Math. Program., Ser. A (2017) 163:85–114
DOI 10.1007/s10107-016-1057-8

FULL LENGTH PAPER

Iteration complexity analysis of block coordinate descent methods

Mingyi Hong[1] · Xiangfeng Wang[2] · Meisam Razaviyayn[3] · Zhi-Quan Luo[4,5]

Received: 28 April 2015 / Accepted: 16 July 2016 / Published online: 19 August 2016

Abstract In this paper, we provide a unified iteration complexity analysis for a family of general block coordinate descent methods, covering popular methods such as the block coordinate gradient descent and the block coordinate proximal gradient, under

The authors would like to thanks Enbin Song for his valuable comments on an early version of the manuscript.

Mingyi Hong: This author is supported by National Science Foundation (NSF) Grant CCF-1526078 and by Air Force Office of Scientific Research (AFOSR) Grant 15RT0767.
Xiangfeng Wang: This author is supported by NSFC, Grant No.11501210, and by Shanghai YangFan, Grant No. 15YF1403400.
Zhi-Quan Luo: This research is supported by NSFC, Grant No. 61571384, and by the Leading Talents of Guang Dong Province program, Grant No. 00201510.

✉ Mingyi Hong
mingyi@iastate.edu

Xiangfeng Wang
xfwang@sei.ecnu.edu.cn

Meisam Razaviyayn
meisam@stanford.edu

Zhi-Quan Luo
luozq@umn.edu

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA, USA

[2] Shanghai Key Laboratory of Trustworthy Computing, School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China

[3] Department of Electrical Engineering, Stanford University, Palo Alto, CA, USA

[4] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

[5] Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA

various different coordinate update rules. We unify these algorithms under the so-called block successive upper-bound minimization (BSUM) framework, and show that for a broad class of multi-block nonsmooth convex problems, all algorithms covered by the BSUM framework achieve a global sublinear iteration complexity of $\mathcal{O}(1/r)$, where r is the iteration index. Moreover, for the case of block coordinate minimization where each block is minimized exactly, we establish the sublinear convergence rate of $O(1/r)$ without per block strong convexity assumption.

Mathematics Subject Classification 49-90

1 Introduction

Consider the problem of minimizing a nonsmooth convex function $f(x)$:

$$\begin{aligned} &\text{minimize } \ f(x) := g\left(x_1, \ldots, x_K\right) + \sum_{k=1}^{K} h_k(x_k) \\ &\text{subject to } x_k \in X_k, \quad k = 1, \ldots, K, \end{aligned} \tag{1.1}$$

where $g(\cdot)$ is a smooth convex function; h_k is a closed nonsmooth convex function (possibly with extended values); $x = (x_1^T, \ldots, x_K^T)^T \in \mathbb{R}^n$ is a partition of the optimization variable x, with $x_k \in X_k \subseteq \mathbb{R}^{n_k}$. Let $X := \prod_{k=1}^{K} X_k$ denote the feasible set for x.

A well known family of algorithms for solving (1.1) is the block coordinate descent (BCD) type method whereby, at every iteration a single block of variables is optimized while the remaining blocks are held fixed. One of the best known algorithms in the BCD family is the block coordinate minimization (BCM) algorithm, where at iteration r, the blocks are updated by solving the following problem exactly [3]

$$x_k^r \in \arg\min_{x_k \in X_k} \ g\left(x_1^r, \ldots, x_{k-1}^r, x_k, x_{k+1}^{r-1}, \ldots, x_K^{r-1}\right) + h_k(x_k), \ \forall\, k. \tag{1.2}$$

When problem (1.2) is not easily solvable, a popular variant is to solve an approximate version of problem (1.2), yielding the block coordinate gradient descent (BCGD) algorithm, or the block coordinate proximal gradient (BCPG) algorithm in the presence of a nonsmooth function [2,28,32,36]. In particular, at a given iteration r, the following problem is solved for each block k:

$$\begin{aligned} x_k^r = \arg\min_{x_k \in X_k} \ &\left\langle \nabla_k g\left(x_1^r, \ldots, x_{k-1}^r, x_k^{r-1}, \ldots, x_K^{r-1}\right), x_k - x_k^{r-1} \right\rangle \\ &+ \frac{L_k}{2}\|x_k - x_k^{r-1}\|^2 + h_k(x_k), \end{aligned}$$

where $L_k > 0$ is some appropriately chosen constant. Other variants of the BCD-type algorithm include those that solve different subproblems [24], or those with different block selection rules, such as the Gauss–Seidel (G–S) rule, the Gauss–Southwell (G–So) rule [31], the randomized rule [22], the essentially cyclic (E-C) rule [29], or the maximum block improvement (MBI) rule [6].

In all the above mentioned variants of BCD method, each step involves solving a simple subproblem of small size, therefore the BCD method can be quite effective for solving large-scale problems; see e.g., [10,22,24,26,28] and the references therein. The existing analysis of the BCD method [4,5,23,29] requires the uniqueness of the minimizer for each subproblem (1.2), or the pseudo-convexity of f [11]. Recently, a unified BCD-type framework, termed the block successive upper-bound minimization (BSUM) method, was proposed in [24]. At each iteration of the BSUM method, certain approximate function of the per-block subproblem (1.2) is constructed and optimized. Due to the flexibility in choosing the approximate function, the BSUM includes many BCD-type algorithms as special cases. It is shown in [24] that the method converges to stationary solutions for nonconvex problems and to global optimal solutions for convex problems, as long as certain regularity conditions are satisfied for the per-block subproblems.

The global rate of convergence for BCD-type algorithm has been studied extensively. When the objective function is strongly convex, the BCM algorithm converges globally linearly [18]. When the objective function is smooth and not strongly convex, Luo and Tseng have shown that the BCD method with the classic G–S/G–So update rules converges linearly, provided that a certain local error bound is satisfied around the solution set [16–19]. In addition, such linear rate is global when the feasible set is compact. This line of analysis has recently been extended to allow certain class of nonsmooth functions in the objective [30,36]. For more general problems where the objective is not strongly convex and the error bound condition does not hold, several recent studies have established the $\mathcal{O}(1/r)$ iteration complexity for various BCD-type algorithms including the randomized BCGD algorithm [22], and for more general settings with nonsmooth objective as well [15,25,28]. When the coordinates are updated according to the traditional G–S/G–So/E-C rule, however, the literature on the iteration complexity for the BCD-type algorithm is scarce. In [26], Saha and Tewari have proven the $\mathcal{O}(1/r)$ rate for the G–S BCPG algorithm when applied to certain special ℓ_1 minimization problem. In [2], Beck and Tetruashvili have shown the $\mathcal{O}(1/r)$ sublinear convergence for the G–S BCGD algorithm for constrained smooth problems. In [1], Beck has shown the sublinear convergence for the G–S BCM algorithm (termed Alternating Minimization method therein) when the number of blocks is two. Although the BCD-type algorithm with G–S rule sometimes has been found to perform better than its randomized counterpart (see, e.g., [26]), establishing its iteration complexity in a general multi-block nonsmooth setting is challenging [22]. To the best of our knowledge, the iteration complexity of the BCD-type algorithm with the classic G–S update rule has not yet been characterized for multi-block nonsmooth problems, not to mention other types of deterministic coordinate selection rules such as G–So, E-C or MBI. Further, there has been no iteration complexity analysis for the classic BCM iteration (1.2) when the number of variable blocks is more than two (i.e., $K \geq 3$).

In this paper, we provide a unified iteration complexity analysis for K-block BCD-type algorithm by utilizing the BSUM framework [24]. Our result covers many different BCD-type algorithms such as BCM, BCPG, and BCGD under a number of deterministic coordinate update rules. First, for a broad class of nonsmooth convex problems, we show that the BSUM algorithm achieves a global sublinear convergence

Table 1 Summary of the Results

Method	Update Rule	Problem	Assumptions	Rate
BSUM	**G–S/E-C**	NS+C+K	u_k valid upper-bound	$\mathcal{O}(1/r)$
BSUM	**G–So/MBI**	NS+C+K	u_k valid upper-bound, h Lipschitz	$\mathcal{O}(1/r)$
BSUM	**G–S**	NS+C+2	u_1 valid upper-bound without BSC, $u_2 = g$	$\mathcal{O}(1/r)$
BSUM	N/A	NS+C+1	u_1 valid upper-bound without BSC	$\mathcal{O}(1/r)$
BCM	**MBI**	NS+C+K	h Lipschitz, without BSC	$\mathcal{O}(1/r)$
BCM	**G–S/E-C**	NS+C+K	$u_k = g$, without BSC	$\mathcal{O}(1/r)$

rate of $\mathcal{O}(1/r)$, provided that *each subproblem* is strongly convex. Second, for the BCM algorithm (1.2), we show the global convergence rate of $\mathcal{O}(1/r)$ without the per-block strong convexity assumption. The main results of this paper are summarized in the following Table 1.[1]

Notations: For a given matrix A, we use $A[i, j]$ to denote its (i, j)th element. For a symmetric matrix A use $\rho_{\max}(A)$ to denote its spectral norm. For a given vector x, we use $x[j]$ to denote its jth component; use $\|x\|$ to denote its ℓ_2 norm. We use $I_X(\cdot)$ to denote the indicator function for a given set X, i.e., $I_X(y) = 1$ if $y \in X$, and $I_X(y) = \infty$ if $y \notin X$. Let x_{-k} denote the vector x with x_k removed. For a given function $f(x_1, \ldots, x_K)$ which contains K block variables, we use $\nabla_k f(x_1, \ldots, x_K)$ to denote the partial gradient with respect to its kth block variable. We use ∂f to denote the subdifferential set of a function f. For a given convex nonsmooth closed function $\ell(\cdot)$, we define the proximity operator $\text{prox}_\ell(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^n$ as

$$\text{prox}_\ell^\beta(x) = \text{argmin}_{u\in\mathbb{R}^n} \ \ \ell(u) + \frac{\beta}{2}\|x - u\|^2.$$

Similarly, for a given closed convex set X, the projection operator $\text{proj}_X(\cdot) : \mathbb{R}^n \mapsto X$ is defined as

$$\text{proj}_X(x) = \text{argmin}_{u\in X} \ \ \frac{1}{2}\|x - u\|^2.$$

2 The BSUM algorithm and preliminaries

2.1 The BSUM algorithm

In this paper, we consider a family of block coordinate descent methods (BCD) for solving problem (1.1). The family of the algorithms we consider falls in the general category of block successive upper-bound minimization (BSUM) method, in which

[1] We have used the following abbreviations: NS = **N**on**s**mooth, C = **C**onstrained, K = **K**-block, BSC = **B**lock-wise **S**trongly **C**onvex, G–So = **G**auss–**So**uthwell, G–S = **G**auss–**S**eidel, E-C = **E**ssentially **C**yclic, MBI = **M**aximum **B**lock **I**mprovement. The notion of *valid upper-bound* as well as the function u_k will be introduced in Sect. 2.

certain *approximate version* of the objective function is optimized one block variable at a time, while fixing the rest of the block variables [24]. In particular, at iteration $r+1$, we first pick an index set $\mathcal{C}^{r+1} \subseteq \{1, \ldots, K\}$. Then the kth block variable is updated by

$$x_k^{r+1} \begin{cases} \in \arg\min_{x_k \in X_k} \; u_k\left(x_k; x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^r, \ldots, x_K^r\right) + h_k(x_k), & \text{if } k \in \mathcal{C}^{r+1}; \\ = x_k^r, & \text{if } k \notin \mathcal{C}^{r+1}, \end{cases} \tag{2.1}$$

where $u_k(\cdot; x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^r, \ldots, x_K^r)$ is an approximation of $g(x)$ at a given iterate $(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^r, \ldots, x_K^r)$. We will see shortly that by properly specifying the approximation function $u_k(\cdot)$ as well as the index set $\mathcal{C}^{r+1}$, we can recover many popular BCD-type algorithms such as the BCM, the BCGD, the BCPG methods and so on.

To simplify notations, let us define a set of auxiliary variables

$$\begin{aligned} w_k^r &:= \left[x_1^r, \ldots, x_{k-1}^r, x_k^{r-1}, x_{k+1}^{r-1}, \ldots, x_K^{r-1}\right], & k &= 1, \ldots, K, \\ w_{-k}^r &:= \left[x_1^r, \ldots, x_{k-1}^r, x_{k+1}^{r-1}, \ldots, x_K^{r-1}\right], & k &= 1, \ldots, K, \\ x_{-k} &:= \left[x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K\right], & k &= 1, \ldots, K. \end{aligned}$$

Clearly we have $w_{K+1}^r := x^r, \; w_1^r := x^{r-1}$. Moreover, at each iteration $r+1$, define a set of new variables $\{\hat{x}_k^{r+1}\}_{k=1}^K$ as follows

$$\hat{x}_k^{r+1} \in \arg\min_{x_k \in X_k} \; u_k\left(x_k; x^r\right) + h_k(x_k), \;\; k = 1, \ldots, K. \tag{2.2}$$

Clearly $\{\hat{x}_k^{r+1}\}_{k=1}^K$ represents a "virtual" update where all variables are optimized in a Jacobi manner based on x^r.

The BSUM algorithm is described formally in the following table.

```
The Block Successive Upper-Bound Minimization (BSUM) Algorithm
At each iteration $r+1$, pick an index set $\mathcal{C}^{r+1}$;
For $k = 1, \ldots, K$, do:
    $x_k^{r+1} \begin{cases} \in \arg\min_{x_k \in X_k} \; u_k\left(x_k; w_k^{r+1}\right) + h_k(x_k), & \text{if } k \in \mathcal{C}^{r+1}; \\ = x_k^r, & \text{if } k \notin \mathcal{C}^{r+1} \end{cases}.$
End For.
```

In this paper, we consider the following well-known block selection rules:

1. *Gauss–Seidel (G–S) rule* At each iteration $r+1$ all the indices are chosen, i.e., $\mathcal{C}^{r+1} = \{1, \ldots, K\}$. Using this rule, the blocks are updated cyclically with fixed order.

2. *Essentially cyclic (E-C) rule* There exists a given period $T \geq 1$ during which each index is updated at least once, i.e.,

$$\bigcup_{i=1}^{T} \mathcal{C}^{r+i} = \{1, \ldots, K\}, \ \forall\, r. \tag{2.3}$$

We call this update rule a *period-T* essentially cyclic update rule. Clearly when $T = 1$ we recover the G–S rule.

3. *Gauss–Southwell (G–So) rule* At each iteration $r+1$, $\mathcal{C}^{r+1}$ contains a single index k^* that satisfies:

$$k^* \in \left\{ k \,\middle|\, \|\hat{x}_k^{r+1} - x_k^r\| \geq q \max_j \|\hat{x}_j^{r+1} - x_j^r\| \right\}, \tag{2.4}$$

for some constant $q \in (0,\ 1]$.

4. *Maximum block improvement (MBI) rule* At each iteration $r+1$, $\mathcal{C}^{r+1}$ contains a single index k^* that satisfies:

$$k^* \in \arg\max_k -f\left(x_{-k}^r, \hat{x}_k^{r+1}\right), \tag{2.5}$$

where $f(x_{-k}^r, \hat{x}_k^{r+1})$ means the function $f(\cdot)$ evaluated at the vector $[x_{-k}^r, \hat{x}_k^{r+1}]$.

2.2 Main assumptions

Suppose f is a closed proper convex function in $\mathbb{R}^n$. Let dom f denote the effective domain of f and let int(dom f) denote the interior of dom f. We make the following standing assumptions regarding problem (1.1):

Assumption A

(a) Problem (1.1) is a convex problem, and its global minimum is attained. The intersection $X \cap \text{int}(\text{dom } f)$ is nonempty. Each h_k is a proper closed and convex function (not necessarily smooth), and its subdifferential set is nonempty at the relative boundary point of X_k.

(b) The gradient $\nabla g(\cdot)$ is block-wise uniformly Lipschitz continuous

$$\begin{aligned} &\|\nabla_k g\left(\left[x_{-k}, x_k\right]\right) - \nabla_k g\left(\left[x_{-k}, x_k'\right]\right)\| \\ &\quad \leq M_k \|x_k - x_k'\|, \ \forall\, x_k, x_k' \in X_k, \ \forall\, x \in X, \ \forall\, k \end{aligned} \tag{2.6}$$

where $M_k > 0$ is a constant. Define $M_{\max} = \max_k M_k$.
The gradient of $g(\cdot)$ is also uniformly Lipschitz continuous

$$\|\nabla g(x) - \nabla g(x')\| \leq M\|x - x'\|, \qquad \forall\, x, x' \in X \tag{2.7}$$

where $M > 0$ is a constant.

Next we make the following assumptions regarding the approximation function $u_k(\cdot;\cdot)$ in (2.1).

Assumption B (a) $u_k(x_k; x) = g(x), \ \forall\, x \in X, \ \forall\, k.$
(b) $u_k(v_k; x) \geq g(v_k, x_{-k}), \ \forall\, v_k \in X_k, \ \forall\, x \in X, \ \forall\, k.$
(c) $\nabla u_k(x_k; x) = \nabla_k g(x), \ \forall\, x \in X, \ \forall\, k$, where the notation $\nabla u_k(x_k; x)$ represents the partial derivative with respect to x_k.
(d) $u_k(v_k; x)$ is continuous in v_k and x. Further, for any given $x \in X$, it is a proper, closed and strongly convex function of v_k, satisfying

$$u_k(v_k; x) \geq u_k(\hat{v}_k; x) + \left\langle \nabla u_k(\hat{v}_k; x), v_k - \hat{v}_k \right\rangle + \frac{\gamma_k}{2}\|v_k - \hat{v}_k\|^2, \ \forall\, v_k, \ \hat{v}_k \in X_k,$$

where $\gamma_k > 0$ is independent of the choice of x.
(e) For any given $x \in X$, $u_k(v_k; x)$ has Lipschitz continuous gradient, that is

$$\|\nabla u_k(v_k; x) - \nabla u_k(\hat{v}_k; x)\| \leq L_k \|v_k - \hat{v}_k\|, \ \forall\, \hat{v}_k, \ v_k \in X_k, \ \forall\, k, \tag{2.8}$$

where $L_k > 0$ is some constant. Further, we have

$$\|\nabla u_k(v_k; x) - \nabla u_k(v_k; y)\| \leq G_k \|x - y\|, \ \forall\, v_k \in X_k, \ \forall\, k, \ \forall\, x, y \in X. \tag{2.9}$$

Define $L_{\max} := \max_k L_k$; $G_{\max} := \max_k G_k$.

We refer to the u_k's that satisfy Assumption B as a *valid upper-bound*.

A few remarks are in order regarding to the assumptions made above.

First of all, Assumption B indicates that for any given x, each $u_k(\cdot; x)$ is a locally tight upper bound for $g(x)$. When the approximation function is chosen as the original function $g(x)$, then we recover the classic BCM algorithm; cf. (1.2). In many practical applications especially nonsmooth problems, minimizing the approximation functions often leads to much simpler subproblems than directly minimizing the original function; see e.g., [12,14,33,34,37]. For example, if $h_k(\cdot) = 0$ for all k, and u_k takes the following form

$$u_k\left(x_k; w_k^{r+1}\right) = g\left(w_k^{r+1}\right) + \left\langle \nabla_k g\left(w_k^{r+1}\right), x_k - x_k^r \right\rangle + \frac{M_k}{2}\|x_k - x_k^r\|^2, \tag{2.10}$$

then we recover the well known BCGD method [2,18,22], in which x_k is updated by

$$x_k^{r+1} = \mathrm{proj}_{X_k}\left[x_k^r - \frac{1}{M_k}\nabla_k g\left(w_k^{r+1}\right)\right]. \tag{2.11}$$

When the nonsmooth components h_k's are present, the above choice of $u_k(\cdot;\cdot)$ in (2.10) leads to the so-called BCPG method [7,24,36], in which x_k is updated by

$$x_k^{r+1} = \mathrm{prox}_{h_k + I_{X_k}}^{M_k}\left[x_k^r - \frac{1}{M_k}\nabla_k g\left(w_k^{r+1}\right)\right]. \tag{2.12}$$

For other possible choices of the approximation function, we refer the readers to [20,24].

Secondly, the strong convexity requirement on $u_k(\cdot; x)$ in Assumption B(d) is quite mild, and is satisfied for the BCPG and BCGD algorithms; see the discussion in the previous remark. When u_k is chosen as the original function $g(x)$, this requirement says that $g(x)$ must be *block-wise strongly convex* (BSC). The BSC condition is in fact satisfied in many practical engineering problems. The following are two interesting examples.

Example 1 Consider the rate maximization problem in an uplink wireless communication network, where K users transmit to a single base station (BS) in the network. Suppose each user has n_t transmit antennas, and the BS has n_r receive antennas. Let $C_k \in \mathbb{R}^{n_t \times n_t}$ denote user k's transmit covariance matrix, P_k denote the maximum transmit power for user k, and $H_k \in \mathbb{R}^{n_r \times n_t}$ denote the channel matrix between user k and the BS. Then the uplink channel capacity optimization problem is given by the following convex program [8,35]

$$\min_{\{C_k\}_{k=1}^K} -\log\det\left|\sum_{k=1}^{K} H_k C_k H_k^T + I_{n_r}\right|, \text{ s.t. } C_k \succeq 0,\ \text{Tr}[C_k] \le P_k,\ \forall\, k, \tag{2.13}$$

where I_{n_r} is the $n_r \times n_r$ identity matrix. The celebrated iterative water-filling algorithm (IWFA) [35] for solving this problem is simply the BSUM algorithm with exact block minimization (i.e. the BCM algorithm) and G–S update rule. It is easy to verify that when $n_t \le n_r$ (i.e., the number of transmit antenna is smaller than that of the receive antenna), and when the channels are generated randomly, then with probability one $H_k^T H_k$ is of full rank, implying that the BSC condition is satisfied. We note that there has been no iteration complexity analysis of the IWFA algorithm for any type of block selection rules.

Example 2 Consider the following LASSO problem:

$$\min_x \|Ax - b\|^2 + \lambda\|x\|_1,$$

where $A \in \mathbb{R}^{M\times K}$, $b \in \mathbb{R}^M$, and $x = [x_1, \ldots, x_K]^T$, with $x_k \in \mathbb{R}$ for all k. That is, each block consists of a single scalar variable. In this case, as long as none of A's columns are zero (in which case we simply remove that column and the corresponding block variable), the problem satisfies the BSC property. Prior to our work, there is no iteration complexity analysis for applying BCD with deterministic block selection rules such as G–S and E-C for LASSO (with general data matrix A).

Note that the BSC property, or more generally the strong convexity assumption on the approximate function u_k, is reasonable as it ensures that each step of the BSUM algorithm is well-defined and has a unique solution. In the ensuing analysis of the BSUM algorithm, we assume that either the BSC property holds true, or u_k is a valid upper-bound. Later in Sects. 4–6, we will consider the case where the BSC assumption is absent.

3 Convergence analysis for BSUM

In this section, we show that under assumptions A and B, the BSUM algorithm with flexible update rules achieves global sublinear rate of convergence.

Let us define X^* as the optimal solution set, and let $x^* \in X^*$ be one of the optimal solutions. For the BSUM algorithm, define the optimality gap as

$$\Delta^r := f(x^r) - f(x^*). \tag{3.1}$$

Despite the generality of the BSUM algorithm, our analysis of BSUM only consists of three simple steps: (S1) estimate the amount of successive decrease of the optimality gaps; (S2) estimate the cost yet to be minimized (i.e., the cost-to-go) after each iteration; (S3) estimate the rate of convergence.

We first characterize the successive difference of the optimality gaps before and after one iteration of the BSUM algorithm, with different update rules.

Lemma 1 (Sufficient Descent) *Suppose Assumption A–B hold. Then*

1. *For BSUM with either G–S rule or the E-C rule, the following is true*

$$\Delta^r - \Delta^{r+1} \geq \sum_{k=1}^{K} \frac{\gamma_k}{2} \|x_k^r - x_k^{r+1}\|^2 \geq \gamma \|x^r - x^{r+1}\|^2, \quad \forall\, r \geq 1, \tag{3.2}$$

   where the constant $\gamma := \frac{1}{2}\min_k \gamma_k > 0$.
2. *For BSUM with G–So rule and MBI rule, the following is true*

$$\Delta^r - \Delta^{r+1} \geq \frac{c_1}{K}\gamma \|x^r - \hat{x}^{r+1}\|^2, \quad \forall\, r \geq 1, \tag{3.3}$$

   where the constant $\gamma := \frac{1}{2}\min_k \gamma_k > 0$*; For G–So rule,* $c_1 = q$*, and for MBI rule,* $c_1 = 1$.

Proof We first show part (1) of the proof. Suppose that $k \notin \mathcal{C}^{r+1}$, then we have the following trivial inequality

$$f\left(w_k^{r+1}\right) - f\left(w_{k+1}^{r+1}\right) \geq \frac{\gamma_k}{2}\|x_k^{r+1} - x_k^r\|^2, \tag{3.4}$$

as both sides of the inequality are zero.

Suppose $k \in \mathcal{C}^{r+1}$. Then using Assumption B, we have that

$$\begin{aligned} & f\left(w_k^{r+1}\right) - f\left(w_{k+1}^{r+1}\right) \\ & \geq u_k\left(x_k^r; w_k^{r+1}\right) + h_k\left(x_k^r\right) - \left(u_k\left(x_k^{r+1}; w_k^{r+1}\right) + h_k\left(x_k^{r+1}\right)\right) \\ & \geq \left\langle \nabla u_k\left(x_k^{r+1}; w_k^{r+1}\right), x_k^r - x_k^{r+1}\right\rangle + h_k\left(x_k^r\right) - h_k\left(x_k^{r+1}\right) + \frac{\gamma_k}{2}\|x_k^{r+1} - x_k^r\|^2 \end{aligned}$$

$$
\begin{aligned}
&\geq \left\langle \nabla u_k \left(x_k^{r+1}; w_k^{r+1}\right) + \zeta_k^{r+1}, x_k^r - x_k^{r+1} \right\rangle + \frac{\gamma_k}{2}\|x_k^{r+1} - x_k^r\|^2 \\
&\geq \frac{\gamma_k}{2}\|x_k^{r+1} - x_k^r\|^2 \qquad (3.5)
\end{aligned}
$$

where the first inequality is due to Assumption B(a)–B(b); the second inequality is due to Assumption B(d); in the third inequality we have defined $\zeta_k^{r+1} \in \partial h_k(x_k^{r+1})$ as any subgradient vector in the subdifferential $\partial h_k(x_k^{r+1})$; in the last inequality we have used the fact that x_k^{r+1} is the optimal solution for the strongly convex problem

$$
\arg\min_{x_k \in X_k} u_k \left(x_k; w_k^{r+1}\right) + h_k(x_k),
$$

and we have specialized ζ_k^{r+1} to the subgradient that satisfying the optimality condition of the problem above. Summing over k, we have

$$
f(x^r) - f(x^{r+1}) \geq \gamma \|x^r - x^{r+1}\|^2, \qquad (3.6)
$$

where $\gamma := \frac{1}{2}\min_k \gamma_k$.

We now show part (2) of the claim. Suppose $k \in \mathcal{C}^{r+1}$, then we have the following series of inequalities for the G–So rule

$$
\begin{aligned}
f(x^r) - f\left(x^{r+1}\right) &= f(x^r) - f\left(x_{-k}^r, \hat{x}_k^{r+1}\right) \\
&\geq u_k\left(x_k^r; x^r\right) + h_k\left(x_k^r\right) - u_k\left(\hat{x}_k^{r+1}; x^r\right) - h_k\left(\hat{x}_k^{r+1}\right) \\
&\geq \frac{1}{2}\gamma_k \|x_k^r - \hat{x}_k^{r+1}\|^2 \\
&\geq \frac{q \min_j \gamma_j}{2K} \sum_{j=1}^{K} \|x_j^r - \hat{x}_j^{r+1}\|^2 = \frac{q}{K}\gamma \|x^r - \hat{x}^{r+1}\|^2. \qquad (3.7)
\end{aligned}
$$

Similar steps lead to the result for the MBI rule. □

Next we show the second step of the proof, which estimates the gap yet to be minimized after each iteration. Let us define the following constants:

$$
\begin{aligned}
R &:= \max_{x \in X} \max_{x^* \in X^*} \left\{ \|x - x^*\| \ : \ f(x) \leq f(x^1) \right\}, \\
Q &:= \max_{x \in X} \left\{ \|\nabla g(x)\| : \ f(x) \leq f(x^1) \right\}. \qquad (3.8)
\end{aligned}
$$

When assuming that the level set $\{x : f(x) \leq f(x^1)\}$ is compact, then all the above constants are finite. Clearly we have

$$
\|x^r - x^*\| \leq R, \quad \|\nabla g(x^r)\| \leq Q, \ \forall\, r = 1, \ldots. \qquad (3.9)
$$

Occasionally we need to further make the assumption that the nonsmooth part $h(x)$ is Lipschitz continuous:

$$\|h(x) - h(y)\| \leq L_h \|x - y\|, \ \forall \, x, y \in X, \tag{3.10}$$

with some $L_h > 0$. Note that such assumption is satisfied by most of the popular nonsmooth regularizers such as the ℓ_1 norm, the ℓ_2 norm and so on. Also note that even with this assumption, our considered problem is still a *constrained* one, as the convex constraints $x_k \in X_k$ have not been moved to the objective as nonsmooth indicator functions.

Lemma 2 (Cost-to-go estimate) *Suppose Assumptions A–B hold. Then*

1. *For the BSUM with G–S update rule, we have*

$$(\varDelta^{r+1})^2 \leq R^2 K G_{\max}^2 \|x^{r+1} - x^r\|^2, \ \forall \, x^* \in X^*.$$

2. *For the BSUM with period-T E-C update rule, we have*

$$(\varDelta^{r+T})^2 \leq T R^2 K G_{\max}^2 \sum_{t=1}^{T} \|x^{r+t} - x^{r+t-1}\|^2, \ \forall \, x^* \in X^*.$$

3. *For the BSUM with G–So and MBI rules, further assume that $h(\cdot)$ is Lipschitz continuous* (*cf.* (3.10)). *Then we have*

$$\begin{aligned}(\varDelta^r)^2 &= \left(f(x^r) - f(x^*)\right)^2 \\ &\leq 2\left((Q + L_h)^2 + L_{\max}^2 K R^2\right) \|\hat{x}^{r+1} - x^r\|^2, \ \forall \, x^* \in X^*.\end{aligned}$$

Proof We first show part (1). We have the following sequence of inequalities

$$\begin{aligned} f(x^{r+1}) - f(x^*) &= g(x^{r+1}) - g(x^*) + h(x^{r+1}) - h(x^*) \\ &\leq \left\langle \nabla g(x^{r+1}), x^{r+1} - x^* \right\rangle + h(x^{r+1}) - h(x^*) \\ &= \sum_{k=1}^{K} \left\langle \nabla_k g(x^{r+1}) - \nabla u_k \left(x_k^{r+1}; w_k^{r+1}\right), x_k^{r+1} - x_k^* \right\rangle \\ &\quad + \sum_{k=1}^{K} \left\langle \nabla u_k \left(x_k^{r+1}; w_k^{r+1}\right), x_k^{r+1} - x_k^* \right\rangle + h(x^{r+1}) - h(x^*). \end{aligned} \tag{3.11}$$

Notice that x_k^{r+1} is the optimal solution for problem: $\operatorname{argmin}_{x_k \in X_k} u_k(x_k; w_k^{r+1}) + h_k(x_k)$. It follows from the optimality condition of this problem that there exists some $\zeta_k^{r+1} \in \partial(h_k(x_k^{r+1}))$ such that

$$
\begin{aligned}
0 &\geq \left\langle \nabla u_k\left(x_k^{r+1}; w_k^{r+1}\right) + \zeta_k^{r+1}, x_k^{r+1} - x_k^*\right\rangle \\
&\geq \left\langle \nabla u_k\left(x_k^{r+1}; w_k^{r+1}\right), x_k^{r+1} - x_k^*\right\rangle + h_k\left(x_k^{r+1}\right) - h_k\left(x_k^*\right), \qquad (3.12)
\end{aligned}
$$

where in the last inequality we have used the definition of subgradient

$$
h_k\left(x_k^{r+1}\right) - h_k\left(x_k^*\right) \leq \left\langle \zeta_k^{r+1}, x_k^{r+1} - x_k^*\right\rangle, \; \forall\, x_k^{r+1}, x_k^* \in X_k. \qquad (3.13)
$$

Combining (3.11) and (3.12), we obtain

$$
\begin{aligned}
\left(f(x^{r+1}) - f(x^*)\right)^2 &\overset{\text{(i)}}{\leq} \left(\sum_{k=1}^K \|\nabla_k g(x^{r+1}) - \nabla u_k\left(x_k^{r+1}; w_k^{r+1}\right)\|\|x_k^{r+1} - x_k^*\|\right)^2 \\
&\overset{\text{(ii)}}{\leq} \left(\sum_{k=1}^K G_k\|x^{r+1} - w_k^{r+1}\|\|x_k^{r+1} - x_k^*\|\right)^2 \\
&\leq R^2 K G_{\max}^2 \|x^{r+1} - x^r\|^2,
\end{aligned}
$$

where in (i) we have used the Cauchy–Schwarz inequality and the Lipschitz continuity of $u_k(\cdot;\cdot)$ in (2.8); in (ii) we have used the Lipschitz continuity of $\nabla g(\cdot)$ in (2.7), and that $\nabla_k g(x^{r+1}) = \nabla u_k(x_k^{r+1}; x^{r+1})$ (cf. Assumption B(c)).

Next we show part (2) of the claim. Let us define an index set $\{r_k\}$ as:

$$
r_k := \arg\max_t \{x_k^t \neq x_k^{r+T}\} + 1, \; k = 1, \ldots, K. \qquad (3.14)
$$

That is, r_k is the latest iteration index (up until $r+T$) in which the kth variable has been updated. From this definition we have $x_k^{r_k} = x_k^{r+T}$, for all k. We have the following sequence of inequalities

$$
\begin{aligned}
f(x^{r+T}) - f(x^*) &= g(x^{r+T}) - g(x^*) + \sum_{k=1}^K \left(h_k(x_k^{r_k}) - h_k(x_k^*)\right) \\
&\leq \left\langle \nabla g(x^{r+T}), x^{r+T} - x^*\right\rangle + \sum_{k=1}^K \left(h_k(x_k^{r_k}) - h_k(x_k^*)\right) \\
&\overset{\text{(i)}}{=} \sum_{k=1}^K \Bigg(\left\langle \nabla_k g(x^{r+T}) - \nabla u_k\left(x_k^{r_k}; w_k^{r_k}\right), x_k^{r+T} - x_k^*\right\rangle \\
&\qquad + \left\langle \nabla u_k\left(x_k^{r_k}; w_k^{r_k}\right), x_k^{r_k} - x_k^*\right\rangle\Bigg) \\
&\qquad + \sum_{k=1}^K \left(h_k\left(x_k^{r_k}\right) - h_k\left(x_k^*\right)\right) \\
&\overset{\text{(ii)}}{\leq} \sum_{k=1}^K \left\langle \nabla_k g(x^{r+T}) - \nabla u_k\left(x_k^{r_k}; w_k^{r_k}\right), x_k^{r+T} - x_k^*\right\rangle,
\end{aligned}
$$

where in (i) we have used the fact that $x_k^{r+T} = x_k^{r_k}$, for all k; in (ii) we have used the optimality of $x_k^{r_k}$. Taking the square on both sides, we obtain

$$
\begin{aligned}
(f(x^{r+T}) - f(x^*))^2 &\le \left(\sum_{k=1}^{K} \|\nabla_k g(x^{r+T}) - \nabla u_k\left(x_k^{r_k}; w_k^{r_k}\right)\| \|x_k^{r+T} - x_k^*\|\right)^2 \\
&\le \left(\sum_{k=1}^{K} G_k \|x^{r+T} - w_k^{r_k}\| \|x_k^{r+T} - x_k^*\|\right)^2 \\
&\le \left(\sum_{k=1}^{K} G_k \left(\|x^{r+T} - x^{r_k}\| + \|x^{r_k} - w_k^{r_k}\|\right) \|x_k^{r+T} - x_k^*\|\right)^2 \\
&\le T K G_{\max}^2 R^2 \sum_{t=1}^{T} \|x^{r+t-1} - x^{r+t}\|^2.
\end{aligned}
$$

Finally we show part (3). We have the following sequence of inequalities

$$
\begin{aligned}
f(x^r) - f(x^*) &= g(x^r) - g(x^*) + h(x^r) - h(x^*) \\
&\overset{\text{(i)}}{\le} \left\langle \nabla g(x^r), x^r - x^* \right\rangle + L_h \|x^r - \hat{x}^{r+1}\| + h(\hat{x}^{r+1}) - h(x^*) \\
&= \left\langle \nabla g(x^r), x^r - \hat{x}^{r+1} \right\rangle + \left\langle \nabla g(x^r), \hat{x}^{r+1} - x^* \right\rangle \\
&\quad + L_h \|x^r - \hat{x}^{r+1}\| + h(\hat{x}^{r+1}) - h(x^*) \\
&\le (L_h + Q)\|x^r - \hat{x}^{r+1}\| + \sum_{k=1}^{K} \left\langle \nabla_k g(x^r) - \nabla u_k\left(\hat{x}_k^{r+1}; x^r\right), \hat{x}_k^{r+1} - x_k^* \right\rangle \\
&\quad + \sum_{k=1}^{K} \left\langle \nabla u_k\left(\hat{x}_k^{r+1}; x^r\right), \hat{x}_k^{r+1} - x_k^* \right\rangle + h(\hat{x}^{r+1}) - h(x^*), \qquad (3.15)
\end{aligned}
$$

where step (i) follows from the Lipschitz continuity assumption (3.10) and the convexity of $g(\cdot)$. Similar to the proof of (3.12) in part (1), we can show that

$$
\sum_{k=1}^{K} \left\langle \nabla u_k\left(\hat{x}_k^{r+1}; x^r\right), \hat{x}_k^{r+1} - x_k^* \right\rangle + h\left(\hat{x}^{r+1}\right) - h\left(x^*\right) \le 0. \qquad (3.16)
$$

Moreover, it follows from Assumption B(c) and B(e) that

$$
\begin{aligned}
&\left(\sum_{k=1}^{K} \left\langle \nabla_k g(x^r) - \nabla u_k\left(\hat{x}_k^{r+1}; x^r\right), x_k^{r+1} - x_k^* \right\rangle\right)^2 \\
&= \left(\sum_{k=1}^{K} \left\langle \nabla u_k\left(x_k^r; x^r\right) - \nabla u_k\left(\hat{x}_k^{r+1}; x^r\right), x_k^{r+1} - x_k^* \right\rangle\right)^2 \\
&\le K \sum_{k=1}^{K} L_k^2 \|x_k^r - \hat{x}_k^{r+1}\|^2 \|x_k^{r+1} - x_k^*\|^2
\end{aligned}
$$

$$\leq KL_{\max}^2\|x^r - \hat{x}^{r+1}\|^2 R^2. \tag{3.17}$$

Putting the above three inequalities together, we have

$$(f(x^r) - f(x^*))^2 \leq 2\left((Q + L_h)^2 + KL_{\max}^2 R^2\right)\|x^r - \hat{x}^{r+1}\|^2. \tag{3.18}$$

This completes the proof. □

We are now ready to prove the $\mathcal{O}(1/r)$ iteration complexity for the BSUM algorithm when applied to problem (1.1). Our results below are more general than the recent analysis on the iteration complexity for BCD-type algorithms. The generality of our results can be seen from several fronts: (1) The family of algorithms we analyze is broad; it includes the classic BCM (with the additional BSC condition), the BCGD method, the BCPG methods as well as their variants based on different coordinate selection rules as special cases, while the existing works only focus on one particular algorithm; (2) When the coordinates are updated in a G–S fashion, our result covers the general multi-block nonsmooth case, where $h_k(x)$ can take any proper closed convex nonsmooth function, while existing works only cover some special cases [1,2,26]; (3) When the coordinates are updated using other update rules such as G–So, MBI, E-C fashion, our convergence results appear to be new.

Theorem 1 *Suppose Assumption A(a), B hold true. We have the following.*

1. *Let $\{x^r\}$ be the sequence generated by the BSUM algorithm with G–S rule. Then we have*

$$\Delta^r = f(x^r) - f^* \leq \frac{c_1}{\sigma_1}\frac{1}{r}, \ \forall\, r \geq 1, \tag{3.19}$$

where the constants are given below

$$c_1 = \max\{4\sigma_1 - 2, f(x^1) - f^*, 2\}, \quad \sigma_1 = \frac{\gamma}{KG_{\max}^2 R^2}. \tag{3.20}$$

2. *Let $\{x^r\}$ be the sequence generated by the BSUM algorithm with E-C rule. Then we have*

$$\Delta^r = f(x^r) - f^* \leq \frac{c_2}{\sigma_2}\frac{1}{r-T}, \ \forall\, r > T, \tag{3.21}$$

where the constants are given below

$$c_2 = \max\{4\sigma_2 - 2, f(x^1) - f^*, 2\}, \quad \sigma_2 = \frac{\gamma}{KTR^2 G_{\max}^2}. \tag{3.22}$$

3. *Suppose the Lipschitz continuity assumption (3.10) holds true. Let $\{\mathbf{x}^r\}$ be the sequence generated by BSUM with G–So and MBI rule. Then we have*

$$\Delta^r = f(x^r) - f^* \leq \frac{1}{\sigma_3 r}, \tag{3.23}$$

where

$$\sigma_3 = \begin{cases} \frac{\gamma q}{2K\left((Q+L_h)^2+L_{\max}^2 K R^2\right)}, & \text{(G–So rule)} \\ \frac{\gamma}{2K\left((Q+L_h)^2+L_{\max}^2 K R^2\right)}, & \text{(MBI rule)} \end{cases}. \tag{3.24}$$

Proof We first show part (1) of the claim by mathematical induction on r. From Lemmas 1 and 2, we have that for the G–S rule, we have

$$\Delta^r - \Delta^{r+1} \geq \frac{\gamma}{K G_{\max}^2 R^2}(\Delta^{r+1})^2 := \sigma_1(\Delta^{r+1})^2, \ \forall\, r \geq 1, \tag{3.25}$$

or equivalently

$$\sigma_1(\Delta^{r+1})^2 + \Delta^{r+1} \leq \Delta^r, \ \forall\, r \geq 1. \tag{3.26}$$

By definition, we have $\Delta^1 = f(x^1) - f^*$. We first argue that

$$\Delta^2 \leq \frac{c_1}{2\sigma_1}, \ \text{with } c_1 := \max\{4\sigma_1 - 2, f(\mathbf{x}^1) - f^*, 2\}. \tag{3.27}$$

From (3.26) and the fact that $\Delta^1 \leq c_1$, we have

$$\Delta^2 \leq \frac{-1+\sqrt{1+4\sigma_1 c_1}}{2\sigma_1} = \frac{2c_1}{1+\sqrt{1+4\sigma_1 c_1}} \leq \frac{2c_1}{1+|4\sigma_1 - 1|},$$

where in the last inequality we have used the fact that $c_1 \geq 4\sigma_1 - 2$. Suppose $4\sigma_1 - 1 \geq 0$, then we immediately have $\Delta^2 \leq \frac{c_1}{2\sigma_1}$. Suppose $4\sigma_1 - 1 < 0$, then

$$\Delta^2 \leq \frac{2c_1}{2-4\sigma_1} \leq \frac{2c_1}{8\sigma_1 - 4\sigma_1} = \frac{c_1}{2\sigma_1}. \tag{3.28}$$

Next we argue that if $\Delta^r \leq \frac{c_1}{r\sigma_1}$, then we must have

$$\Delta^{r+1} \leq \frac{c_1}{(r+1)\sigma_1}. \tag{3.29}$$

Using the condition (3.26) and the inductive hypothesis $\Delta^r \leq \frac{c_1}{r\sigma_1}$, we have

$$\begin{aligned} \Delta^{r+1} &\leq \frac{-1+\sqrt{1+\frac{4c_1}{r}}}{2\sigma_1} = \frac{2c_1}{r\sigma_1\left(1+\sqrt{1+\frac{4c_1}{r}}\right)} \\ &\leq \frac{2c_1}{\sigma_1\left(r+\sqrt{r^2+4r+4}\right)} = \frac{c_1}{\sigma_1(r+1)}, \end{aligned} \tag{3.30}$$

where the last inequality is due to the fact that $c_1 \geq 2$, and $r \geq 2$. Consequently, we have shown that for all $r \geq 1$

$$\Delta^r = f(x^r) - f^* \leq \frac{c_1}{\sigma_1}\frac{1}{r}. \tag{3.31}$$

For the E-C rule, first note that from Lemma 1, we have

$$\Delta^r - \Delta^{r+T} \geq \frac{\gamma}{TKR^2G_{\max}^2}(\Delta^{r+T})^2 := \sigma_2(\Delta^{r+T})^2, \ \forall\, r \geq 1. \tag{3.32}$$

Then using the similar argument as for the G–S rule, we can obtain the desired result.

Next we show part (3) of the claim. For the G–So rule, we have from Lemma 2, the second part of Lemma 1, that for all $r \geq 1$

$$\begin{aligned}\Delta^r - \Delta^{r+1} &\geq \frac{q}{K}\gamma\|\hat{x}^{r+1} - x^r\|^2 \geq \frac{\gamma q}{2K\left((Q+L_h)^2 + L_{\max}^2 KR^2\right)}(\Delta^r)^2\\ &:= \sigma_3(\Delta^r)^2. \end{aligned}\tag{3.33}$$

Similar relation can be shown for the MBI rule as well. The rest of the proof follows standard argument, see for example [22, Theorem 1]. □

Below we provide further remarks on some special cases of BSUM.

- One popular choice of the upper bound function $u_k(\cdot,\cdot)$ is [2,12,22,33,37]

$$u_k(z_k; x) := g(x) + \langle \nabla_k g(x), z_k - x_k\rangle + \frac{L_k}{2}\|z_k - x_k\|^2, \tag{3.34}$$

  where the constant $L_k \geq \rho_{\max}(\nabla^2 g(x))$, is often chosen to be largest eigenvalue of the Hessian of $g(x)$. In this case, evidently we have $\gamma_k = L_k = M_k \leq M$, for all k, and $G_{\max} \leq M$. We can also verify that $G_k \leq 2M$ for all k. Using this choice of $u_k(\cdot;\cdot)$ and L_k, the first result in Theorem 1 reduces to

$$\Delta^r \leq 2\frac{c_1KM^2R^2}{M_{\min}}\frac{1}{r}, \tag{3.35}$$

  where $M_{\min} := \min_k M_k$. Let us compare the order given in (3.35) with the one stated in [2, Theorem 6.1], which is the best known complexity bound for the G–S BCD algorithm for *smooth* problems (i.e., when h_k is not present). The bound derived in [2] for smooth constrained problem (resp. smooth unconstrained problem) is in the order of $\frac{KM^2R^2}{M_{\min}}\frac{1}{r}$ (resp. $\frac{M_{\max}KM^2R^2}{M_{\min}^2}\frac{1}{r}$). These orders are approximately the same as (3.35). However, our proof covers the general nonsmooth cases, and is simpler. Similarly, when $u_k(\cdot;\cdot)$ takes the form (3.34), the bounds for the BSUM with the E-C/G–So/MBI rules shown in Theorem 1 can also be simplified.

- The results derived in Theorem 1 is equally applicable to the BCM scheme (1.2) with various block selection rules discussed above. In particular, we can specialize the upper-bound function u_k to be the original smooth function g. As long as $g(x_1, \ldots, x_K)$ satisfies the BSC property, Theorem 1 carries over. As mentioned in Sect. 2.2, the BSC property is fairly mild and is satisfied in many engineering applications. Nevertheless, we will further relax the BSC condition in the subsequent sections.

4 The BSUM for single block problems

4.1 The SUM algorithm

In this section, we consider the following single-block problem with $K = 1$:

$$\min \quad f(x) := g(x) + h(x), \quad \text{s.t.} \quad x \in X. \tag{4.1}$$

In this case the BSUM algorithm reduces to to the so-called successive upper-bound minimization (SUM) algorithm [24], listed in the following table.

The Successive Upper-Bound Minimization (SUM) Algorithm
At each iteration $r + 1$, do:

$$x^{r+1} \in \arg\min_{x \in X} \; u\left(x; x^r\right) + h(x). \tag{4.2}$$

Let us make the following assumptions on the function $u(v; x)$.

Assumption C (a) $u(x; x) = g(x), \quad \forall\, x \in X.$
(b) $u(v; x) \geq g(v), \quad \forall\, v \in X, \; \forall\, x \in X.$
(c) $\nabla u(x; x) = \nabla g(x), \quad \forall\, x \in X.$
(d) For any given x, $u(v; x)$ is convex in v and satisfies the following

$$\|\nabla u(v; x) - \nabla u(\hat{v}; x)\| \leq L\|v - \hat{v}\|, \; \forall\, \hat{v}, \; v \in X, \forall\, x \in X, \tag{4.3}$$

where $L > 0$ is some constant.

Compared to Assumptions B and C does not require $u(v; x)$ to be strongly convex in v, nor $\nabla u(v; x)$ to be Lipschitz continuous over x.

Proposition 1 *Suppose $g(x)$ is convex, and $u(v; x)$ satisfies Assumption C. Then we must have*

$$\|\nabla g(v) - \nabla g(x)\| \leq L\|v - x\|, \quad \forall\, x, v \in X. \tag{4.4}$$

That is, ∇g is Lipschitz continuous with the coefficient no larger than L.

Proof Utilizing Assumption C, we must have

$$\begin{aligned} g(v) - g(x) &\le u(v; x) - u(x; x) \\ &\le \langle \nabla u(x; x), v - x\rangle + \frac{L}{2}\|x - v\|^2 \\ &= \langle \nabla g(x), v - x\rangle + \frac{L}{2}\|x - v\|^2, \ \forall\, x, v \in X. \end{aligned}$$

Further, using the convexity of g we have

$$g(v) - g(x) \ge \langle \nabla g(x), v - x\rangle, \ \forall\, x, v \in X.$$

Combining these two inequalities we obtain

$$0 \le g(v) - g(x) - \langle \nabla g(x), v - x\rangle \le \frac{L}{2}\|x - v\|^2, \ \forall\, x, v \in X. \tag{4.5}$$

Similar to [21, Theorem 2.1.5], we construct $\phi(x) := g(x) - \langle \nabla g(v), x\rangle$. Clearly $v \in \arg\min \phi(x)$. We have

$$\phi(v) \le \phi\left(x - \frac{1}{L}\nabla\phi(x)\right) \le \phi(x) - \frac{1}{2L}\|\nabla\phi(x)\|^2, \tag{4.6}$$

where the first inequality is due to the optimality of v and the second inequality uses (4.5). Plugging in the definition of $\phi(x)$ and $\phi(v)$ we have

$$g(v) - \langle \nabla g(v), v\rangle \le g(x) - \langle \nabla g(v), x\rangle - \frac{1}{2L}\|\nabla g(v) - \nabla g(x)\|^2.$$

Since the above inequality is true for any $x, v \in X$, we can interchange x and v and obtain

$$g(x) - \langle \nabla g(x), x\rangle \le g(v) - \langle \nabla g(x), v\rangle - \frac{1}{2L}\|\nabla g(v) - \nabla g(x)\|^2.$$

Adding these two inequalities we obtain

$$\frac{1}{L}\|\nabla g(x) - \nabla g(v)\|^2 \le \langle \nabla g(x) - \nabla g(v), x - v\rangle \le \|\nabla g(x) - \nabla g(v)\|\|x - v\|.$$

Cancelling $\|\nabla g(x) - \nabla g(v)\|$ we arrive at the desired results. □

We remark that this result is only true when $g(\cdot)$ is a convex function.

Our main result is that the SUM algorithm converges sublinearly under Assumption C, *without* the strong convexity of the upper-bound function $u(v; x)$ in v. The proof of this claim is an extension of Theorem 1, therefore we will only provide its key steps. Observe that the following is true

$$\begin{aligned} f(x^r) - f(x^{r+1}) &\overset{\text{(i)}}{\geq} f(x^r) - \left(u(x^{r+1}; x^r) + h(x^{r+1})\right) \\ &\overset{\text{(ii)}}{\geq} f(x^r) - \left(u(\widetilde{x}^{r+1}; x^r) + h(\widetilde{x}^{r+1})\right) \overset{\text{(iii)}}{\geq} \frac{\gamma}{2}\|x^r - \widetilde{x}^{r+1}\|^2, \end{aligned} \tag{4.7}$$

where $\widetilde{x}^{r+1}$ is obtained by solving the following problem for any $\gamma > 0$

$$\widetilde{x}^{r+1} = \arg\min_{x \in X} u(x; x^r) + h(x) + \frac{\gamma}{2}\|x - x^r\|^2. \tag{4.8}$$

In (4.7), (i) is true because $u(x; y)$ is an upper-bound function for $g(x)$ satisfying Assumption C(b); (ii) is true because x^{r+1} is a minimizer of problem (4.2); (iii) is true due to the fact that $\widetilde{x}^{r+1}$ is the optimal solution of (4.8) while x^r is a feasible solution.

Then we bound $f(x^{r+1})$ using $f(\widetilde{x}^{r+1})$. We have

$$\begin{aligned} f(x^{r+1}) &\leq u(x^{r+1}; x^r) + h(x^{r+1}) \overset{\text{(i)}}{\leq} u(\widetilde{x}^{r+1}; x^r) + h(\widetilde{x}^{r+1}) \\ &\overset{\text{(ii)}}{\leq} u(x^r; x^r) + \left\langle \nabla u(x^r; x^r), \widetilde{x}^{r+1} - x^r \right\rangle + \frac{L}{2}\|\widetilde{x}^{r+1} - x^r\|^2 + h(\widetilde{x}^{r+1}) \\ &\overset{\text{(iii)}}{\leq} g(\widetilde{x}^{r+1}) + \left\langle \nabla u(x^r; x^r), \widetilde{x}^{r+1} - x^r \right\rangle + \left\langle \nabla g(\widetilde{x}^{r+1}), x^r - \widetilde{x}^{r+1} \right\rangle \\ &\quad + L\|\widetilde{x}^{r+1} - x^r\|^2 + h(\widetilde{x}^{r+1}) \\ &\overset{\text{(iv)}}{=} g(\widetilde{x}^{r+1}) + \left\langle \nabla g(\widetilde{x}^{r+1}) - \nabla g(x^r), x^r - \widetilde{x}^{r+1} \right\rangle \\ &\quad + L\|\widetilde{x}^{r+1} - x^r\|^2 + h(\widetilde{x}^{r+1}) \\ &\overset{\text{(v)}}{\leq} f(\widetilde{x}^{r+1}) + L\|\widetilde{x}^{r+1} - x^r\|^2, \end{aligned}$$

where (i) is due to the optimality of x^{r+1} for problem (4.2); (ii) uses the gradient Lipschitz continuity of $u(\cdot; x^r)$; (iii) uses the fact that $u(x^r; x^r) = g(x^r)$, the gradient Lipschitz continuity of $g(\cdot)$ derived in Proposition 1; (iv) uses the fact that $\nabla u(x^r; x^r) = \nabla g(x^r)$ (cf. Assumption C(c)); (v) uses the convexity of $g(\cdot)$.

Utilizing this bound, we derive the estimate of the cost-to-go

$$\begin{aligned} f(x^{r+1}) - f(x^*) &\leq f(\widetilde{x}^{r+1}) - f(x^*) + L\|\widetilde{x}^{r+1} - x^r\|^2 \\ &\leq \left\langle \nabla g(\widetilde{x}^{r+1}), \widetilde{x}^{r+1} - x^* \right\rangle + h(\widetilde{x}^{r+1}) - h(x^*) + L\|\widetilde{x}^{r+1} - x^r\|^2 \\ &= \left\langle \nabla g(\widetilde{x}^{r+1}) - \nabla g(x^r), \widetilde{x}^{r+1} - x^* \right\rangle + L\|\widetilde{x}^{r+1} - x^r\|^2 \\ &\quad + \left\langle \nabla g(x^r) - \nabla\left(u(\widetilde{x}^{r+1}; x^r) + \frac{\gamma}{2}\|\widetilde{x}^{r+1} - x^r\|^2\right), \widetilde{x}^{r+1} - x^* \right\rangle \\ &\quad + h(\widetilde{x}^{r+1}) - h(x^*) \end{aligned}$$

$$
\begin{aligned}
&+\left\langle \nabla\left(u(\widetilde{x}^{r+1};x^r)+\frac{\gamma}{2}\|\widetilde{x}^{r+1}-x^r\|^2\right),\widetilde{x}^{r+1}-x^*\right\rangle\\
&\overset{\text{(i)}}{\leq}\left\langle \nabla g(\widetilde{x}^{r+1})-\nabla g(x^r),\widetilde{x}^{r+1}-x^*\right\rangle+L\|\widetilde{x}^{r+1}-x^r\|^2\\
&\quad+\left\langle \nabla u(x^r;x^r)-\nabla u(\widetilde{x}^{r+1};x^r),\widetilde{x}^{r+1}-x^*\right\rangle\\
&\quad-\gamma\left\langle \widetilde{x}^{r+1}-x^r,\widetilde{x}^{r+1}-x^*\right\rangle\\
&\overset{\text{(ii)}}{\leq}(2L+\gamma)\|\widetilde{x}^{r+1}-x^r\|R+L\|\widetilde{x}^{r+1}-x^r\|\|\widetilde{x}^{r+1}-x^*+x^*-x^r\|\\
&\leq(4L+\gamma)\|\widetilde{x}^{r+1}-x^r\|R.
\end{aligned}
$$

Here (i) is due to the optimality of $\widetilde{x}^{r+1}$ to the problem (4.8); in (ii) we have used (4.4), Cauchy–Schwartz inequality and the definition of R (it is easy to show that $f(\widetilde{x}^{r+1})\leq f(x^r)\leq f(x^0)$, hence $\|\widetilde{x}^{r+1}-x^*\|\leq R$ for all r).

Combining the above two inequalities, we obtain

$$
\Delta^r-\Delta^{r+1}\geq\frac{\gamma}{2R^2(4L+\gamma)^2}(\Delta^{r+1})^2,\quad\forall\gamma>0. \tag{4.9}
$$

Maximizing over γ (with $\gamma=4L$), we have

$$
\Delta^r-\Delta^{r+1}\geq\frac{1}{32R^2L}(\Delta^{r+1})^2:=\sigma_4(\Delta^{r+1})^2. \tag{4.10}
$$

Using the same derivation as in Theorem 1, we obtain

$$
\Delta^{r+1}\leq\frac{c_4}{\sigma_4}\frac{1}{r},\quad\text{with}\quad\sigma_4=\frac{1}{32R^2L},\quad c_4:=\max\{4\sigma_4-2,f(x^1)-f^*,2\}. \tag{4.11}
$$

4.2 Application

To see the importance of the above result, consider the well-known method of iterative reweighted least squares (IRLS) [1,9]. The IRLS is a popular algorithm used for solving problems such as sparse recovery and Fermat-Weber problem; see [1, Section 4] for applications. Consider the following problem

$$
\min_x\quad h(x)+\sum_{j=1}^{\ell}\|A_jx+b_j\|_2,\quad\text{s.t.}\quad x\in X, \tag{4.12}
$$

where $A_j\in\mathbb{R}^{k_i\times m}$, $b_j\in\mathbb{R}^{k_i}$, $X\subseteq\mathbb{R}^m$, and $h(x)$ is some convex function not necessarily smooth. Let us introduce a constant $\eta>0$ and consider a *smooth approximation* of problem (4.12):

$$\min_{x} \quad h(x) + g(x) := h(x) + \sum_{j=1}^{\ell} \sqrt{\|A_j x + b_j\|_2^2 + \eta^2}, \quad \text{s.t.} \quad x \in X. \tag{4.13}$$

The IRLS algorithm generates the following iterates

$$x^{r+1} = \arg\min_{x \in X} \left\{ h(x) + \frac{1}{2} \sum_{j=1}^{\ell} \frac{\|A_j x + b_j\|^2 + \eta^2}{\sqrt{\|A_j x^r + b_j\|^2 + \eta^2}} \right\}. \tag{4.14}$$

It is known that the IRLS is equivalent to a BCM method applied to the following two-block problem (i.e., the first block is x and the second is $\{z_j\}_{j=1}^{\ell}$)

$$\begin{aligned} \min \quad & h(x) + \frac{1}{2} \sum_{j=1}^{\ell} \left(\frac{\|A_j x + b_j\|^2 + \eta^2}{z_j} + z_j \right) \\ \text{s.t.} \quad & x \in X, \quad z_j \in [\eta/2, \infty), \ \forall\, j. \end{aligned} \tag{4.15}$$

Utilizing such two-block BCM interpretation, the author of [1] shows that the IRLS converges sublinearly when $h(x)$ has Lipschitz continuous gradient; see [1, Theorem 4.1].

Differently from [1], here we take a new perspective. We argue that the IRLS is in fact the SUM algorithm in disguise, therefore our simple iteration complexity analysis given in Sect. 4.1 for SUM can be directly applied.

Let us consider the following function:

$$u(x; x^r) = \frac{1}{2} \sum_{j=1}^{\ell} \left(\frac{\|A_j x + b_j\|^2 + \eta^2}{\sqrt{\|A_j x^r + b_j\|^2 + \eta^2}} + \sqrt{\|A_j x^r + b_j\|^2 + \eta^2} \right). \tag{4.16}$$

It is clear that $g(x^r) = u(x^r; x^r)$, so Assumption C(a) is satisfied. To verify Assumption C(b), we apply the arithmetic-geometric inequality, and have

$$\begin{aligned} u(x; x^r) &= \frac{1}{2} \sum_{j=1}^{\ell} \left(\frac{\|A_j x + b_j\|^2 + \eta^2}{\sqrt{\|A_j x^r + b_j\|^2 + \eta^2}} + \sqrt{\|A_j x^r + b_j\|^2 + \eta^2} \right) \\ &\geq \sum_{j=1}^{\ell} \sqrt{\|A_j x + b_j\|^2 + \eta^2} = g(x), \ \forall\, x \in X. \end{aligned}$$

Assumptions C(c), (d) are also easy to verify. Note that the matrices A_j's do not necessarily have full column rank, so $u(x; x^r)$ may not be strongly convex over $x \in X$. Nevertheless, $u(x; x^r)$ defined in (4.16) is indeed an upper bound function for the smooth function $g(x)$, and we have shown that it satisfies Assumptions C. It follows that the iteration (4.14) corresponds to a single-block BSUM algorithm. Our analysis leading to (4.11) suggests that this algorithm converges in a sublinear rate, even when $h(x)$ is a nonsmooth function. To be more specific, for this problem we have

$$L = \frac{1}{\eta}\rho_{\max}\left(\sum_{j=1}^{\ell} A_j^T A_j\right).$$

Therefore the rate can be expressed as

$$\Delta^{r+1} \le \max\{4\sigma_4 - 2, f(x^1) - f(x^*), 2\}\frac{32R^2\rho_{\max}\left(\sum_{j=1}^{\ell} A_j^T A_j\right)}{\eta r}. \tag{4.17}$$

Note that compared with the result derived in [1, Theorem 4.1] which is based on transforming the IRLS algorithm to the two-block BCM problem (4.15), our analysis is based on the key insight of the equivalence between IRLS and the single block BSUM, and it is significantly simpler. Further we do not require $h(x)$ to be smooth, while the result in [1, Theorem 4.1] additionally requires that the gradient of $h(x)$ is Lipschitz continuous.[2]

5 The BSUM for two block problems

5.1 Iteration complexity for 2-block BSUM

In this section, we consider the following two-block problem ($K = 2$), which is a special case of problem (1.1):

$$\begin{aligned} \min \quad & f(x_1, x_2) := g(x_1, x_2) + h_1(x_1) + h_2(x_2) \\ \text{s.t.} \quad & x_1 \in X_1, \ x_2 \in X_2. \end{aligned} \tag{5.1}$$

This problem has many applications, such as the special case of Example 1 with two users, the two-block formulation of the IRLS algorithm (4.15) or the example presented in [1, Section 5]. Throughout this section, we assume that Assumption A(a) is true. We make the following additional assumptions about problem (5.1).

Assumption D (a) The problem $\min_{x_2 \in X_2} f(x_1, x_2)$ has a unique solution.
(b) The gradient of $g(x_1, x_2)$ with respect to x_1 is Lipschitz continuous, i.e.,

$$\|\nabla_1 g(x_1, x_2) - \nabla_1 g(v_1, x_2)\| \le M_1 \|x_1 - v_1\|.$$

Note that here we do not require that the gradient of $g(\cdot)$ with respect to the second block to be Lipschitz continuous.

[2] It appears that the proof in [1, Theorem 4.1] can be modified to allow nonsmooth h, just that it is not explicitly mentioned in the paper. But as it stands, the bound in [1, Theorem 4.1] is explicitly dependent on the Lipschitz constant of the gradient of h, while the bound we derived here in (4.17) is not.

We show that for this problem BSUM with G–S update rule is able to achieve sublinear rate without the BSC condition or the Lipschitz continuity of $\nabla_2 g(x_1, x_2)$. In the table given below we list the two-block BSUM algorithm with G–S update rule.

The G–S 2-block BSUM for problem (5.1)

At each iteration $r+1$, update the variable blocks by:

$$\begin{aligned} x_2^{r+1} &= \arg\min_{x_2 \in X_2} u_2\left(x_2; x_1^r, x_2^r\right) + h_2(x_2) \\ x_1^{r+1} &\in \arg\min_{x_1 \in X_1} u_1\left(x_1; x_1^r, x_2^{r+1}\right) + h_1(x_1). \end{aligned} \tag{5.2}$$

Unfortunately for the problem of interest here the rate analysis provided in Theorem 1 is no longer applicable because $\nabla_2 g(x_1, x_2)$ may not be Lipschitz continuous, and both subproblems may not be strongly convex. To analyze the convergence rate, let us consider the following special choices of the upper bound where $u_1(x_1; x)$ satisfies Assumption B(a)–(c) and the Lipschtiz continuous gradient condition (2.8), restated below for convenience

$$\|\nabla u_1(x_1; x) - \nabla u_1(v_1; x)\| \le L_1 \|x_1 - v_1\|, \ \forall\, x_1, v_1 \in X_1, \ \forall\, x \in X. \tag{5.3}$$

By utilizing the argument in Proposition 1, we can show that $L_1 \ge M_1$, and therefore the following is true as well

$$\|\nabla_1 g(x_1, x_2) - \nabla_1 g(v_1, x_2)\| \le L_1 \|x_1 - v_1\|.$$

Further we do not use any upper bound for the second block, i.e., we let

$$u_2(v_2; x) = g(v_2, x_1), \ \forall\, x_1 \in X_1, \ v_2 \in X_2.$$

This suggests that the x_2-block is minimized exactly.

To analyze the algorithm, it is convenient to consider an equivalent *single-block* problem, which only takes x_1 as its variable:

$$\min_{x_1 \in X_1} \quad \ell(x_1) + h_1(x_1) := \min_{x_1 \in X_1} \min_{x_2 \in X_2} f(x_1, x_2), \tag{5.4}$$

where we have defined $\ell(x_1) := \min_{x_2 \in X_2} g(x_1, x_2) + h_2(x_2)$. Let us denote an optimal solution of the inner problem $\min_{x_2 \in X_2} f(x_1, x_2)$ by the mapping: $x_2^*(x_1) : X_1 \to X_2$, which is a singleton for any $x_1 \in X_1$ by Assumption D(a). Next we analyze problem (5.4).

Let us define a new function

$$u(v_1; x_1) := u_1\left(v_1; x_1, x_2^*(x_1)\right) + h_2\left(x_2^*(x_1)\right). \tag{5.5}$$

First we argue that for all $x_1, v_1 \in X_1$, $u(v_1; x_1)$ is an upper bound for $\ell(v_1)$, and it satisfies Assumption C given in Sect. 4.1. Clearly Assumption C(a) is true because

$$\begin{aligned}\ell(x_1) &= g\left(x_1, x_2^*(x_1)\right) + h_2\left(x_2^*(x_1)\right) \\ &= u_1\left(x_1; x_1, x_2^*(x_1)\right) + h_2\left(x_2^*(x_1)\right) = u(x_1; x_1) \qquad (5.6)\end{aligned}$$

where the second equality is due to the fact that $u_1(x_1; x)$ is an upper bound function for $g(\cdot, x_2)$. The last equality is from the definition of $u(\cdot; \cdot)$.

Assumption C(b) is true because

$$\begin{aligned}u(v_1; x_1) &= u_1\left(v_1; x_1, x_2^*(x_1)\right) + h_2\left(x_2^*(x_1)\right) \\ &\ge g\left(v_1, x_2^*(x_1)\right) + h_2\left(x_2^*(x_1)\right) \ge \min_{x_2} g(v_1, x_2) + h_2(x_2). \qquad (5.7)\end{aligned}$$

To verify Assumption C(c), recall that by Assumption D, $\min_{x_2 \in X_2} f(x_1, x_2)$ has a *unique* solution, or equivalently for any given $x_1 \in X_1$, the mapping $x_2^*(x_1)$ is a singleton. By applying [13, Corollary 4.5.2–4.5.3], we obtain

$$\nabla \ell(x_1) = \nabla_1 g\left(x_1, \widetilde{x}_2\right), \;\forall\, x_1 \in X_1, \qquad (5.8)$$

where $\widetilde{x}_2 = \arg\min_{x_2 \in X_2} f(x_1, x_2)$. Therefore, we must have

$$\nabla \ell(x_1) = \nabla_1 g\left(x_1, \widetilde{x}_2\right) = \nabla u_1(x_1; x_1, \widetilde{x}_2) = \nabla u_1\left(x_1; x_1, x_2^*(x_1)\right) = \nabla u(x_1; x_1),$$

where the second equality comes from the fact that $u_1(\cdot; \cdot)$ satisfies Assumption B(c); the third inequality is because $\widetilde{x}_2 = x_2^*(x_1)$ by definition; the last equality is from (5.9). This verifies Assumption C(c).

The Lipschitz continuous gradient condition (with constant L_1) in Assumption C(d) can be verified by combining (5.3) and the following equality

$$\nabla u_1\left(v_1; x_1, x_2^*(x_1)\right) = \nabla u(v_1; x_1), \;\forall\, v_1, x_1 \in X_1. \qquad (5.9)$$

Now that we have verified that $u(v_1; x_1)$ given in (5.5) satisfies Assumption C, then Proposition 1 implies $\ell(\cdot)$ also has Lipschitz continuous gradient with constant L_1, that is

$$\|\nabla \ell(x_1) - \nabla \ell(v_1)\| \le L_1 \|x_1 - v_1\|, \;\forall\, v_1, x_1 \in X.$$

At this point it is clear that the 2-block BSUM algorithm with G–S update rule is in fact the SUM algorithm given in Sect. 4.1, where the iterates are generated by

$$x_1^{r+1} \in \arg\min u\left(x_1; x_1^r\right). \qquad (5.10)$$

By applying the argument leading to (4.11), we conclude that the 2-block BSUM in which the second block performs an exact minimization converges sublinearly. Also note that neither subproblems in (5.1) is required to be strongly convex, which suggests that the BCM applied to problem (5.1) converges sublinearly without block strong convexity. The precise statement is given in the following corollary.

Corollary 1 *Assume that Assumption A(a) and D hold for problem* (5.1). *Then we have the following.*

1. *Suppose that* $u_2(v_2; x) = g(x_1, v_2)$ *for all* $v_2 \in X_2,\ x \in X$ *and that* $u_1(v_1; x)$ *satisfies Assumption B(a)–(c) and the Lipschtiz continuous gradient condition* (2.8). *Then the 2-block BSUM algorithm with G–S rule is equivalent to the SUM algorithm and converges sublinearly, i.e.,*

$$\Delta^{r+1} \leq \frac{c_4}{\sigma_4}\frac{1}{r}, \tag{5.11}$$

   where c_4 *and* σ_4 *is given in* (4.11)*, with* L *in* (4.11) *replaced by* L_1.
2. *The BCM algorithm applied to* (5.1) *converges sublinearly with the same rate, again with* L *in* (4.11) *replaced by* L_1.

To conclude this section, we note that the schemes and analysis developed in this section are special in the sense that they heavily rely on the fact that $K = 2$, and the resulting transformation to the single block problem. At this point, it is unclear whether the same $\mathcal{O}(1/r)$ iteration complexity holds for a general K without the BSC condition. In the next section we will address the first issue and show that without BSC one can still achieve $\mathcal{O}(1/r)$ complexity.

6 Analysis of the BCM without per-block strong convexity

In this section, we consider the BCM algorithm below, which is the BSUM algorithm without using approximation for each block. We analyze its iteration complexity *without* the BSC assumption.

The Block Coordinate Minimization (BCM) Algorithm

At each iteration $r + 1$, pick an index set $\mathcal{C}^{r+1}$; update the variable blocks:

$$x_k^{r+1} \begin{cases} \in \arg\min_{x_k \in X_k}\ g\left(x_k, w_{-k}^{r+1}\right) + h_k(x_k), & \text{if } k \in \mathcal{C}^{r+1}; \\ = x_k^r, & \text{if } k \notin \mathcal{C}^{r+1}. \end{cases}$$

In the absence of the BSC property, there can be multiple optimal solutions for each subproblem. This makes it tricky to establish the convergence rate of BCM. Specifically, in the context of the three-step analysis framework presented herein, it is difficult to bound the sufficient descent of the objective using the size of of the successive iterates (as per Lemma 1). In this section, we overcome this obstacle by developing a variant of the sufficient descent estimate step. We will show that BCM with MBI, G–S and E-C rules has an iteration complexity of $O(1/r)$ for problem (1.1) without the BSC condition. Throughout this section we will impose Assumption A.

We first consider the MBI rule. We notice that the following is true

$$f(x^r) - f(x^{r+1}) \overset{\text{(i)}}{\geq} f(x^r) - f(\bar{x}^{r+1}) \overset{\text{(ii)}}{\geq} \frac{\gamma}{K}\|x^r - \hat{x}^{r+1}\|^2, \tag{6.1}$$

where $\bar{x}^{r+1}$ is the iterates obtained by any BSUM algorithm with MBI rule; $\hat{x}^{r+1}$ is defined in (2.2). In the above expression (ii) can be obtained using Lemma 1, while (i) is true because we used the exact minimization in each step. Then it is straightforward to establish, using the additional assumption that h is Lipschitz continuous, the same rate stated in part (3) of Theorem 1.

Next we show that the BCM algorithm with the G–S and E-C rules also achieves an $\mathcal{O}(1/r)$ iteration complexity, without the BSC assumption. These are the key results of this section.

The main difficulty in analyzing the BCM without the BSC is that the size of the difference of the successive iterates is no longer a good measure of the "sufficient descent". Indeed, due to the lack of per-block strong convexity, it is possible that a block variable travels a long distance without changing the objective value (e.g., it stays in the per-block optimal solution set).

Below we analyze the iteration complexity of BCM. We need to make use of the following key inequality due to Nesterov [21]; also see (4.5) for a proof. From Assumption A we know that g is convex and has Lipschitz continuous gradient with constant M, then we must have

$$g(x) - g(v) \geq \langle \nabla g(v), x - v\rangle + \frac{1}{2M}\|\nabla g(v) - \nabla g(x)\|^2, \ \forall\, v, x \in X. \tag{6.2}$$

Utilizing this inequality, the sufficient descent estimate is given by the following lemma.

Lemma 3 *Suppose Assumption A holds. Then for BCM with either G–S rule or the E-C rule, we have that for all $r \geq 1$*

$$\Delta^r - \Delta^{r+1} \geq \frac{1}{2M}\sum_{k=1}^{K}\|\nabla g\left(w_k^{r+1}\right) - \nabla g\left(w_{k+1}^{r+1}\right)\|^2. \tag{6.3}$$

Proof Suppose that $k \notin \mathcal{C}^{r+1}$, then we have the following trivial inequality

$$f\left(w_k^{r+1}\right) - f\left(w_{k+1}^{r+1}\right) \geq \frac{1}{2M}\|\nabla g\left(w_k^{r+1}\right) - \nabla g\left(w_{k+1}^{r+1}\right)\|^2 \tag{6.4}$$

as both sides of the inequality are zero.

Suppose $k \in \mathcal{C}^{r+1}$. Then by (6.2), we have that

$$\begin{aligned} &f\left(w_k^{r+1}\right) - f\left(w_{k+1}^{r+1}\right) \\ &\quad \geq \left\langle \nabla g\left(w_{k+1}^{r+1}\right), w_k^{r+1} - w_{k+1}^{r+1}\right\rangle + h_k\left(w_k^{r+1}\right) - h_k\left(w_{k+1}^{r+1}\right) \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2M}\|\nabla g\left(w_k^{r+1}\right)-\nabla g\left(w_{k+1}^{r+1}\right)\|^2\\
&\overset{\text{(i)}}{\geq}\langle\nabla_k g\left(w_{k+1}^{r+1}\right),x_k^r-x_k^{r+1}\rangle+h_k\left(x_k^r\right)-h_k\left(x_k^{r+1}\right)\\
&+\frac{1}{2M}\|\nabla g\left(w_k^{r+1}\right)-\nabla g\left(w_{k+1}^{r+1}\right)\|^2\\
&\overset{\text{(ii)}}{\geq}\frac{1}{2M}\|\nabla g\left(w_k^{r+1}\right)-\nabla g\left(w_{k+1}^{r+1}\right)\|^2, \qquad (6.5)
\end{aligned}
$$

where (i) is because w_{k+1}^{r+1} and w_k^{r+1} only differs by a single block; (ii) is due to the optimality of x_k^{r+1}. Summing over k, we have

$$
f(x^r)-f(x^{r+1})\geq\sum_{k=1}^{K}\frac{1}{2M}\|\nabla g\left(w_k^{r+1}\right)-\nabla g\left(w_{k+1}^{r+1}\right)\|^2. \qquad (6.6)
$$

This completes the proof of this lemma. □

Lemma 4 *Suppose Assumptions A is satisfied. Then*

1. *For the BCM with the G–S update rule, we have*

$$
(\Delta^{r+1})^2\leq 2K^2R^2\sum_{k=1}^{K}\|\nabla g\left(w_{k+1}^{r+1}\right)-\nabla g\left(w_k^{r+1}\right)\|^2,\ \forall x^*\ \in X^*.
$$

2. *For the BCM with the period-T E-C update rule, we have*

$$
(\Delta^{r+T})^2\leq 2TK^2R^2\sum_{k=1}^{K}\sum_{t=1}^{T}\|\nabla g\left(w_{k+1}^{r+t}\right)-\nabla g\left(w_k^{r+t}\right)\|^2,\ \forall\, x^*\in X^*.
$$

Proof We only show the second part of the claim, as the proof for the first part is simply a special case. Define a new index set $\{r_k\}$ as in (3.14). Recall that we have $x_k^{r_k}=x_k^{r+T}$, for all k. We have the following series of inequalities

$$
\begin{aligned}
f(x^{r+T})-f(x^*)&\leq\sum_{k=1}^{K}\left\langle\nabla_k g(x^{r+T}),x_k^{r+T}-x_k^*\right\rangle+\sum_{k=1}^{K}h_k\left(x_k^{r_k}\right)-h_k\left(x_k^*\right)\\
&=\sum_{k=1}^{K}\left\langle\nabla_k g(x^{r+T})-\nabla_k g\left(w_{k+1}^{r_k}\right),x_k^{r+T}-x_k^*\right\rangle\\
&\quad+\left\langle\nabla_k g\left(w_{k+1}^{r_k}\right),x_k^{r+T}-x_k^*\right\rangle+h_k\left(x_k^{r_k}\right)-h_k\left(x_k^*\right)\\
&\overset{\text{(i)}}{\leq}\sum_{k=1}^{K}\left\langle\nabla_k g(x^{r+T})-\nabla_k g\left(w_{k+1}^{r_k}\right),x_k^{r+T}-x_k^*\right\rangle
\end{aligned}
$$

$$\begin{aligned}
&\leq \sum_{k=1}^{K} \|\nabla g(x^{r+T}) - \nabla g\left(w_{k+1}^{r_k}\right)\| \|x_k^{r+T} - x_k^*\| \\
&\leq \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{j=1}^{K} \|\nabla g\left(w_{j+1}^{r+t}\right) - \nabla g\left(w_j^{r+t}\right)\| \|x_k^{r+T} - x_k^*\| \\
&\leq \sum_{t=1}^{T}\sum_{j=1}^{K} \|\nabla g\left(w_{j+1}^{r+t}\right) - \nabla g\left(w_j^{r+t}\right)\| \sum_{k=1}^{K} \|x_k^{r+T} - x_k^*\|
\end{aligned}$$

where in (i) we have used the optimality of $x_k^{r_k}$ and $x_k^{r_k} = x_k^{r+T}$, for all k. Then taking the square on both sides, we obtain

$$\left(f(x^{r+T}) - f(x^*)\right)^2 \leq TK^2R^2 \sum_{t=1}^{T}\sum_{k=1}^{K} \|\nabla g\left(w_{k+1}^{r+t}\right) - \nabla g\left(w_k^{r+t}\right)\|^2. \tag{6.7}$$

The proof is complete. □

Combining these two results, and utilizing the technique in Theorem 1, we readily have the following main result for BCM.

Theorem 2 *Suppose Assumption A holds true. We have the following.*

1. *Let $\{x^r\}$ be the sequence generated by BCM with G–S rule. Then we have*

$$\Delta^r = f(x^r) - f^* \leq \frac{c_5}{\sigma_5}\frac{1}{r}, \ \forall\, r \geq 1, \tag{6.8}$$

where the constants are given below

$$c_5 = \max\{4\sigma_5 - 2, f(x^1) - f^*, 2\}, \quad \sigma_5 = \frac{1}{2MK^2R^2}, \tag{6.9}$$

2. *Let $\{x^r\}$ be the sequence generated by BCM with E-C rule. Then we have*

$$\Delta^r = f(x^r) - f^* \leq \frac{c_6}{\sigma_6}\frac{1}{r-T}, \ \forall\, r > T, \tag{6.10}$$

where the constants are given below

$$c_6 = \max\{4\sigma_6 - 2, f(x^1) - f^*, 2\}, \quad \sigma_6 = \frac{1}{2K^2TR^2M}. \tag{6.11}$$

Remark 1 Our analysis above implies that when using the BCM (or equivalently the IWFA algorithm [35]) to solve the rate optimization problem given in Example 1, a sublinear rate can be obtained regardless of the rank of the channel matrices $\{H_k\}$. In fact, it is easy to check that Assumption A is satisfied for this problem; see for example a related discussion in [27, Section V-A]. Then Theorem 2 implies that IWFA converges in a rate $O(1/r)$, regardless of the rank of the channel matrices. Prior to our work, no convergence rate analysis has been done for the IWFA when solving problem (2.13).

7 Discussion and concluding remarks

In this paper we have analyzed the iteration complexity of a family of BCD-type algorithms for solving general convex nonsmooth problems of the form (1.1). Using a three-step argument, we show that the family of BCD-type algorithms, which includes BCM, BCGD, BCPG algorithms with G–S, E-C, G–So and MBI update rules, converges globally in a sublinear rate of $\mathcal{O}(1/r)$. It should be noted that in case of the classical BCM algorithm, the sublinear rate can be achieved even without the per-block strong convexity.

We note that the structure of the three-step approach, i.e., estimate the sufficient descent, estimate the cost-to-go and obtain the rate of convergence, is not new. For example Luo and Tseng in [18] has developed a three-step argument for proving linear convergence rate of certain descent method (including BCD) for certain non-strongly convex problems. Beck and Tetruashvili [2] has used such argument for showing sublinear convergence for using cyclic BCPG to solve smooth constrained problem. However it is important to note that these works differ significantly in how each step is proved. For example, the proof of cost-to-go in Lemma 2 differs from its counterpart in [2] and [18] because we have to take into consideration the nonsmooth function h_k's, as well as various different update rules. The proof of both sufficient descent and cost-to-go steps in Lemmas 3–4 differ from those in [2, 18], because without per-block strong convexity a different measure is used to show sufficient descent, which to the best of our knowledge has not been used in any related analysis before.

As a future work, it will be interesting to see whether the three-step approach can be extended to establish the iteration complexity bounds for other first order methods. Also we will investigate whether the problem dependent constants in front of the $1/r$ can be further reduced, or even be made independent of problem dimension K.

References

1. Beck, A.: On the convergence of alternating minimization with applications to iteratively reweighted least squares and decomposition schemes. SIAM J. Optim. **25**(1), 185–209 (2015)
2. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM J. Optim. **23**(4), 2037–2060 (2013)
3. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont (1999)
4. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont (1996)
5. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods, 2nd edn. Athena Scientific, Belmont (1997)
6. Chen, B., He, S., Li, Z., Zhang, S.: Maximum block improvement and polynomial optimization. SIAM J. Optim. **22**(1), 87–107 (2012)
7. Combettes, P., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer Optimization and Its Applications, pp. 185–212. Springer, New York (2011)
8. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, Hoboken (2005)

9. Daubechies, I., DeVore, R., Fornasier, M., Gunturk, C.S.: Iteratively reweighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
10. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
11. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. Oper. Res. Lett. **26**, 127–136 (2000)
12. He, B., Liao, L., Han, D., Yang, H.: A new inexact alternating directions method for monotone variational inequalities. Math. Program. **92**(1), 103–118 (2002)
13. Hiriart-Urruty, J.-B., Lemarechal, C.: Convex Analysis and Minimization Algorithms I: Fundamentals. Springer, Berlin (1996)
14. Hong, M., Razaviyayn, M., Luo, Z.-Q., Pang, J.-S.: A unified algorithmic framework for block-structured optimization involving big data. IEEE Signal Process. Mag. **33**(1), 57–77 (2016)
15. Lu, Z., Xiao, L.: On the complexity analysis of randomized block-coordinate descent methods. Math. Program. **152**(1), 615–642 (2015)
16. Luo, Z.-Q., Tseng, P.: On the convergence of the coordinate descent method for convex differentiable minimization. J. Optim. Theory Appl. **72**(1), 7–35 (1992)
17. Luo, Z.-Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. **30**(2), 408–425 (1992)
18. Luo, Z.-Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46–47**, 157–178 (1993)
19. Luo, Z.-Q., Tseng, P.: On the convergence rate of dual ascent methods for strictly convex minimization. Math. Oper. Res. **18**(4), 846–867 (1993)
20. Mairal, J.: Optimization with first-order surrogate functions. In: The Proceedings of the International Conference on Machine Learning (ICML) (2013)
21. Nesterov, V.: Introductory Lectures on Convex Optimization: A Basic Course. Springer, Berlin (2004)
22. Nesterov, Y.: Efficiency of coordiate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
23. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, Cambridge (1972)
24. Razaviyayn, M., Hong, M., Luo, Z.-Q.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Optim. **23**(2), 1126–1153 (2013)
25. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**, 1–38 (2014)
26. Saha, A., Tewari, A.: On the nonasymptotic convergence of cyclic coordinate descent method. SIAM J. Optim. **23**(1), 576–601 (2013)
27. Scutari, G., Facchinei, F., Song, P., Palomar, D.P., Pang, J.-S.: Decomposition by partial linearization: parallel optimization of multi-agent systems. IEEE Trans. Signal Process. **63**(3), 641–656 (2014)
28. Shalev-Shwartz, S., Tewari, A.: Stochastic methods for ℓ_1 regularized loss minimization. J. Mach. Learn. Res. **12**, 1865–1892 (2011)
29. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **103**(9), 475–494 (2001)
30. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. **125**(2), 263–295 (2010)
31. Tseng, P., Yun, S.: Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. J. Optim. Theory Appl. **140**, 513–535 (2009)
32. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**, 387–423 (2009)
33. Wang, X., Yuan, X.: The linearized alternating direction method of multipliers for dantzig selector. SIAM J. Sci. Comput. **34**(5), 2792–2811 (2012)
34. Yang, J., Zhang, Y., Yin, W.: A fast alternating direction method for TVL1-L2 signal reconstruction from partial fourier data. IEEE J. Sel. Top. Signal Process. **4**(2), 288–297 (2010)
35. Yu, W., Rhee, W., Boyd, S., Cioffi, J.M.: Iterative water-filling for Gaussian vector multiple-access channels. IEEE Trans. Inf. Theory **50**(1), 145–152 (2004)
36. Zhang, H., Jiang, J., Luo, Z.-Q.: On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. J. Oper. Res. Soc. China **1**(2), 163–186 (2013)
37. Zhang, X., Burger, M., Osher, S.: A unified primal-dual algorithm framework based on Bregman iteration. J. Sci. Comput. **46**(1), 20–46 (2011)