq \kappa_i\beta_i^k$ to guarantee the convexity of the function $x_i \mapsto \xi(x_i) = \frac{1}{2}\kappa_i L_i^k \beta_i^k (x_i)^2 - \alpha_i^k f(x_i, x_{\neq i}^{k,i-1})$. Note that $\xi(x_i)$ has $\kappa_i L_i^k \beta_i^k$-Lipschitz gradient. Hence, similarly to Section 4.2.1, we can take $A_i^k = \kappa_i L_i^k \beta_i^k$ for this case.

Remark 4.1. We have derived the values of η_i^k and γ_i^k by using Theorem 2.7, and specific values of A_i^k and ρ_i^k of Theorem 2.7 were given. The cases (i) f and g_i are non-convex, (ii) the block functions of f are convex but g_i are not, and (iii) f is non-convex but g_i are convex, were analysed in Section 4.2. When F possesses the strong property that the block functions of f, $i \in [m]$, are convex and g_i are convex, we can obtain a better bound for γ_i^k and η_i^k that allow larger extrapolation parameters. Let us choose $\mathcal{G}_i^k$ as in (20). It was established in the proof within [19, Remark 3] that

(22)
$$F(x^{k,i-1}) + \frac{L_i^k}{2}\big((\tau_i^k)^2 + \frac{(\beta_i^k - \tau_i^k)^2}{\nu}\big)\|x_i^k - x_i^{k-1}\|^2 \geq F(x^{k,i}) + \frac{(1-\nu)L_i^k}{2}\|x_i^{k+1} - x_i^k\|^2,$$

where $0 < \nu < 1$ is a constant. Hence, in this case

$$\gamma_i^k = \frac{L_i^k}{2}\Big((\tau_i^k)^2 + \frac{(\beta_i^k - \tau_i^k)^2}{\nu}\Big), \quad \eta_i^k = \frac{(1-\nu)L_i^k}{2}. \tag{23}$$

4.3. TITAN with Bregman surrogates. The Bregman surrogate for relative smooth functions, which has been used for example in [2, 18, 41], has the form

$$u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i\rangle + \kappa_i L_i^{(y)} D_{\varphi_i^{(y)}}(x_i, y_i),$$

where $\kappa_i \geq 1$, the block function $x_i \mapsto f(x_i, y_{\neq i})$ is differentiable, $\varphi_i^{(y)}$ is a differentiable convex function such that the function $x_i \mapsto L_i^{(y)}\varphi_i^{(y)}(x_i) - f(x_i, y_{\neq i})$ is convex, and $D_{\varphi_i^{(y)}}$ is the block Bregman divergence associated with $\varphi_i^{(y)}$ defined by

$$D_{\varphi_i^{(y)}}(x_i, v_i) = \varphi_i^{(y)}(x_i) - [\varphi_i^{(y)}(v_i) + \langle \nabla\varphi_i^{(y)}(v_i), x_i - v_i\rangle]. \tag{24}$$

It is assumed that $\varphi_i^{(y)}$ is a $\rho_{\varphi_i^{(y)}}$-strongly convex function on $\mathbb{E}_i$ and its gradient is Lipschitz continuous on bounded subsets of $\mathbb{E}_i$. We have

$$\nabla_{x_i} h_i(x_i, y) = \kappa_i L_i^{(y)}(\nabla\varphi_i^{(y)}(x_i) - \nabla\varphi_i^{(y)}(y_i)) + \nabla_i f(y) - \nabla_i f(x_i, y_{\neq i}).$$

Hence, Assumption 2.2(B) is satisfied with $\bar{h}_i(x_i, y) = h_i(x_i, y)$.

Similarly to Section 4.2, if $g_i(x_i)$ is convex then $x_i \mapsto u_i(x_i, y) + g_i(x_i)$ is a $\kappa_i L_i^{(y)}\rho_{\varphi_i}$-strongly convex function. In this case we can choose $\kappa_i = 1$ and Assumption 2.6 is satisfied with $\rho_i^{(y)} = L_i^{(y)}\rho_{\varphi_i^{(y)}}$. Considering the case g_i is not convex, as we have $h_i(\cdot, y)$ is a $(\kappa_i - 1)L_i^{(y)}\rho_{\varphi_i^{(y)}}$-strongly convex function, we need to choose $\kappa_i > 1$ and then Assumption 2.5 is satisfied with $\rho_i^{(y)} = (\kappa_i - 1)L_i^{(y)}\rho_{\varphi_i^{(y)}}$. Taking $y = x^{k,i-1}$, the formulas of η_i^k and γ_i^k are determined as in Theorem 2.7. In the following, we consider two specific choices for $\mathcal{G}_i^k$ and determine the corresponding values for A_i^k.

Weak inertial force. Let us choose $\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \beta_i^k(x_i^{k-1} - x_i^k)$, where β_i^k are some extrapolation parameters, then we recover the block inertial Bregman proximal algorithm in [3]. In this case, we have $A_i^k = \beta_i^k$.

Heavy ball type acceleration with back-tracking. Now let us choose

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \kappa_i L_i^k(\nabla\varphi_i^k(\bar{x}_i^k) - \nabla\varphi_i^k(x_i^k)),$$

where $\varphi_i^k = \varphi_i^{(x^{k,i-1})}$, $\bar{x}_i^k = x_i^k + \tau_i^k(x_i^k - x_i^{k-1})$ with τ_i^k being extrapolation parameters. We recall that it is assumed that $\varphi_i^k(\cdot)$ is strongly convex and differentiable on $\mathbb{E}_i$, hence $\nabla\varphi_i^k(\bar{x}_i^k)$ is well-defined. The update (3) becomes

$$\begin{aligned}&\operatorname*{argmin}_{x_i} \left\langle \nabla_i f(x^{k,i-1}), x_i - x_i^k \right\rangle + g_i(x_i) + \kappa_i L_i^k \big(\varphi_i^k(x_i) - \langle \nabla\varphi_i^k(\bar{x}_i^k), x_i - \bar{x}_i^k\rangle - \varphi_i^k(\bar{x}_i^k)\big)\\ &= \operatorname*{argmin}_{x_i} \left\langle \nabla_i f(x^{k,i-1}), x_i - x_i^k \right\rangle + g_i(x_i) + D_{\varphi_i^k}(x_i, \bar{x}_i^k),\end{aligned}$$

which has the form of a heavy ball acceleration. Note that we do not assume $\nabla\varphi_i^k$ is globally Lipschitz continuous. Therefore, we propose to apply line-search to determine the extrapolation parameter τ_i^k as follows. Starting with $\tau_i^k = 1$, we decrease τ_i^k by multiplying it with a constant $\bar{\tau} < 1$ until the following condition is satisfied

$$\kappa_i L_i^k \|\nabla\varphi_i^k(\bar{x}_i^k) - \nabla\varphi_i^k(x_i^k)\|^2 \le C\|x_i^k - x_i^{k-1}\|^2 \rho_i^k \rho_i^{k+1}.$$

This process terminates after finite steps as we assume $\nabla\varphi_i^k(x_i)$ is Lipschitz continuous on bounded sets. Then the condition in (13) is satisfied with $A_i^k = \frac{\|\nabla\varphi_i^k(\bar{x}_i^k) - \nabla\varphi_i^k(x_i^k)\|}{\|x_i^k - x_i^{k-1}\|}$.

4.4. TITAN with quadratic surrogates. The quadratic surrogate, which has been used for example in [14, 33], has the following form

$$u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i\rangle + \frac{\kappa_i}{2}(x_i - y_i)^T H_i^{(y)}(x_i - y_i), \tag{25}$$

where $\kappa_i \ge 1$, f is twice differentiable and $H_i^{(y)}$ is a positive definite matrix such that $(H_i^{(y)} - \nabla_i^2 f(x_i, y_{\neq i}))$ is positive definite ($H_i^{(y)}$ may depend on y). Taking $y = x^{k,i-1}$, we note that although the quadratic surrogate is a special case of the Bregman surrogate (Section 4.3) with $\varphi_i^k(x) = x_i^T H_i^k x_i$, $L_i^k = 1$ and $\rho_{\varphi_i^k}$ being the smallest eigenvalue of H_i^k, the kernel function $\varphi_i^k(x_i) = \left\langle x_i, H_i^k x_i \right\rangle$ is globally $\|H_i^k\|$-Lipschitz smooth. We choose G_i^k as follows

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \kappa_i(H_i^k(\bar{x}_i^k) - H_i^k(x_i^k)) = \kappa_i\tau_i^k H_i^k(x_i^k - x_i^{k-1}),$$

where $\bar{x}_i^k = x_i^k + \tau_i^k(x_i^k - x_i^{k-1})$. In this case, $A_i^k = \kappa_i\tau_i^k\|H_i^k\|$. The update in (3) has the form of a heavy ball acceleration

$$\operatorname*{argmin}_{x_i} \left\langle \nabla_i f_i(x_i^k), x_i - x_i^k \right\rangle + g_i(x_i) + \frac{\kappa_i}{2}(x_i - \bar{x}_i^k)^T H_i^k(x_i - \bar{x}_i^k).$$

4.5. TITAN with composite surrogates. Suppose f has the form

$$f(x) = \psi(x) + \phi(r(x)), \tag{26}$$

where

- $\psi : \mathcal{X} \to \mathbb{R}$ is a non-smooth non-convex function that has block surrogate functions $u_i^\psi(x_i, y)$, $i = 1, \ldots, m$,
- $r = (r_1, ..., r_m)$, where $r_i : \mathcal{X}_i \to \mathcal{Y}_i \subset \mathbb{F}_i$ are Lipschitz continuous (that is, $\|r_i(x_i) - r_i(y_i)\| \le L_{r_i}\|x_i - y_i\|$ for $x_i, y_i \in \mathcal{X}_i$) and $\mathbb{F}_i$ ($i = 1, \ldots, m$) are finite dimensional real linear spaces, and
- $\phi : \mathcal{Y} := \mathcal{Y}_1 \times ... \times \mathcal{Y}_m \to \mathbb{R}_+$ is a continuously differentiable and block-wise concave function with Lipschitz gradient.

There are several practical problems in machine learning that minimize the objective functions of the form (26); see for example [13, 15, 36]. Considering f of the form (26), we propose to use the following composite surrogate function for f

$$u_i(x_i, y) = u_i^{\psi}(x_i, y) + \phi(r(y)) + \langle \nabla_i \phi(r(y)), r_i(x_i) - r_i(y_i) \rangle.$$

Indeed, since the block function of ϕ is concave, we have

$$(\phi \circ r)(x_i, y_{\neq i}) \leq \phi(r(y)) + \langle \nabla_i \phi(r(y)), r_i(x_i) - r_i(y_i) \rangle, \tag{27}$$

where $\langle \nabla_i \phi(r(y))$ is the gradient of ϕ at $r(y)$ with respect to block i. We will illustrate this case on the MCP in Section 6.

Let us assume the block surrogate functions $u_i^{\psi}(\cdot, \cdot)$ of $\psi(\cdot)$ satisfy Assumption 2.2(B). We prove that the block surrogate functions u_i of f also satisfy Assumption 2.2(B). Indeed, we have

$$\begin{aligned} h_i(x_i, y) &= u_i(x_i, y) - f_{\neq i}(x_i, y) \\ &= u_i^{\psi}(x_i, y) - \psi(x_i, y_{\neq i}) + \phi(r(y)) + \langle \nabla_i \phi(r(y)), r_i(x_i) - r_i(y_i) \rangle - \phi \circ r(x_i, y_{\neq i}). \end{aligned}$$

Moreover, as we assume $\nabla_i \phi$ is Lipschitz continuous, we have

$$\phi(r(y)) + \langle \nabla_i \phi(r(y)), r_i(x_i) - r_i(y_i) \rangle - (\phi \circ r)(x_i, y_{\neq i}) \leq \frac{L_i^{\phi}}{2} \|r_i(x_i) - r_i(y_i)\|^2,$$

for some constant L_i^{ϕ}. Therefore, we obtain

$$\begin{aligned} h_i(x_i, y) &\leq u_i^{\psi}(x_i, y) - \psi(x_i, y_{\neq i}) + \tfrac{L_i^{\phi}}{2} \|r_i(x_i) - r_i(y_i)\|^2 \\ &\leq u_i^{\psi}(x_i, y) - \psi(x_i, y_{\neq i}) + \tfrac{L_i^{\phi} (L_{r_i})^2}{2} \|x_i - y_i\|^2, \end{aligned} \tag{28}$$

where we use the Lipschitz continuity of $r_i(\cdot)$ in the last inequality. Since $u_i^{\psi}(\cdot, \cdot)$ satisfies Assumption 2.2(B), it follows from (28) that $u_i(\cdot, \cdot)$ satisfies Assumption 2.2(B).

Let us determine the values of ρ_i^k of Theorem 2.7 for the two cases (i) u_i^{ψ} satisfies Assumption 2.5, and (ii) $u_i^{\psi}(\cdot, y)$ satisfies Assumption 2.6 and $x_i \mapsto \langle \nabla_i \phi(r(y)), r_i(x_i) \rangle$ is convex. For the first case, we see that $u_i(x_i, y)$ also satisfies Assumption 2.5. Indeed, it follows from Inequality (27) that

$$h_i(x_i, y) \geq u_i^{\psi}(x_i, y) - \psi(x_i, y_{\neq i}) \geq \frac{\rho_i^{(y)}}{2} \|x_i - y_i\|^2.$$

For the second case, we see that $u_i(x_i, y) + g_i(x_i)$ is also a ρ_i^y-strongly convex function. The formulas of η_i^k and γ_i^k are determined as in Theorem 2.7. The values of A_i^k of Theorem 2.7 depends on how we choose block surrogate functions for ψ and then choose $\mathcal{G}_i^k$. Specific examples and their corresponding values of A_i^k that were presented in Section 4.2, Section 4.3 and Section 4.4 can be used for ψ.

Remark 4.2. Let us consider the case when $g_i(x_i)$ and $x_i \mapsto \langle \nabla_i \phi(r(y)), r_i(x_i) \rangle$, for $i \in [m]$, are convex, $\psi(x)$ is a block wise convex function and its block function $x_i \mapsto \psi(x_i, y_{\neq i})$ is continuously differentiable with $L_i^{(y)}$-Lipschitz gradient. We choose the Lipschitz gradient surrogate for ψ and choose $\mathcal{G}_i^k$ as in (20). Let $y = x^{k,i-1}$ and $L_i^k = L_i^{(x^{k,i-1})}$. Using the same technique as in the proof within [19, Remark 3] we get the

following inequality (we also can take $F = \psi(x) + \sum_{i=1}^m \big(\langle \nabla_i \phi(r(y)), r_i(x_i)\rangle + g_i(x_i)\big)$ in (22) to get the result)

$$\begin{aligned}&\psi(x^{k,i-1}) + \langle \nabla_i \phi(r(y)), r_i(x_i^k)\rangle + g_i(x_i^k) + \tfrac{L_i^k}{2}\big((\tau_i^k)^2 + \tfrac{(\beta_i^k - \tau_i^k)^2}{\nu}\big)\|x_i^k - x_i^{k-1}\|^2\\ &\geq \psi(x^{k,i}) + \langle \nabla_i \phi(r(y)), r_i(x_i^{k+1})\rangle + g_i(x_i^{k+1}) + \tfrac{(1-\nu)L_i^k}{2}\|x_i^{k+1} - x_i^k\|^2.\end{aligned}$$

Together with (27), we obtain

$$\begin{aligned}&\psi(x^{k,i-1}) + \phi(r(y)) + g_i(x_i^k) + \tfrac{L_i^k}{2}\big((\tau_i^k)^2 + \tfrac{(\beta_i^k - \tau_i^k)^2}{\nu}\big)\|x_i^k - x_i^{k-1}\|^2\\ &\geq \psi(x^{k,i}) + (\phi \circ r)(x_i^{k+1}, y_{\neq i}) + g_i(x_i^{k+1}) + \tfrac{(1-\nu)L_i^k}{2}\|x_i^{k+1} - x_i^k\|^2.\end{aligned} \tag{29}$$

Moreover, recall that $F(x) = \psi(x) + \phi(r(x)) + \sum_{i=1}^m g_i(x_i)$. Therefore, Inequality (29) recovers Inequality (22), and we can take η_i^k and γ_i^k as in (23).

5. Extension to essentially cyclic rule. In this section, we extend TITAN to allow the essentially cyclic rule in the block updates. Instead of cyclically updating the m blocks as in Algorithm 1, a block i_k among $\{1, \ldots, m\}$ is randomly or deterministically chosen to be updated at a time while fixing the values of the other blocks. The essentially cyclic rule with interval $T \geq m$ imposes that, for any T consecutive times of choosing an individual block, each of the m blocks must be updated at least once. Starting with two initial points x^{-1} and x^0, at iteration k, $k \geq 0$, TITAN with essentially cyclic rule will update x^k as follows.

$$x_{i_k}^{k+1} \in \operatorname*{argmin}_{x_{i_k} \in \mathcal{X}_{i_k}} \Big\{ u_{i_k}(x_{i_k}, x^k) - \langle \mathcal{G}_{i_k}^k(x_{i_k}^k, x_{i_k}^{prev}), x_{i_k}\rangle + g_{i_k}(x_{i_k}) \Big\}, \tag{30}$$

and set $x_a^{k+1} = x_a^k$ for all $a \neq i_k$. Here we use $x_{i_k}^{prev}$ to denote the value of block i_k before it was updated to $x_{i_k}^k$. To simplify the presentation of our upcoming analysis, we propose to use the following additional notations.

- Starting from x^0, we divide the generated sequence $\{x^k\}$ into consecutive intervals of T iterates and we use $\mathbf{x}^k$ to record the last iterate after every interval of T points of $\{x^k\}$, that is, $\mathbf{x}^k = x^{kT}$ for $k \geq 0$, and let $\mathbf{x}^{-1} = x^{-1}$.
- $\mathbf{x}^{k,j}$, $j = 1, \ldots, T$, are the points of $\{x^k\}$ lying between $\mathbf{x}^k$ and $\mathbf{x}^{k+1}$, that is, $\mathbf{x}^{k,j} = x^{kT+j}$.
- Since the values of a block can be unchanged in some consecutive iterations, we use $\bar{\mathbf{x}}_i^{k,l}$ to denote the value of block i after it has been updated l times during the k-th interval $[\mathbf{x}^k, \mathbf{x}^{k,1}, \ldots, \mathbf{x}^{k,T-1}, \mathbf{x}^{k+1} = \mathbf{x}^{k,T}]$. In other words, $\bar{\mathbf{x}}_i^{k,l}$ records the value of the i-th block when it is actually updated. The previous value of block i before it is updated to $\bar{\mathbf{x}}_i^{k,l}$ (which is $x_i^{k,j}$ for some j) is $\bar{\mathbf{x}}_i^{k,l-1}$ (which is $x_i^{k,j-1}$). Correspondingly, we use d_i^k to denote the total number of times the i-th block is updated during the k-th interval.
- $\mathbf{x}_{prev}^k$ stores the previous values of the blocks of $\mathbf{x}^k$, that is, $(\mathbf{x}_{prev}^{k+1})_i = \bar{\mathbf{x}}_i^{k,d_i^k - 1}$.

Using these notations, we express the generated sequence $\{x^n\}_{n\geq 0}$ as the sequence $\{\mathbf{x}^{k,j}\}_{k\geq 0, j=0,\ldots,T-1}$:

$$\ldots, \mathbf{x}^k = \mathbf{x}^{k,0}, \mathbf{x}^{k,1}, \ldots, \mathbf{x}^{k,T-1}, \mathbf{x}^{k+1} = \mathbf{x}^{k,T}, \ldots \tag{31}$$

So $x^n = \mathbf{x}^{k,j}$ with $k = [\frac{n}{T}]$, $j = n - kT$. Let us now translate Condition (4) using the new notations. The inequality in (4) for updating block i in the k-th period becomes

$$F(\mathbf{x}^{k,j-1}) + \frac{\gamma_i^{(\mathbf{x}^{k,j-1})}}{2}\|\mathbf{x}_i^{k,j-1} - x_i^{prev}\|^2 \geq F(\mathbf{x}^{k,j}) + \frac{\eta_i^{(\mathbf{x}^{k,j-1})}}{2}\|\mathbf{x}_i^{k,j} - \mathbf{x}_i^{k,j-1}\|^2. \tag{32}$$

Note that x_i^{prev}, $\mathbf{x}_i^{k,j-1}$ and $\mathbf{x}_i^{k,j}$ are three consecutive points of $\{\bar{\mathbf{x}}_i^{k,l}\}_{l=-1,\dots,d_i^k}$. We remark that $\bar{\mathbf{x}}_i^{k,-1} = (\mathbf{x}_{prev}^k)_i$. So if $\mathbf{x}_i^{k,j-1}$ is $\bar{\mathbf{x}}_i^{k,l-1}$ then $\bar{\mathbf{x}}_i^{k,l-2} = x_i^{prev}$ and $\bar{\mathbf{x}}_i^{k,l} = \mathbf{x}_i^{k,j}$. Inequality (32) is rewritten as

$$F(\mathbf{x}^{k,j-1}) + \frac{\bar{\gamma}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l-1} - \bar{\mathbf{x}}_i^{k,l-2}\|^2 \geq F(\mathbf{x}^{k,j}) + \frac{\bar{\eta}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2, \tag{33}$$

where $\bar{\eta}_i^{k,l-1} = \eta_i^{(\mathbf{x}^{k,j-1})}$ and $\bar{\gamma}_i^{k,l-1} = \gamma_i^{(\mathbf{x}^{k,j-1})}$. All the convergence results so far still hold for TITAN with the essentially cyclic update rule. For example, the following proposition has the same essence as Proposition 3.1.

PROPOSITION 5.1. *Considering TITAN with essentially cyclic rule, let $\{\mathbf{x}^{k,l}\}$ be the generated sequence of TITAN, see (31). Assume that the parameters are chosen such that the conditions in (33)(or its equivalent form in (32)) and Assumption 2.2 are satisfied. Furthermore, suppose for $k = 0, 1, \dots$ and $l = 1, \dots, d_i^k$ we have*

$$C\frac{\bar{\eta}_i^{k,l-1}}{2} \geq \frac{\bar{\gamma}_i^{k,l}}{2}, \tag{34}$$

for some constant $0 < C < 1$. Let $\bar{\eta}_i^{0,-1} = \bar{\gamma}_i^{0,0}/C$.

(A) We have

$$F(\mathbf{x}^K) + (1-C)\sum_{k=0}^{K-1}\sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\frac{\bar{\eta}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 \leq F(\mathbf{x}^0) + C\sum_{i=1}^{m}\frac{\bar{\eta}_i^{0,-1}}{2}\|\mathbf{x}_i^0 - \mathbf{x}_i^{-1}\|^2. \tag{35}$$

(B) If there exists positive number $\underline{l}$ such that $\min_{i,k,l}\{\frac{\bar{\eta}_i^{k,l}}{2}\} \geq \underline{l}$ then we have $\sum_{k=0}^{+\infty}\sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 < +\infty$.

Proof. See Appendix C.2 ∎

A subsequence $\{x^{k_n}\}$ of $\{x^n\}_{n\geq 0}$ when being expressed as $\mathbf{x}^{k,l}$ (see (31)) is $\{\mathbf{x}^{\bar{k}_n,l_n}\}$ with $\bar{k}_n = [\frac{k_n}{T}]$ and $l_n = k_n - T[\frac{k_n}{T}]$. We derive from Proposition 5.1 that if $\bar{\mathbf{x}}_i^{k,l_k}$ converges to x_i^* as k goes to 0, then $\bar{\mathbf{x}}_i^{k,l}$ also converges to x_i^* for $l = 1, \dots, d_i^k$. From this fact, we use the same technique in the proof of Theorem 3.2 to establish the subsequential convergence. We omit the details here.

For the global convergence, we define $\Phi^\delta(x, y) := \Phi(x) + \sum_{i=1}^m \frac{\delta_i}{2}\|x_i - y_i\|^2$, take

$$\varphi_k^2 = \sum_{i=1}^{m}\sum_{l=0}^{d_i^k}\tfrac{1}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 = \sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\tfrac{1}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 + \tfrac{1}{2}\|\mathbf{x}^k - \mathbf{x}_{prev}^k\|^2$$

and let $\mathbf{z}^k = (\mathbf{x}^k, \mathbf{x}_{prev}^k)$. Then, we have

$$\Phi^\delta(\mathbf{z}^k) - \Phi^\delta(\mathbf{z}^{k+1}) = F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) + \sum_{i=1}^{m}\tfrac{\delta_i}{2}\|\mathbf{x}_i^k - (\mathbf{x}_{prev}^k)_i\|^2 - \sum_{i=1}^{m}\tfrac{\delta_i}{2}\|\mathbf{x}_i^{k+1} - (\mathbf{x}_{prev}^{k+1})_i\|^2.$$

Similarly to Theorem 3.5 we assume there exists $\bar{l}$ such that $\max_{i,k,l}\{\frac{\bar{\eta}_i^{k,l}}{2}\} \leq \bar{l}$. Then, as for Theorem 3.5, we can prove that the whole sequence $\{\mathbf{x}^k\}$ converges to x^* when one of the two methods are used: $C < \underline{l}/\bar{l}$ or applying restarting step for (30). Hence each sequence $\{\mathbf{x}_i^k\}_{k\geq 0}$ converges to x_i^* for $i \in [m]$. Finally, note that

$$\begin{aligned}\|\mathbf{x}^{k,j-1} - x^*\|^2 &\leq (T-j+2)\big(\textstyle\sum_{a=j-1}^{T-1}\|\mathbf{x}^{k,a} - \mathbf{x}^{k,a+1}\|^2 + \|\mathbf{x}^{k+1} - x^*\|^2\big)\\ &\leq (T-j+2)\big(\textstyle\sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\|\bar{\mathbf{x}}_i^{k,l-1} - \bar{\mathbf{x}}_i^{k,l}\|^2 + \|\mathbf{x}^{k+1} - x^*\|^2\big).\end{aligned}$$

Together with Proposition 5.1(B) it implies that the whole sequence $\{x^k\}$ converges.

6. Numerical results. In this section, we apply TITAN to sparse NMF and the MCP. All tests are preformed using Matlab R2019a on a PC 2.3 GHz Intel Core i5 of 8GB RAM. The code is available from https://github.com/nhatpd/TITAN

6.1. Sparse Non-negative Matrix Factorization. Let us consider the following sparse NMF problem [35]

$$\min_{U,V}\Big\{\frac{1}{2}\|M-UV\|^2 : U\in\mathbb{R}_+^{\mathbf{m}\times\mathbf{r}}, V\in\mathbb{R}_+^{\mathbf{r}\times\mathbf{n}}, \|U_{:,i}\|_0\le s, i=1,\dots,\mathbf{r}\Big\}, \tag{36}$$

where $M\in\mathbb{R}_+^{\mathbf{m}\times\mathbf{n}}$ is a data matrix, $\mathbf{r}$ is a given positive integer, $U_{:,i}$ denotes the i-th column of U and $\|U_{:,i}\|_0$ denotes the number of non-zero entries of $U_{:,i}$. Problem (36) has the form of Problem (1) with $\mathcal{X}_1=\mathbb{R}^{\mathbf{m}\times\mathbf{r}}$, $\mathcal{X}_2=\mathbb{R}^{\mathbf{r}\times\mathbf{n}}$, $f(U,V)=\frac{1}{2}\|M-UV\|^2$, $g_1(U)$ is the indicator function of the set $\{U : U\in\mathbb{R}_+^{\mathbf{m}\times\mathbf{r}}, \|U_{:,i}\|_0\le s, i=1,\dots,\mathbf{r}\}$, and $g_2(V)$ is the indicator function of the set $\{V : V\in\mathbb{R}_+^{\mathbf{r}\times\mathbf{n}}\}$.

Noting that $\nabla_U f(U,V)=(UV-M)V^T$ is Lipschitz continuous with constant $L_1=\|VV^T\|$ and $\nabla_V f(U,V)=U^T(UV-M)$ is Lipschitz continuous with constant $L_2=\|U^TU\|$, we choose the block Lipschitz surrogate for f as in Section 4.2. Let us choose the Nesterov-type acceleration as in Section 4.2.1. The corresponding update in (3) for U is

$$U^{k+1}=\operatorname*{argmin}_U\big\langle\nabla_U f(\bar{U}^k,V^k),U\big\rangle+\frac{\kappa_1 L_1^k}{2}\|U-\bar{U}^k\|^2+g_1(U),$$

where $\kappa_1>1$ is a constant, $\bar{U}^k=U^k+\beta_1^k(U^k-U^{k-1})$, $L_1^k=\|V^k(V^k)^T\|$, and the corresponding update for V is

$$\begin{aligned}V^{k+1}&=\operatorname{argmin}_V\big\langle\nabla_V f(U^{k+1},\bar{V}^k),V\big\rangle+\tfrac{L_2^k}{2}\|V-\bar{V}^k\|^2+g_2(V)\\&=\big[\bar{V}^k-\tfrac{1}{L_2^k}\nabla_V f(U^{k+1},\bar{V}^k)\big]_+,\end{aligned}$$

where $\bar{V}^k=V^k+\beta_2^k(V^k-V^{k-1})$, $L_2^k=\|(U^{k+1})^TU^{k+1}\|$ and $[a]_+$ denotes $\max\{a,0\}$. It was shown in [12] that the update of U has the form

$$U^{k+1}=\mathcal{T}_s\Big(\big[\bar{U}^k-\frac{1}{\kappa_1 L_1^k}\nabla_U f(\bar{U}^k,V^k)\big]_+\Big),$$

where $\mathcal{T}_s(a)$ keeps the s largest values of a and sets the remaining values of a to zero.

Let us now determine η_i^k and γ_i^k, for $i=1,2$ (there are two blocks: U and V) of Condition (13). Note that $f(\cdot,V)$, $f(U,\cdot)$ and $g_2(V)$ are convex functions but $g_1(U)$ is non-convex. It follows from Section 4.2.1 that $\rho_1^k(V)=(\kappa_1-1)L_1^k$ and $A_1^k=\kappa_1\beta_1^kL_1^k$ for the block U surrogate functions. Applying Theorem 2.7, we get η_i^k and γ_i^k, and the condition (13) for block U becomes $\beta_1^k\le\frac{\kappa_1-1}{\kappa_1}\sqrt{\frac{C\nu_1(1-\nu_1)L_1^{k-1}}{L_1^k}}$, where $0<C,\nu_1<1$. Considering block V, as both $f(U,\cdot)$ and $g_2(V)$ are convex, it follows from Section 4.2 that $\gamma_2^k=\frac{1}{2}L_2^k(\beta_i^k)^2$ and $\eta_2^k=\frac{1}{2}(1-\nu_2)L_2^k$. Hence, the condition (13) for block V becomes $\beta_2^k\le\sqrt{\frac{C(1-\nu_2)L_2^{k-1}}{L_2^k}}$, where $0<C,\nu_2<1$. In our experiments, we choose

$$\begin{aligned}&C=0.9999^2,\mu_0=1,\mu_k=\tfrac{1}{2}(1+\sqrt{1+4\mu_{k-1}^2}),\nu_1=1/2,\nu_2=10^{-15},\\&\beta_1^k=\min\Big\{\tfrac{\mu_k-1}{\mu_k},\tfrac{\kappa_1-1}{\kappa_1}\sqrt{\tfrac{C\nu_1(1-\nu_1)L_1^{k-1}}{L_1^k}}\Big\},\beta_2^k=\min\Big\{\tfrac{\mu_k-1}{\mu_k},\sqrt{\tfrac{C(1-\nu_2)L_2^{k-1}}{L_2^k}}\Big\}.\end{aligned}$$

Since TITAN also works with essentially cyclic rule, in our experiment, we update U several times before updating V and vice versa. As explained in [16], repeating update U or V accelerates the algorithm compared to the cyclic update since the terms VV^T and MV^T in the gradient of U (resp. the terms U^TU and U^TM in the gradient of V) do not need to be re-evaluated hence the next evaluation of the gradient only requires $O(\mathbf{mr}^2)$ (resp. $O(\mathbf{nr}^2)$) operations in the update of U (resp. V) compared to $O(\mathbf{mnr})$ of the cyclic update. In our experiments, we test two values of κ_1: $\kappa_1 = 1.0001$ and $\kappa_1 = 1.5$. We use "TITAN - $\kappa = 1.0001$" and "TITAN - $\kappa = 1.5$" to denote the respective TITAN algorithms. As we do not use restarting, the two TITAN algorithms guarantee a sub-sequential convergence.

To verify the effect of inertial terms, we compare our TITAN algorithms with its non-inertial version, which is the proximal alternating linearized minimization (PALM) proposed in [12]. We test the algorithms on two image data sets cbclim[7] and Umist[8]. We choose $\mathbf{r} = 25$ and take a sparsity of s equal 25% (it means each basis face contains 25% non-zero pixels). For each data set, we run all the algorithms 30 times and use the same initialization, which is generated by the Matlab commands $W = rand(\mathbf{m}, \mathbf{r})$ and $H = rand(\mathbf{r}, \mathbf{n})$, for all algorithms in each run. We run each algorithm 50 seconds for the cbclim data set and 200 seconds for the Umist data set. We define the relative error as $\|M - UV\|_F/\|M\|_F$. We report the evolution with respect to time of the average values of $E(k) := \|M - U^kV^k\|_F/\|M\|_F - e_{\min}$, where $e_{\min}$ is the smallest value of all the relative errors in all runs, in Figure 1. We also report the average and the standard deviation (std) of the relative errors in Table 1.

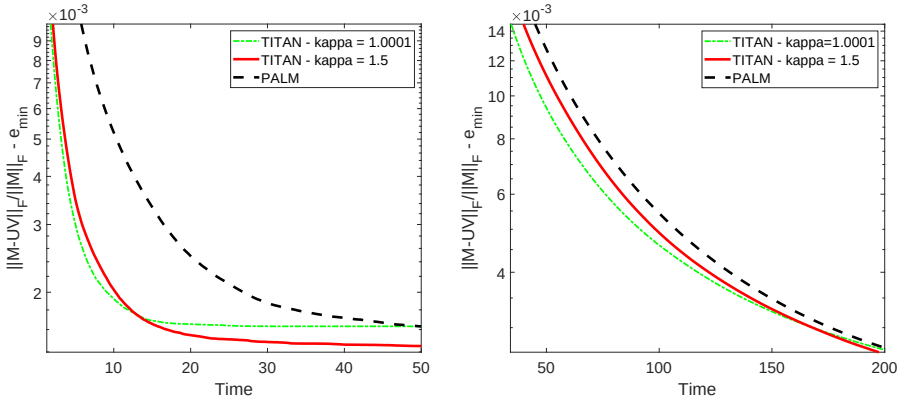

FIG. 1. *TITAN and PALM applied on sparse NMF. The plots show the evolution of the average value of $E(k)$ with respect to time on the cbclim data set (left) and the Umist data set (right).*

We observe that the two TITAN algorithms converges faster than PALM, in which TITAN - $\kappa = 1.0001$ converges initially the fastest. In term of the accuracy of solutions, TITAN - $\kappa = 1.5$ provides better relative errors in average.

6.2. Matrix Completion Problem. In this section, we illustrate the advantages of using block surrogate functions by deploying TITAN for the following MCP:

$$\min_{U\in\mathbb{R}^{\mathbf{m}\times\mathbf{r}}, V\in\mathbb{R}^{\mathbf{r}\times\mathbf{n}}} \left\{ \frac{1}{2}\|\mathcal{P}(A - UV)\|_F^2 + \mathcal{R}(U, V) \right\}, \tag{37}$$

[7] http://cbcl.mit.edu/software-datasets/FaceData2.html
[8] https://cs.nyu.edu/~roweis/data.html

TABLE 1
Average and std of relative errors obtained by TITAN and PALM applied on sparse NMF (36). Bold values correspond to the best results for each dataset

Dataset	Method	mean ± std
cbclim	PALM	$1.1960\,10^{-1} \pm 7.9357\,10^{-4}$
	TITAN - $\kappa = 1.0001$	$1.1961\,10^{-1} \pm 9.1071\,10^{-4}$
	TITAN - $\kappa = 1.5$	$\mathbf{1.1942\,10^{-1} \pm 7.9847\,10^{-4}}$
Umist	PALM	$1.1998\,10^{-1} \pm 9.8806\,10^{-4}$
	TITAN - $\kappa = 1.0001$	$1.2002\,10^{-1} \pm 1.0464\,10^{-4}$
	TITAN - $\kappa = 1.5$	$\mathbf{1.1989\,10^{-1} \pm 8.7035\,10^{-4}}$

where $A \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ is a given data matrix, R is a regularization term, and $\mathcal{P}(Z)_{ij} = Z_{ij}$ if A_{ij} is observed and is equal to 0 otherwise. The MCP (37) is one of the workhorse approaches in recommendation system; see, e.g., [24]. The application of MCP can also be found in sensor networks [9], social network analysis [22], and image processing [26]. We are interested in the exponential regularization (see, e.g., [13]) $\mathcal{R} = \phi \circ r$, where ϕ and r are given by

$$\begin{aligned} \phi(U,V) &= \lambda\Big(\textstyle\sum_{ij}\left(1-\exp(-\theta u_{ij})\right) + \sum_{ij}\left(1-\exp(-\theta v_{ij})\right)\Big), \\ r(U,V) &= (r_1(U), r_2(V)) = (|U|, |V|), \end{aligned} \tag{38}$$

where u_{ij} are elements of U and $|U|$ is the matrix whose elements are $|u_{ij}|$, λ and θ are tuning parameters. We see that Problem (37) in this case has the form of Problem (1) with $g = 0$, $\mathcal{X}_1 = \mathbb{R}^{\mathbf{m}\times\mathbf{r}}$, $\mathcal{X}_2 = \mathbb{R}^{\mathbf{r}\times\mathbf{n}}$ and $f(U,V) = \psi(U,V) + \phi(r(U,V))$, where $\psi(U,V) := \frac{1}{2}\|\mathcal{P}(A-UV)\|_F^2$ is the data-fitting term, as in Section 4.5. Since $\psi(U,V)$ is continuously differentiable and $\mathcal{R}(U,V)$ is a block separable function, hence F (in this case $F = f$) satisfies the condition in (2).

Since $\nabla_U\psi(U,V) = -\mathcal{P}(A-UV)V^T$ is Lipschitz continuous with constant $L_1 = \|VV^T\|$ and $\nabla_V\psi(U,V) = -U^T\mathcal{P}(A-UV)$ is Lipschitz continuous with constant $L_2 = \|U^TU\|$, we thus choose the block Lipschitz gradient surrogate functions u_i^{ψ}, $i = 1, 2$, in Section 4.2, for ψ. Moreover, ϕ is block-wise concave and differentiable with Lipschitz gradient on $\mathbb{R}_+^{\mathbf{m}\times\mathbf{n}}$. Hence, we select the composite surrogate function for $f = \psi + \phi\circ r$ as in Section 4.5. From Section 4.5 we have Assumption 2.2 is satisfied. Let us choose the Nesterov type acceleration. We note that $\nabla_U\phi(r(U^k,V^k)) = \lambda\theta\left(\exp(-\theta\|u_{ij}^k\|)\right)$. The update in (3) for U is

$$U^{k+1} \in \operatorname*{argmin}_U \left\langle \nabla_U\psi(\bar{U}^k, V^k), U\right\rangle + \frac{L_1^k}{2}\|U-\bar{U}^k\|^2 + \langle \nabla_U\phi(r(U^k,V^k)), |U|\rangle, \tag{39}$$

where $L_1^k = \|V^k(V^k)^T\|$, $\bar{U}^k = U^k + \beta_1^k(U^k - U^{k-1})$. The solution of (39) is given by

$$U^{k+1} = \mathcal{S}_{1/L_1^k}\left(P^k, \nabla_U\phi\left(g\left(U^k,V^k\right)\right)\right), \tag{40}$$

where $P^k = \bar{U}^k - \frac{1}{L_1^k}\nabla_U\psi(\bar{U}^k, V^k)$ and $\mathcal{S}_\tau$ is the soft-thresholding with parameter τ,

$$\mathcal{S}_\tau(P,W)_{ij} = [|p_{ij}| - \tau w_{ij}]_+ \text{sign}(p_{ij}). \tag{41}$$

Similarly, we derive the corresponding update for V as

$$V^{k+1} = \mathcal{S}_{1/L_2^k}\left(Q^k, \nabla_V\phi\left(r\left(U^{k+1},V^k\right)\right)\right), \tag{42}$$

where $L_2^k = \|(U^{k+1})^T U^{k+1}\|$, $Q^k = \bar{V}^k - \frac{1}{L_2^k}\nabla_V \psi(U^{k+1}, \bar{V}^k)$ and $\bar{V}^k = V^k + \beta_2^k(V^k - V^{k-1})$. Let us now determine η_i^k and γ_i^k, for $i = 1, 2$, of Condition (13). Note that $x_i \mapsto \langle \nabla_i \phi(r(y)), r_i(x_i)\rangle$ are convex, where x_i $(i = 1, 2)$ represent U and V. Furthermore, $\psi(U, V)$ is a block wise convex function. Therefore, it follows from Remark 4.2 that we can take we get η_i^k and γ_i^k as in (23). Note that $\tau_i^k = \beta_i^k$ since we choose Nesterov type acceleration. Then Condition (13) becomes $\beta_i^k \leq \sqrt{C(1-\nu)L_i^{k-1}/L_i^k}$, where $0 < C, \nu < 1$. In our experiments, we choose

$$C(1-\nu) = 0.9999^2, \mu_0 = 1, \mu_k = \tfrac{1}{2}(1 + \sqrt{1 + 4\mu_{k-1}^2}),$$
$$\beta_i^k = \min\left\{\tfrac{\mu_k - 1}{\mu_k}, \sqrt{C(1-\nu)L_i^{k-1}/L_i^k}\right\}.$$

We compare TITAN without extrapolation, which is denoted by TITAN-NO, and TITAN with the extrapolation, which is denoted by TITAN-EXTRA, with PALM that alternatively updates U and V by solving the following sub-problems

$$\min_U \left\langle \nabla_U \psi(U^k, V^k), U\right\rangle + \tfrac{L_1(V^k)}{2}\|U - U^k\|^2 + \lambda \textstyle\sum_{ij}\Big(1 - \exp(-\theta|u_{ij}|)\Big),$$
$$\min_V \left\langle \nabla_V \psi(U^{k+1}, V^k), V\right\rangle + \tfrac{L_2(U^{k+1})}{2}\|V - V^k\|^2 + \lambda \textstyle\sum_{ij}\Big(1 - \exp(-\theta|v_{ij}|)\Big).$$

These sub-problems can be separated into one-dimensional non-convex problems

$$\min_{x\in\mathbb{R}} \frac{1}{2}\|x - v\|^2 - \gamma \exp(-\theta|x|). \tag{43}$$

Although its solution can be computed via the LambertW function [23], it does not have a closed-form solution.

In our experiments, we choose $\lambda = 0.1$ and $\theta = 5$. We note that we do not optimize numerical results by tweaking the parameters as this is beyond the scope of this work. Rather, we simply chose the parameters that are typically used in the literature, see, e.g., [13]. It is important noting that we evaluate the algorithms on the same models. We carried out the experiments on the two most widely used datasets in the field of recommendation systems, MovieLens and Netflix, which contain ratings of different users. The characteristics of the datasets are given in Table 2. We respectively choose $\mathbf{r} = 5, 8$, and 13 for MovieLens 1M, 10M, and Netflix data set. We randomly used 70% of the observed ratings for training and the rest for testing. The process was repeated twenty times. We run each algorithm 20, 200, and 3600 seconds for MovieLens 1M, 10M, and Netflix data set, respectively. We are interested in the root mean squared error on the test set: $RMSE = \sqrt{\|\mathcal{P}_T(A - UV)\|^2/N_T}$, where $\mathcal{P}_T(Z)_{ij} = Z_{ij}$ if A_{ij} belongs to the test set and 0 otherwise, N_T is the number of ratings in the test set. We plotted the curves of the average value of RMSE and the objective function value (log scale) versus training time in Figure 2 and report the average and the standard deviation of RMSE and the objective function value in Table 3.

We observe that TITAN-EXTRA converges the fastest on all the data sets, providing a significant acceleration of TITAN-NO. TITAN-EXTRA achieves not only the best final objective function values but also the best RMSE on the test set. This illustrates the usefulness of the inertial terms. Moreover, TITAN-NO outperforms PALM on the three data sets which illustrates the usefulness of properly choosing the surrogate function.

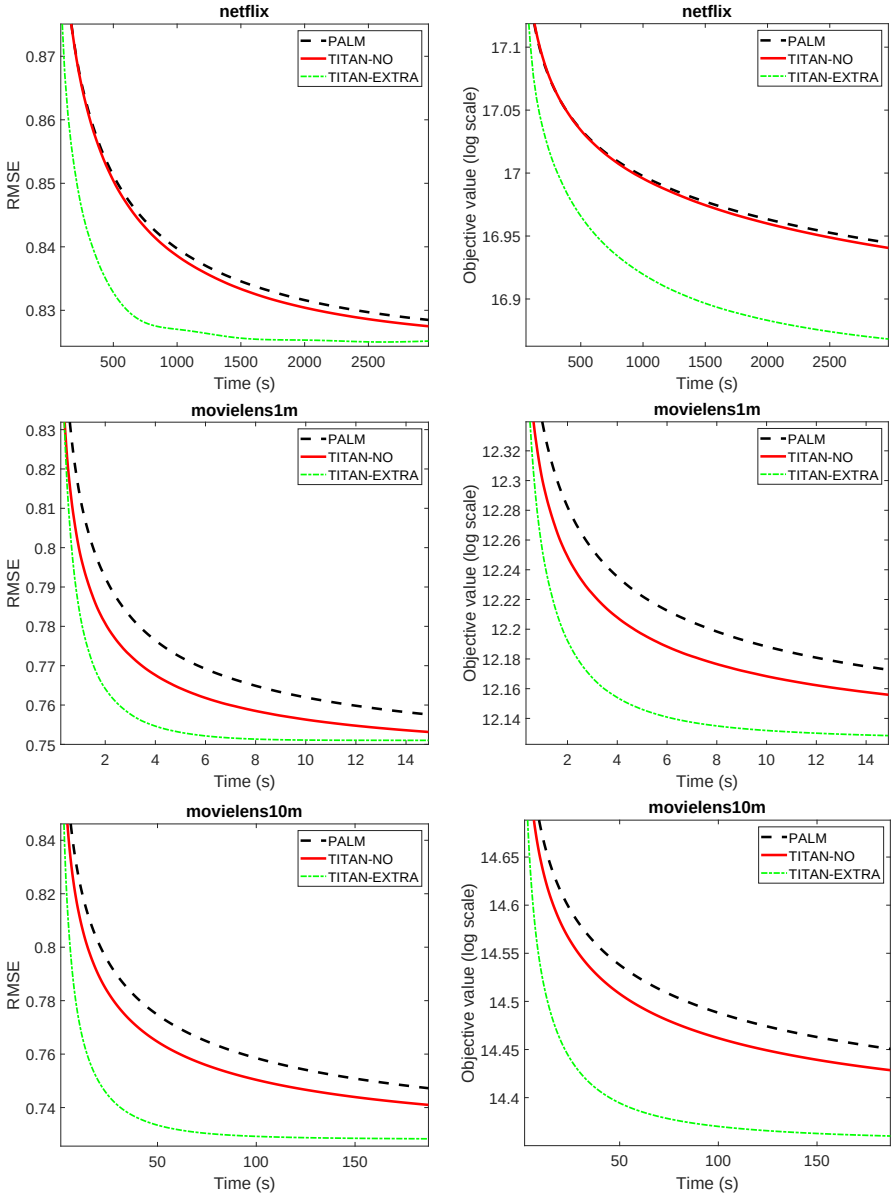

FIG. 2. *TITAN and PALM applied on the MCP (37). Evolution of the average value of the RMSE on the test set and the objective function value with respect to time*

TABLE 2
Dataset Descriptions. The number of users, items, and ratings used in each dataset

Dataset		#users	#items	#ratings
MovieLens	1M	6,040	3,449	999,714
	10M	69,878	10,677	10,000,054
Netflix		480,189	17,770	100,480,507

TABLE 3
Comparison of TITAN and PALM applied on the MCP (37): RMSE and final objective function values obtained within the allotted time. Bold values indicate the best results for each dataset.

Dataset	Method	RMSE mean ± std	Objective value (mean ± std)×10^{-5}
MovieLens 1M	PALM	0.7550 ± 0.0016	1.9155 ± 0.0088
	TITAN-NO	0.7514 ± 0.0013	1.8879 ± 0.0066
	TITAN-EXTRA	**0.7509** ± 0.0008	**1.8483** ± 0.0038
MovieLens 10M	PALM	0.7462 ± 0.0006	18.8038 ± 0.0348
	TITAN-NO	0.7402 ± 0.0006	18.4027 ± 0.0375
	TITAN-EXTRA	**0.7283** ± 0.0005	**17.2277** ± 0.0236
Netflix	PALM	0.8274 ± 0.0006	226.4846 ± 1.1898
	TITAN-NO	0.8265 ± 0.0006	225.4806 ± 1.1808
	TITAN-EXTRA	**0.8250** ± 0.0004	**210.4999** ± 0.3569

7. Conclusion. We have analysed TITAN, a novel inertial block majorization minimization algorithm that is a unified framework of many inertial block coordinate methods. We proved sub-sequential convergence of TITAN under mild assumptions and global convergence of TITAN under some stronger assumptions. We applied TITAN to the sparse NMF and MCP to illustrate the benefit of using inertial terms in BCD methods, and of using proper surrogate functions. Especially, the way we choose the surrogate functions and the corresponding extrapolation operators to derive the TITAN algorithms for the MCP strongly confirms the advantages of using TITAN algorithms compared to the typical proximal BCD methods. Our future research direction is to develop TITAN algorithms for solving more specific practical problems, for which using the typical proximal BCD methods may be not favorable.

Appendix A. Preliminaries of non-convex non-smooth optimization.
Let $g : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

DEFINITION A.1. *(i) 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 $\hat{\partial} g(x) = \emptyset$.
(ii) The limiting-subdifferential $\partial g(x)$ of g at $x \in \operatorname{dom} g$ is defined as follows.

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

DEFINITION A.2. *We call $x^* \in \operatorname{dom} \mathrm{F}$ a critical point of F if $0 \in \partial F\left(x^*\right)$.*

We note that if x^* is a local minimizer of F then x^* is a critical point of F.

DEFINITION A.3. *A function $\phi(x)$ is said to have the 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 $x \in U \cap [\phi(\bar{x}) < \phi(x) < \phi(\bar{x}) + \eta]$, we have*

$$\xi'\left(\phi(x) - \phi(\bar{x})\right) \operatorname{dist}\left(0, \partial \phi(x)\right) \geq 1. \tag{44}$$

$\operatorname{dist}(0, \partial\phi(x)) = \min\{\|y\| : y \in \partial\phi(x)\}$. *If $\phi(x)$ has the KL property at each point of* $\operatorname{dom}\partial\phi$ *then ϕ is a KL function.*

Many non-convex non-smooth functions in practical applications belong to the class of KL functions, for examples, real analytic functions, semi-algebraic functions, and locally strongly convex functions [10, 12].

Appendix B. Global convergence recipe.

THEOREM B.1. *[19, Theorem 2] 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 $\{z^k\}$ by $z^0 \in \mathbb{R}^N$, $z^{k+1} \in \mathcal{A}(z^k)$, $k = 0, 1, \ldots$ Assume that there exist positive constants ρ_1, ρ_2 and ρ_3 and a non-negative sequence $\{\varphi_k\}_{k\in\mathbb{N}}$ such that the following conditions are satisfied*

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

$$\rho_1\|z^k - z^{k+1}\|^2 \le \rho_2\varphi_k^2 \le \Phi(z^k) - \Phi(z^{k+1}), k = 0, 1, \ldots$$

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

$$\|\omega^{k+1}\| \le \rho_3\varphi_k, \omega^k \in \partial\Phi(z^k) \quad \textit{for} \quad k = 0, 1, \ldots$$

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

(B4) ***A continuity condition****: If a subsequence $\{z^{k_n}\}$ converges to $\bar{z}$ then $\Phi(z^{k_n})$ converges to $\Phi(\bar{z})$ as n goes to ∞.*

Then we have $\sum_{k=1}^{\infty}\varphi_k < \infty$ and $\{z^k\}$ converges to a critical point of Φ.

Appendix C. Technical proofs.

C.1. Proof of Theorem 3.5. Let us make some remarks before proving the theorem. Let x^* be a limit point of x^k. First, we note that when f is continuously differential we have $\partial\Phi(x) = \{\partial_{x_1}(F(x) + \mathcal{I}_{\mathcal{X}_1}(x_1)\} \times \ldots \times \{\partial_{x_m}F(x) + \mathcal{I}_{\mathcal{X}_m}(x_m)\}$ for all $x \in \mathcal{X}$. On the other hand, it follows from the fact that x_i^* is a minimizer of Problem (18), we have $0 \in \partial_{x_i}(F(x^*) + \mathcal{I}_{\mathcal{X}_i}(x_i^*))$. Hence, x^* is a critical point of Φ.

Second, note that since Assumption 2.2(B) is satisfied with $\bar{h}(x_i, y) = h(x_i, y)$ we have $\nabla_i u_i(y_i, y) = \nabla_i f(y) + \nabla_i h_i(y_i, y) = \nabla_i f(y_i, y_{\neq i})$ for a given $y \in \mathcal{X}$. Hence, the limiting subgradient $\partial_{x_i}\big(u_i(x_i, y) + \mathcal{I}_{\mathcal{X}_i}(x_i) + g_i(x_i)\big)$ at y_i is identical with the limiting subgradient $\partial_{x_i}\big(f(x_i, y_{\neq i}) + \mathcal{I}_{\mathcal{X}_i}(x_i) + g_i(x_i)\big)$ at y_i.

Let us now prove the theorem. As the generated sequence $\{x^k\}$ is assumed to be bounded, in the following, we only work on the bounded set that contains $\{x^k\}$.

Case 1: $C < \underline{l}/\bar{l}$. Define $\Phi^\delta(x, y) := \Phi(x) + \sum_{i=1}^m \frac{\delta_i}{2}\|x_i - y_i\|^2$. Let $z^k = (x^k, x^{k-1})$ and $\varphi_k^2 = \frac{1}{2}\|x^{k+1} - x^k\|^2 + \frac{1}{2}\|x^k - x^{k-1}\|^2$. We verify the conditions of Theorem B.1 for $\Phi^\delta(x^k, x^{k-1})$ with $\delta_i = (\underline{l} + C\bar{l})/2$.

(B1) *Sufficient decrease property.* From Inequality (15), we have

$$F(x^{k+1}) + \underline{l}\|x^{k+1} - x^k\|^2 \le F(x^k) + C\bar{l}\|x^k - x^{k-1}\|^2.$$

Hence, $\Phi^\delta(z^k) - \Phi^\delta(z^{k+1}) \ge (\underline{l} - C\bar{l})\varphi_k^2$.

(B2) *Boundedness of subgradient.* We note that

$$\partial_x\Phi^\delta(x, y) = \partial\Phi(x) + [\delta_i(x_i - y_i)|_{i=1,\ldots,m}], \quad \partial_y\Phi^\delta(x, y) = [\delta_i(y_i - x_i)|_{i=1,\ldots,m}]. \tag{45}$$

From (3) we have $0 \in \partial_{x_i}\big(u_i(x_i^{k+1}, x^{k,i-1}) + \mathcal{I}_{\mathcal{X}_i}(x_i^{k+1}) + g_i(x_i^{k+1}) - \mathcal{G}_i^k(x_i^k, x_i^{k-1})\big)$, which leads to $\mathcal{G}_i^k(x_i^k, x_i^{k-1}) \in \partial_{x_i}\big(u_i(x_i^{k+1}, x^{k,i-1}) + \mathcal{I}_{\mathcal{X}_i}(x_i^{k+1}) + g_i(x_i^{k+1})\big)$. Let $\mathbf{s}_i \in$

$\partial_{x_i}\big(u_i(x_i^{k+1}, x^{k+1}) - u_i(x_i^{k+1}, x^{k,i-1})\big)$. Then we have

$$\mathbf{s}_i + \mathcal{G}_i^k(x_i^k, x_i^{k-1}) \in \partial_{x_i}\big(u_i(x_i^{k+1}, x^{k+1}) + \mathcal{I}_{\mathcal{X}_i}(x_i^{k+1}) + g_i(x_i^{k+1})\big).$$

As we assume Assumption 3.4 holds, and x^{k+1} and $x^{k,i-1}$ are bounded, we have $\|\mathbf{s}_i\| \leq B_i\|x^{k+1} - x^{k,i-1}\| \leq B_i \sum_{j=i}^m \|x_j^{k+1} - x_j^k\|$. The boundedness of the subgradient is derived from the fact that $\partial_{x_i}\big(u_i(x_i^{k+1}, x^{k+1}) + \mathcal{I}_{\mathcal{X}_i}(x_i^{k+1}) + g_i(x_i^{k+1})\big) = \partial_{x_i}\Phi(x^{k+1})$ and $\|\mathbf{s}_i + \mathcal{G}_i^k(x_i^k, x_i^{k-1})\| \leq \|\mathbf{s}_i\| + A_i^k\|x_i^k - x_i^{k-1}\|$, where A_i^k is bounded.

(B3) *KL property.* As F is a KL function, Φ^δ is also a KL function.

(B4) *A continuity condition.* Suppose $z^{k_n} \to z^*$. From Proposition 3.1, we have that if x^{k_n} converges to x^* then x^{k_n-1} also converges to x^*. Hence $z^* = (x^*, x^*)$. On the other hand, similarly to the proof of Theorem 3.2, we can derive from (7) that, for $i \in [m]$, $u_i(x_i^{k_n}, x^{k_n-1,i-1}) + g_i(x_i^{k_n})$ converges to $u_i(x_i^*, x^*) + g_i(x_i^*)$. As we assume $u_i(\cdot,\cdot)$ is continuous we have $u_i(x_i^{k_n}, x^{k_n-1,i-1})$ converges to $u_i(x_i^*, x^*) = f(x^*)$. Hence, $g_i(x_i^{k_n}) \to g_i(x_i^*)$. We then have $F(x^{k_n}) = f(x^{k_n}) + \sum g_i(x_i^{k_n})$ converges to $F(x^*)$, which leads to $\Phi^\delta(z^{k_n+1})$ converges to $\Phi^\delta(z^*)$

Applying Theorem B.1, we get $0 \in \partial\Phi^\delta(x^*, x^*)$, which leads to $0 \in \partial\Phi(x^*)$.

Case 2: With restart. We use the technique in the proof of [12, Theorem 1] with some modification. A restarting step would be taken when $F(x^{k+1}) \geq F(x^k)$. When restarting happens, Condition (4) is satisfied with $\gamma_i^k = 0$, then Inequality (15) still holds. Thus the result in Proposition 3.1 and Theorem 3.2 do not change. So, as in the first case, we have $F(x^{k_n}) \to F(x^*)$. So, since $F(x^k)$ is non-increasing we have $F(x^k) \to F(x^*)$. This also means $\Phi(x)$ is constant on the set Ω of all limit points of x^k. From Proposition 3.1, we have $\|x^k - x^{k-1}\| \to 0$. Hence, [12, Lemma 5] yields that Ω is a compact and connected set.

Note that when restarting happens we can let $C = 0$ in Inequality (15). Therefore, as long as $x^{k+1} \neq x^k$, $F(x^k)$ is strictly decreasing (that is $F(x^{k+1}) < F(x^k)$). Hence, if there exists an integer $\bar{k}$ such that $F(x^{\bar{k}}) = F(x^*)$ then we have $F(x^k) = F(x^*)$ and $x^k = x^{\bar{k}}$ for all $k \geq \bar{k}$. So this case is trivial.

Let us consider $F(x^k) > F(x^*)$ for all k. Then there exists a positive integer k_0 such that $F(x^k) < F(x^*) + \eta$ for all $k > k_0$. On the other hand, there exists a positive integer k_1 such that $\text{dist}(x^k, \Omega) < \varepsilon$ for all $k > k_1$. Applying [12, Lemma 6] we have

$$\xi'\big(\Phi\left(x^k\right) - \Phi(x^*)\big)\,\text{dist}\left(0, \partial\Phi(x^k)\right) \geq 1, \text{for any } k > \mathbf{k} := \max\{k_0, k_1\}. \tag{46}$$

On the other hand, in the proof of Case 1, we proved that $\exists\varpi > 0$ such that for some $\omega^{k+1} \in \partial\Phi(x^{k+1})$ we have $\|\omega^{k+1}\| \leq \varpi\varphi_k \leq \frac{\varpi}{\sqrt{2}}(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|)$. Therefore, it follows from (46) that

$$\xi'\big(\Phi\left(x^k\right) - \Phi(x^*)\big)(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|) \geq \tfrac{\sqrt{2}}{\varpi} \tag{47}$$

From Inequality (15) we get

$$\Phi(x^k) - \Phi(x^{k+1}) \geq \textstyle\sum_{i=1}^m \frac{\eta_i^k}{2}\|x_i^{k+1} - x_i^k\|^2 - C\sum_{i=1}^m \frac{\eta_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|^2 \tag{48}$$

Denote $A_{i,j} = \xi(\Phi(x^i) - \Phi(x^*)) - \xi(\Phi(x^j) - \Phi(x^*))$. From the concavity of ξ we get $A_{k,k+1} \geq \xi'\big(\Phi\left(x^k\right) - \Phi(x^*)\big)\big(\Phi(x^k) - \Phi(x^{k+1})\big)$. Together with (47) and (48) we get

$$\sum_{i=1}^m \frac{\eta_i^k}{2}\|x_i^{k+1} - x_i^k\|^2 \leq C\sum_{i=1}^m \frac{\eta_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|^2 + \frac{\varpi}{\sqrt{2}}A_{k,k+1}(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|) \tag{49}$$

Denote $\Upsilon^k = \sum_{i=1}^m \frac{\eta_i^k}{2}\|x_i^{k+1} - x_i^k\|^2$. Using inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\sqrt{ab} \leq ta + b/4t$, for $t > 0$, from (49) we get

$$\begin{aligned}\sqrt{\Upsilon^k} &\leq \sqrt{C\Upsilon^{k-1}} + \sqrt{\tfrac{\varpi A_{k,k+1}}{\sqrt{2}}(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|)} \\ &\leq \sqrt{C\Upsilon^{k-1}} + \tfrac{(1-\sqrt{C})\sqrt{\underline{l}}}{3}(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|) + \tfrac{3\varpi A_{k,k+1}}{4\sqrt{2}\sqrt{\underline{l}}(1-\sqrt{C})}\end{aligned}$$

Summing up this inequality from $k = \mathbf{k}+1$ to K we obtain

$$\begin{aligned}\sqrt{\Upsilon^K} + \sum_{k=\mathbf{k}+1}^{K-1}(1-\sqrt{C})\sqrt{\Upsilon^k} \leq \sqrt{C\Upsilon^{\mathbf{k}}} + \frac{(1-\sqrt{C})\sqrt{\underline{l}}}{3}\sum_{k=\mathbf{k}+1}^{K}(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|) \\ + \frac{3\varpi}{4\sqrt{2}\sqrt{\underline{l}}(1-\sqrt{C})}A_{\mathbf{k}+1,K+1}.\end{aligned}$$

On the other hand, we note that $\sqrt{\Upsilon^k} \geq \sqrt{\underline{l}}\|x^{k+1} - x^k\|$. Therefore, we get

$$\frac{2}{3}(1-\sqrt{C})\sqrt{\underline{l}}\sum_{k=\mathbf{k}+1}^{K}\|x^{k+1} - x^k\| \leq \frac{(1-\sqrt{C})\sqrt{\underline{l}}}{3}\sum_{k=\mathbf{k}+1}^{K}\|x^k - x^{k-1}\| + \frac{3\varpi A_{\mathbf{k}+1,K+1}}{4\sqrt{2}\sqrt{\underline{l}}(1-\sqrt{C})},$$

which implies that $\sum_{k=\mathbf{k}+1}^{K}\|x^{k+1} - x^k\| \leq \|x^{\mathbf{k}+1} - x^{\mathbf{k}}\| + \frac{9\varpi}{4\sqrt{2}(1-\sqrt{C})^2\underline{l}}A_{\mathbf{k},K+1}$. Hence, $\sum_{k=1}^{\infty}\|x^{k+1} - x^k\| < +\infty$. The result follows.

C.2. Proof of Proposition 5.1. Let us prove Statement (A). Statement (B) of Proposition 5.1 is a consequence of Statement (A). From Inequality (33) we get

$$F(\mathbf{x}^{k,j}) + \frac{\bar{\eta}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 \leq F(\mathbf{x}^{k,j-1}) + \frac{C\bar{\eta}_i^{k,l-2}}{2}\|\bar{\mathbf{x}}_i^{k,l-1} - \bar{\mathbf{x}}_i^{k,l-2}\|^2. \tag{50}$$

Summing up Inequality (50) from $j = 1$ to T we obtain

$$F(\mathbf{x}^{k+1}) + \sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\frac{\bar{\eta}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 \leq F(\mathbf{x}^k) + C\sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\frac{\bar{\eta}_i^{k,l-2}}{2}\|\bar{\mathbf{x}}_i^{k,l-1} - \bar{\mathbf{x}}_i^{k,l-2}\|^2.$$

Therefore,

$$\begin{aligned}&F(\mathbf{x}^{k+1}) + C\sum_{i=1}^{m}\frac{\bar{\eta}_i^{k,d_i^k-1}}{2}\|\bar{\mathbf{x}}_i^{k,d_i^k} - \bar{\mathbf{x}}_i^{k,d_i^k-1}\|^2 + (1-C)\sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\frac{\bar{\eta}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 \\ &\leq F(\mathbf{x}^k) + C\sum_{i=1}^{m}\frac{\bar{\eta}_i^{k,-1}}{2}\|\bar{\mathbf{x}}_i^{k,0} - \bar{\mathbf{x}}_i^{k,-1}\|^2.\end{aligned} \tag{51}$$

Note that $\bar{\mathbf{x}}_i^{k,0} = \bar{\mathbf{x}}_i^{k-1,d_i^{k-1}}$, $\bar{\mathbf{x}}_i^{k,-1} = \bar{\mathbf{x}}_i^{k-1,d_i^{k-1}-1} = (\mathbf{x}_{prev}^{k-1})_i$ and $\bar{\eta}_i^{k+1,-1} = \bar{\eta}_i^{k,d_i^k}$. Hence, from (51) we obtain

$$\begin{aligned}&F(\mathbf{x}^{k+1}) + C\sum_{i=1}^{m}\frac{\bar{\eta}_i^{k+1,-1}}{2}\|\mathbf{x}_i^{k+1} - (\mathbf{x}_{prev}^{k+1})_i\|^2 + (1-C)\sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\frac{\bar{\eta}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 \\ &\leq F(\mathbf{x}^k) + C\sum_{i=1}^{m}\bar{\eta}_i^{k,-1}\|\mathbf{x}_i^k - (\mathbf{x}_{prev}^k)_i\|^2.\end{aligned} \tag{52}$$

Summing up Inequality (52) from $k = 0$ to $K-1$ we get

$$\begin{aligned}&F(\mathbf{x}^K) + C\sum_{i=1}^{m}\tfrac{\bar{\eta}_i^{K,-1}}{2}\|\mathbf{x}_i^K - (\mathbf{x}_{prev}^K)_i\|^2 + (1-C)\textstyle\sum_{k=0}^{K-1}\sum_{i=1}^{m}\sum_{l=1}^{d_i^k}\frac{\bar{\eta}_i^{k,l-1}}{2}\|\bar{\mathbf{x}}_i^{k,l} - \bar{\mathbf{x}}_i^{k,l-1}\|^2 \\ &\leq F(\mathbf{x}^0) + C\sum_{i=1}^{m}\tfrac{\bar{\eta}_i^{0,-1}}{2}\|\mathbf{x}_i^0 - (\mathbf{x}_{prev}^0)_i\|^2,\end{aligned}$$

which gives the result.

REFERENCES

[1] S. Adly and H. Attouch. Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping. *SIAM Journal on Optimization*, 30(3):2134–2162, 2020.

[2] M. Ahookhosh, L. T. K. Hien, N. Gillis, and P. Patrinos. Multi-block Bregman proximal alternating linearized minimization and its application to sparse orthogonal nonnegative matrix factorization. arXiv:1908.01402, 2019.

[3] M. Ahookhosh, L. T. K. Hien, N. Gillis, and P. Patrinos. A block inertial Bregman proximal algorithm for nonsmooth nonconvex problems with application to symmetric nonnegative matrix tri-factorization. *arXiv:2003.03963*, 2020.

[4] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, Jan 2009.

[5] 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.

[6] 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.

[7] 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.

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

[9] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Trans. Sen. Netw.*, 2(2):188–220, 2006.

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

[11] 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.

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

[13] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceeding of international conference on machine learning ICML'98*, 1998.

[14] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward–backward algorithm. *Journal of Global Optimization*, 66:457–485, 2016.

[15] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Stat. Ass.*, 96(456):1348–1360, 2001.

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

[17] 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.

[18] L. T. K. Hien and N. Gillis. Algorithms for nonnegative matrix factorization with the Kullback-Leibler divergence. arXiv:2010.01935, 2020.

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

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

[21] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163:85–114, 2017.

[22] M. Kim and J. Leskovec. The network completion problem: Inferring missing nodes and edges in networks. In *Proceedings of the 11th International Conference on Data Mining*, pages 47–58, 2011.

[23] R. M. C. G. H. G. D. E. G. H. D. J. J. D. E. Knuth. On the lambertwfunction. *Advances in Computational Mathematics*, 5, 1996.

[24] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

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

[26] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[27] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

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

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

[30] 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.

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

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

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

[34] 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.

[35] R. Peharz and F. Pernkopf. Sparse nonnegative matrix factorization with $\ell 0$-constraints. *Neurocomputing*, 80:38 – 46, 2012. Special Issue on Machine Learning for Signal Processing 2010.

[36] D. N. Phan and H. A. Le Thi. Group variable selection via $\ell_{p,0}$ regularization and application to optimal scoring. *Neural Networks*, 118:220 – 234, 2019.

[37] 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.

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

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

[40] 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.

[41] M. Teboulle and Y. Vaisbourd. Novel proximal gradient methods for nonnegative matrix factorization with sparsity constraints. *SIAM Journal on Imaging Sciences*, 13(1):381–421, 2020.

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

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

[44] 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.

[45] 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.

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