Block Coordinate Proximal Gradient Method for Nonconvex Optimization Problems: Convergence Analysis

Xiangfeng Wang∗ Xiaoming Yuan† Shangzhi Zeng‡ Jin Zhang§ Jinchuan Zhou¶

February, 2018

Abstract

We propose a block coordinate proximal gradient method for a composite minimization problem with two nonconvex function components in the objective while only one of them is assumed to be differentiable. Under some per-block Lipschitz-like conditions based on Bregman distance, but without the global Lipschitz continuity of the gradient of the differentiable function, we prove that any accumulation point of the sequence is a stationary point of the model. We further show that the stationarity is the “best” one if the global Lipschitz continuity is additionally assumed, and even the local minimizer for some special cases. Convergence analysis without the global Lipschitz continuity and the enhanced stationarity analysis make our results different from existing results in both the convex and nonconvex contexts.

1 Introduction

Many applications can be modelled as a composite minimization problem with an objective function as the sum of two functions, one is possibly non-differentiable and the other is differentiable. That is,

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x).$$

Among many algorithms applicable to this composite minimization problem is the well-studied proximal gradient method, see, e.g. [27]. In this paper, we focus on the specific structured scenario:

$$(P) : \min_{x \in \mathbb{R}^n} F(x) = f(x_1, \cdots, x_m) + \sum_{i=1}^m g_i(x_i),$$

where $x = (x_1, \cdots, x_m) \in \mathbb{R}^n$, $x_i \in \mathbb{R}^{n_i}$ and $\sum_{i=1}^m n_i = n$. We assume that the functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^{n_i} \to \mathbb{R}$ are proper and lower semi-continuous (l.s.c.); $f$ is continuously differentiable. To

∗Shanghai Key Lab for Trustworthy Computing, School of Computer Science and Software Engineering, East China Normal University, China. Email: xfwang@sei.ecnu.edu.cn
†Department of Mathematics, The University of Hong Kong, Hong Kong, China. Email: xmyuan@hku.hk
‡Department of Mathematics, The University of Hong Kong, Hong Kong, China. Email: zengsz@connect.hku.hk
§Department of Mathematics, Hong Kong Baptist University, Hong Kong. HKBU Institute of Research and Continuing Education, Shenzhen, China. Email: zhangjin@hkbu.edu.hk
¶Department of Statistics, School of Mathematics and Statistics, Shandong University of Technology, Zibo, China. Email: jinchuanzhou@163.com
avoid triviality, \( v(\mathcal{P}) := \min \{ F(x) \} > -\infty \) is assumed. Note that \( f \) and all \( g_i \)'s in (1.1) are possibly nonconvex in our discussion.

The specific model (1.1) captures various applications, especially some arising from big-data scenarios. For such a case, the variable \( x \) can (should) be treated separately as \( m \) blocks in smaller dimensions and the minimization should be performed block by block among the variables, rather than as a whole. This idea had inspired the longstanding family of efficient methods known as block coordinate descent (BCD) methods. The subproblems over variable blocks can be solved in different ways. Consequently, specific BCD type methods have been proposed in the literature such as the block coordinate minimization (BCM), block coordinate gradient descent (BCGD) and block coordinate proximal gradient (BCPG) methods. We will briefly review BCD type methods in Section 2.

For analyzing the convergence for various first-order methods including the mentioned BCD type methods, the gradient of \( f \), i.e., \( \nabla f \), is usually required to be globally Lipschitz continuous on the entire space defining the function \( f \). Some nice theoretical convergence results have been obtained in the literature, see, e.g., [14] and references therein. Meanwhile, the global Lipschitz continuity assumption excludes many important applications (e.g., [9, 13, 33]) whose data-fidelity functions satisfy the so-called Kullback-Liebler divergence (KL-divergence, see [16]), and it is known that the KL-divergence does not necessarily imply the global Lipschitz continuity. It thus becomes interesting to know whether or not the convergence of some BCD type methods for the model (1.1) remains without the global Lipschitz continuity of \( \nabla f \). Our primary purpose is to answer this question.

Motivated by recent work [3, 7] for the proximal gradient method, we introduce some similar Lipschitz-type condition (see its definition in Section 4) and propose a Bregman-distance-based block coordinate proximal gradient (BCPG) method for (1.1). Without the global Lipschitz continuity of \( \nabla f \), we show that any accumulation point of the sequence generated by the proposed algorithm is a stationary point in sense of the limiting subdifferential (see its definition in Section 3). This convergence result generalizes known results of the classical BCPG method in convex settings to the nonconvex setting.

If \( \nabla f \) is indeed globally Lipschitz continuous, we shall show that the stationarity in sense of the limiting subdifferential can be enhanced to a strongly stationary point in sense of the Fréchet (regular) subdifferential (see its definition in Section 3). Since the Fréchet (regular) subdifferential is more compact than the limiting subdifferential (see (3.1)) and the former characterizes the local optimality more precisely [24], this is a stronger convergence result. In fact, according to [18], the strong stationarity can be regarded the “best” one because all other stationary points fail to exclude the existence of descent directions and thus they may not be good enough. Note that when the convergence of various first-order methods is analyzed for nonconvex optimization problems, usually only the stationarity in sense of the limiting subdifferential is obtained. Therefore, our analysis for the strong stationarity seems to be novel in the literature. Furthermore, for some special cases, e.g., the function \( g \) in (5.2) is the indicator function of a set, then our analysis implies that the strongly stationary point is just the local minimization point despite the nonconvexity of the problem.

Our contributions may be summarized as follows. (i) We propose a Bregman-distance-based BCPG method for the nonconvex optimization problem (1.1) without the global Lipschitz continuity of \( \nabla f \) and analyze the convergence of its sequence to a stationary point. (ii) We initiate a new analytic framework to show that the stationarity can be enhanced if \( \nabla f \) is indeed globally Lipschitz continuous (but both \( f \) and \( g \) in (1.1) are still nonconvex). To the best of our knowledge, these results seem to be new.

For completeness, we test the Possion linear inverse problem with \( \ell_1 \) regularization (for some imaging datasets) in Section 6 and verify the convergence of the proposed Bregman-distance-based BCPG method. Finally, some conclusions are made in Section 7. Some supplementary materials for the proof of some lemmas and theorems are included as well.
2 Related Work

In this section, we briefly review BCD type methods for both the convex and nonconvex cases of the model (1.1). For different block selection schemes of BCD type methods, we refer to, e.g., [10, 12, 20, 21, 25, 26, 30, 31, 32]; and [14, 34] for some review papers. Also, there is a rich literature discussing the convergence (and convergence rates) for various BCD type methods in both the convex and nonconvex settings. We refer to [23, 28] and references therein for a comprehensive analysis. For the convex case, some results about the iteration complexity for various BCD type methods have been studied in, e.g., [5, 15, 25, 30]. Recently, based on the Kurdyka-Łojasiewicz property [17, 22], convergence of BCD type methods (and other first order methods) has been well studied in numerous literatures, e.g., [1, 6, 11, 19, 35]. Note that all these works require the globally Lipschitz continuity of $\nabla f$.

3 Notation and Preliminaries

We recall some notation and preliminaries that will be used in later analysis.

**Definition 3.1.** Let $\varphi : \mathbb{R}^d \to (-\infty, \infty]$ be a lsc function and $x_0 \in \text{dom} \varphi$. The Fréchet (regular) subdifferential of $\varphi$ at $x_0$ is the set

$$\partial^F \varphi(x_0) = \left\{ \xi \in \mathbb{R}^d \mid \liminf_{h \to 0} \frac{\varphi(x_0 + h) - \varphi(x_0) - \langle \xi, h \rangle}{\| h \|} \geq 0 \right\}.$$ 

The limiting (Mordukhovich or basic) subdifferential of $\varphi$ at $x_0$ is the set

$$\partial \varphi(x_0) := \left\{ \xi \in \mathbb{R}^d \mid \exists x_k \to x_0, \xi_k \to \xi \text{ with } \xi_k \in \partial^F \varphi(x_k) \right\}.$$ 

By the definitions, obviously the following inclusion holds:

$$\partial^F \varphi(x_0) \subset \partial \varphi(x_0). \tag{3.1}$$

**Definition 3.2.** Legendre function. Let $h : X \to (-\infty, \infty]$ be a lsc proper convex function. It is called

1. essentially smooth if $h$ is differentiable on $\text{int} \text{ dom} h$, and $\| \nabla h(x^k) \| \to \infty$ for every sequence $\{x^k\} \subset \text{int} \text{ dom} h$ converging to a boundary point of $\text{dom} h$ as $k \to +\infty$;

2. Legendre type if $h$ is essentially smooth and strictly convex on $\text{int} \text{ dom} h$.

**Definition 3.3.** Bregman distance. Let $h$ be a given Legendre function. For all $x \in \text{dom} h$ and $y \in \text{int} \text{ dom} h$, the Bregman distance is given by

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle,$$

where $D_h$ is strictly convex with respect to its first argument. Moreover, $D_h(x, y) \geq 0$ for all $(x, y) \in \text{dom} h \times \text{int dom} h$, and it equals to zero if and only if $x = y$. However, $D_h$ is in general asymmetric, i.e., $D_h(x, y) \neq D_h(y, x)$.

**Lemma 3.4.** Three points identity [8]. For any $x \in \text{dom} h$ and $y, z \in \text{int} \text{ dom} h$, the following identity holds

$$D_h(x, z) - D_h(x, y) - D_h(y, z) = \langle \nabla h(y) - \nabla h(z), x - y \rangle.$$

**Definition 3.5.** Symmetry coefficient. [3]. Given a Legendre function $h$, its symmetry coefficient is defined by

$$\alpha(h) = \inf \left\{ \frac{D_h(x, y)}{D_h(y, x)} \mid x \neq y \right\} \in [0, 1].$$
To simplify notation, let us define a set of auxiliary variables and for all $i = 1, \cdots, m$

$$
(x_i, z_{\neq i}) = (z_{1, \cdots, i-1, i, i+1, \cdots, z_{m}}),

(x_{<i}, z_{\geq i}) = (x_{1, \cdots, i-1, z_{i, \cdots, z_{m}}}).
$$

The set of stationary points of the problem (1.1) is defined as $X^* := \{ x \mid 0 \in \nabla f(x) + \partial g(x) \}$, where $\partial g(x)$ represents the limiting subdifferential of $g$ at $x$.

## 4 Bregman-distance-based BCPG

We are now ready to present a Bregman-distance-based BCPG for the nonconvex case of (1.1) without the global Lipschitz continuity of $\nabla f$.

### 4.1 Per-block Lipschitz-like condition

Inspired by [3, 7], we introduce the so-called per-block Lipschitz-like condition (pbL-condition) for our analysis.

**Definition 4.1.** Per-Block Lipschitz-like Condition of $(f, h_i)$ on block $x_i$: The function $f(x_{\ell}, z_{\neq \ell}) : \mathbb{R}^n \to (-\infty, \infty]$ is said to satisfy the pbL-condition with constant $L_i$ based on $h_i$, i.e., $\exists L_i > 0$ such that $L_i h_i(x_i) - f(x_{\ell}, z_{\neq \ell})$ is convex on $\text{int dom } h_i$, where $h_i : \mathbb{R}^n \to (-\infty, \infty]$ is a Legendre function.

It is notable that when $h_i(x_i) = \frac{1}{2} \| x_i \|^2$, the per-block Lipschitz-like condition recovers the property that $\nabla_i f(x_{\ell}, z_{\neq \ell})$ being Lipschitz continuous with respect to $x_i$ with a constant $L_i$. Furthermore, we have the following descent lemma.

**Lemma 4.2.** Descent Lemma without Lipschitz Gradient Continuity. Let $h_i$ be a Legendre function, $f$ be a function and continuously differentiable on $\text{int dom } h_i$ with respect to $x_i$. Then, the pbL condition of $(f, h_i)$ on block $x_i$ is equivalent to $f(x_i, z_{\neq i}) \leq f(y_i, z_{\neq i}) + \langle \nabla f(y_i, z_{\neq i}), x_i - y_i \rangle + L_i D_{h_i}(x_i, y_i).

**Proof.** $L_i h_i(x_i) - f(x_{\ell}, z_{\neq \ell})$ is convex, the by the convex property, we have

$$
[L_i h_i(x_i) - f(x_{\ell}, z_{\neq \ell})] \geq [L_i h_i(y_i) - f(y_{\ell}, z_{\neq \ell})] + \langle L_i \nabla h_i(y_i), \nabla f(y_{\ell}, z_{\neq \ell}), x_i - y_i \rangle.
$$

With proper rearrangement, we can guarantee that

$$
f(x_i, z_{\neq i}) \leq f(y_i, z_{\neq i}) + \langle \nabla f(y_i, z_{\neq i}), x_i - y_i \rangle + L_i [h_i(x_i) - h_i(y_i)] - \nabla h_i(y_i)

\leq f(y_i, z_{\neq i}) + \langle \nabla f(y_i, z_{\neq i}), x_i - y_i \rangle + L_i D_{h_i}(x_i, y_i),$$

which is the descent inequality we need.

Given a Legendre function $h_i$, for all $x_i \in \text{int dom } h$ and $z \in \mathbb{R}^n$, we define the following Block Bregman Proximal Gradient Mapping as following:

$$
\mathcal{T}^h_{\lambda}(x_i, z_{\neq i}) := \arg \min_u \left\{ g_i(u) + \langle \nabla f(x_i, z_{\neq i}), u - x_i \rangle + \frac{1}{\lambda} D_{h_i}(u, x_i) \right\}.
$$

(4.1)

Note that when $h_i(x) = \frac{1}{2} \| x \|^2$, (4.1) reduces to the classical block proximal gradient mapping. The solvability of (4.1) depends on $g_i$ and $h_i$. In general, because of the nonconvexity of $g_i$, finding a global minimizer of (4.1) is difficult. But we shall show by some examples in Section 4.2 that this problem could be easy enough to have closed-form solutions for some special cases of $g_i$ and $h_i$ which have wide applications. So our analysis is conducted under the assumption that (4.1) is solvable. This essentially requires that each mapping $\mathcal{T}^h_{\lambda}$ is nonempty and it maps $\text{int dom } h_i$ in $\text{int dom } h_i$. 

4
4.2 Algorithm

In order to expose our ideas more clearly, we only focus on the cyclic scheme for selecting blocks of variables. But our analysis can be extended for other selection schemes such as the essentially cyclic, Gauss-Southwell, maximum block improvement schemes, and even the randomized scheme. The cyclic Bregman-distance-based BCPG for the nonconvex case of (1.1) without the global Lipschitz continuity of \( \nabla f \) is presented below.

**Algorithm 1 Cyclic-NonLip-Noconvex-Bregman-BCPG**

**Require:** Choose Legendre function \( \{ h_i \} \) such that we have per-block Lipschitz-like condition of \( \{ (f, h_i) \} \); Start with any \( x^0 \) with \( x^0_i \in \text{int dom } h_i \). Let \( \mathcal{T}_{\lambda}^{h_i} \) be defined in (4.1);

1: while not convergent do
2:  
3:     For each \( k = 1, \cdots \) with \( \lambda_k \in \left( 0, \frac{1}{\max\{L_i\}} \right) \), generate the sequence \( \{ x^k \} \) via
4:       \[ \begin{align*}
x_i^{k+1} & \in \mathcal{T}_{\lambda_k}^{h_i} \left( x_{<i}^{k+1}, x_{\geq i}^k \right), \quad i = 1, \cdots, m. \end{align*} \] (4.2)

3:  end while

It is easy to see that if we choose specific \( h_i \) in (4.1), then the step (4.2) reduces to the per-block proximal gradient step of the classical BCPG method (see, e.g., [5, 14, 15]). Hence, Algorithm 1 can be considered as a generalization of the classical BCPG method in Bregman-distance framework.

We elaborate on the subproblem (4.2) and show that it can be solved via two steps which hereafter are called the **Bregman gradient step** and **Bregman proximal step**, respectively. First, the optimality condition of (4.2) is

\[
0 \in \partial g_i(x_i^{k+1}) + \nabla_i f(x_{<i}^{k+1}, x_{\geq i}^k) + \frac{\nabla h_i(x_i^{k+1}) - \nabla h_i(x_i^k)}{\lambda_k}.
\] (4.3)

For convenience, let us denote

\[
p_{\lambda_k}^{h_i}(x_{<i}^{k+1}, x_{\geq i}^k) := \nabla h_i^* \left[ \nabla h_i(x_i^k) - \lambda_k \nabla_i f(x_{<i}^{k+1}, x_{\geq i}^k) \right],
\] (4.4)

where \( h_i^* \) is the conjugate function of \( h_i \). Then, (4.4) can be considered as a **Bregman gradient step**. Now, it follows from \( \nabla h_i^{-1} = \nabla h_i^* \) that the optimality condition can be rewritten as

\[
0 \in \partial g_i(x_i^{k+1}) + \nabla h_i(x_i^{k+1}) - \nabla h_i \left( p_{\lambda_k}^{h_i}(x_{<i}^{k+1}, x_{\geq i}^k) \right).
\]

Therefore, \( x_i^{k+1} \) can be calculated by

\[
x_i^{k+1} = \arg \min_{x_i} \left\{ g_i(x_i) + \frac{1}{\lambda_k} D_{h_i} \left( x_i, p_{\lambda_k}^{h_i}(x_{<i}^{k+1}, x_{\geq i}^k) \right) \right\},
\]

which indicates that the subproblem (4.2) can be solved if \( p_{\lambda_k}^{h_i}(x_{<i}^{k+1}, x_{\geq i}^k) \) and

\[
\text{Prox}_{\lambda_k g_i}^{h_i}(u) := \arg \min_{x_i} \left\{ g_i(x_i) + \frac{D_{h_i}(x_i, u)}{\lambda_k} \right\}
\] (4.5)

are both solvable \((i = 1, \cdots, m)\). We call (4.5) a **Bregman proximal step** because of its Bregman distance. Indeed, conceptually these two steps can be merged as

\[
x_i^{k+1} \in \text{Prox}_{\lambda_k g_i}^{h_i} \left[ p_{\lambda_k}^{h_i}(x_{<i}^{k+1}, x_{\geq i}^k) \right].
\]
Below we give some concrete examples to show that the subproblem (4.2) can be solved easily via these two steps. Hence, for some applications, Algorithm 1 is implementable.

**Bregman gradient step.** Given \( h_i \), a key issue for implementing (4.4) is identifying \( h^*_i \) and then calculating \( \nabla h^*_i \). We list several popular choices of \( h_i \) in Table 1, for which \( \nabla h^*_i \) can be computed easily.

<table>
<thead>
<tr>
<th>( h_i(x_i) )</th>
<th>( h^*_i(y_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{2} | x_i |^2 )</td>
<td>( \frac{1}{2} | y_i |^2 )</td>
</tr>
<tr>
<td>( \sum_{j=1}^{m} (x_i)_j \log (x_i)_j )</td>
<td>( \sum_{j=1}^{m} (\exp(y_i)_j - 1) )</td>
</tr>
<tr>
<td>( - \sum_{j=1}^{m} \log (x_i)_j )</td>
<td>( - \sum_{j=1}^{m} \log (-y_i)_j )</td>
</tr>
</tbody>
</table>

Table 1: Examples of \( h_i \) and \( h^*_i \)

**Bregman proximal step.** To execute (4.5), we can choose appropriate \( h_i \) in accordance with the given function \( g_i \) so that (4.5) could be solved efficiently. For instance, if \( g_i(x_i) = \|x_i\|_1 \) or \( \|x_i\|_2 \), \( h_i(x_i) \) can be chosen as \( \frac{1}{2} \|x_i\|^2 \) (\( \ell_2 \)-norm), then (4.5) reduces to the classical proximal gradient operation. If \( g_i(x_i) = \|x_i - a\|_1 \), then we can choose \( h(x_i) = \sum_{j=1}^{m} (x_i)_j \log (x_i)_j \) and for \( j = 1, \cdots, n_i \), we have

\[
\text{Prox}_{\lambda_k g_i}^h (y_i) = \begin{cases} 
\frac{(y_i)_j}{1 - \lambda_k (y_i)_j} & \text{if } (y_i)_j < \frac{a_j}{1 - \lambda_k a_j}, \\
\frac{a_j}{1 - \lambda_k a_j} & \text{if } (y_i)_j \in \left[ \frac{a_j}{1 - \lambda_k a_j}, \frac{a_j}{1 + \lambda_k a_j} \right], \\
\frac{a_j}{1 + \lambda_k a_j} & \text{if } (y_i)_j > \frac{a_j}{1 - \lambda_k a_j}.
\end{cases}
\]

Else, we can also take \( h(x_i) = - \sum_{j=1}^{m} \log (x_i)_j \). Then, for \( j = 1, \cdots, n_i \), we have

\[
\text{Prox}_{\lambda_k g_i}^h (y_i) = \begin{cases} 
\frac{(y_i)_j}{1 - \lambda_k (y_i)_j} & \text{if } (y_i)_j < \frac{a_j}{1 - \lambda_k a_j}, \\
\frac{a_j}{1 - \lambda_k a_j} & \text{if } (y_i)_j \in \left[ \frac{a_j}{1 - \lambda_k a_j}, \frac{a_j}{1 + \lambda_k a_j} \right], \\
\frac{a_j}{1 + \lambda_k a_j} & \text{if } (y_i)_j > \frac{a_j}{1 - \lambda_k a_j}.
\end{cases}
\]

We refer to [3] and references therein for more such examples.

### 4.3 Convergence

We first analyze the convergence of the sequence generated by the proposed Algorithm 1. Since the nonconvex and non-global Lipschitz continuity case of (1.1) is considered, only some weaker convergence result can be obtained. The convergence result in this general setting is summarized in the following theorem.

**Theorem 4.3.** Let \( \{x^k\} \) be the sequence generated by Algorithm 1 with \( \lambda_k \in \left(0, \frac{1}{\max_{L_k}}\right) \). Then the following assertions can be guaranteed:

1. The sequence \( \{F(x^k)\} \) is non-increasing;
2. \( \sum_{k=1}^{\infty} \left[ \sum_{i=1}^{m} D_{h_i} \left( x^k_i, x^{k-1}_i \right) \right] < \infty \), and hence the sequence \( \left\{ \sum_{i=1}^{m} D_{h_i} \left( x^k_i, x^{k-1}_i \right) \right\} \) converges to zero;
3. \( \min_{1 \leq k \leq n} \left[ \sum_{i=1}^{m} D_{h_i} \left( x^k_i, x^{k-1}_i \right) \right] \leq \frac{F(x^0) - v(P)}{n^2} \);  
4. Any limiting point \( \bar{x} \) of \( \{x^k\} \) is a stationary point in sense of the limiting subdifferential defined in (1.1), i.e., \( \bar{x} \in X^* \).
Proof. Because of the page limitation, we only present a brief framework of the proof. More details can be found in the supplemental materials. The most important step in this theorem is establishing the so-called sufficient decrease property, i.e.,

\[ F(x^{k+1}) \leq F(x^k) - \sum_{i=1}^{m} \left[ \left( \frac{1}{\lambda_k} - L_i \right) D_{h_i}(x^{k+1}_i, x^k_i) \right]. \] (4.6)

Recall the subproblem in each iteration

\[ x^{k+1}_i \in \arg \min_{x_i \in X_i} \left\{ g_i(x_i) + \left\langle \nabla_i f(x^{k+1}_i, x^k_{i+1}), x_i - x^k_i \right\rangle + \frac{1}{\lambda_k} D_{h_i}(x^{k+1}_i, x^k_i) \right\}, \]

which indicates that

\[ g_i(x^{k+1}_i) + \left\langle \nabla_i f(x^{k+1}_i, x^k_{i+1}), x^{k+1}_i - x^k_i \right\rangle + \frac{1}{\lambda_k} D_{h_i}(x^{k+1}_i, x^k_i) \leq g_i(x^k_i). \] (4.7)

Recall Lemma 4.2, we have

\[ f(x_i, z_{\neq i}) \leq f(y_i, z_{\neq i}) + \langle \nabla_i f(y_i, z_{\neq i}), x_i - y_i \rangle + L_i D_{h_i}(x_i, y_i). \]

Let \( x_i = x^{k+1}_i, y_i = x^k_i, z = (x^{k+1}_{i+1}, x^k_{i+1}) \), we obtain

\[ f(x^{k+1}_i, x^k_{i+1}) \leq f(x^{k+1}_i, x^k_{i+1}) + \langle \nabla_i f(x^{k+1}_i, x^k_{i+1}), x^{k+1}_i - x^k_i \rangle + L_i D_{h_i}(x^{k+1}_i, x^k_i). \] (4.8)

By combing with (4.7), we have

\[ f(x^{k+1}_i, x^k_{i+1}) \stackrel{(4.7)}{\leq} f(x^{k+1}_i, x^k_{i+1}) + g_i(x^k_i) - g_i(x^{k+1}_i) - \left( \frac{1}{\lambda_k} - L_i \right) D_{h_i}(x^{k+1}_i, x^k_i). \] (4.9)

Summing (4.9) from \( i = 1 \) to \( i = m \), we can guarantee

\[ F(x^{k+1}) \leq F(x^k) - \sum_{i=1}^{m} \left[ \left( \frac{1}{\lambda_k} - L_i \right) D_{h_i}(x^{k+1}_i, x^k_i) \right], \] (4.10)

while \( (x^{k+1}_{m+1}, x^k_{m+1}) := x^{k+1} \). If we choose

\[ \lambda_k \in \left( 0, \frac{1}{\max \{L_i\}} \right), \]

we obtain the decrease of the sequence \( \{F(x^k)\} \).

Discussions: \( g_i \) is assumed convex. Recall the global optimality condition of each iteration of Algorithm 1, we can obtain

\[ g_i(x_i) - g_i(x^{k+1}_i) + \langle \nabla_i f(x^{k+1}_i, x^k_{i+1}), x_i - x^{k+1}_i \rangle - \frac{1}{\lambda_k} \left( \nabla h_i(x^{k+1}_i) - \nabla h_i(x^k_i) \right) \geq 0, \quad \forall x_i \in X_i. \] (4.11)
Similarly combined with (4.8), we guarantee that
\[
\begin{align*}
  f \left( x_{k+1}^{i+1}, x_k^i \right) & \leq f \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) + \left( \nabla f \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \cdot x_{i_1} ^{k_1} - x_k^i \right) + L_i D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \\
  & \overset{(4.11)}{=} f \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) + g_i(x_k^i) - g_i(x_{i_1} ^{k_1}) + \frac{1}{\lambda_k} \left( \nabla h_i(x_{k+1}^{i_1}) - \nabla h_i(x_k^i) \right) \cdot x_{i_1} ^{k_1} - x_k^i \\
  & \quad + L_i D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \\
  & = f \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) + g_i(x_k^i) - g_i(x_{i_1} ^{k_1}) - \frac{1}{\lambda_k} \left( D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) + D_{h_i} \left( x_k^i, x_{i_1} ^{k_1} \right) \right) \\
  & \quad + L_i D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \\
  & \leq f \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) + g_i(x_k^i) - g_i(x_{i_1} ^{k_1}) - \left( 1 + \frac{a(h_i)}{\lambda_k} - L_i \right) D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right), \quad (4.12)
\end{align*}
\]
where the third equation is obtained through three points identity and the last inequality is obtained by the definition of symmetry coefficient. (4.12) indicates that
\[
F(x^{k+1}) \leq F(x^k) - \sum_{i=1}^{m} \left[ \left( \frac{1 + a(h_i)}{\lambda_k} - L_i \right) D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \right].
\]
If we choose
\[
\lambda_k \in \left( 0, \min \left\{ \frac{1 + a(h_i)}{L_i} \right\} \right),
\]
whose range seems to be more wide than the non-convex case.

In summary, we obtain the decrease of the sequence \( \{ F(x^k) \} \), i.e., (4.6). Further we prove the main results in this theorem, i.e.,

1. Fix \( k \geq 1 \). Under our assumptions and using (4.10), we can obtain
\[
F(x^k) - F(x^{k-1}) \leq - \sum_{i=1}^{m} \left[ \left( \frac{1}{\lambda_{k-1}} - L_i \right) D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \right].
\]
Further with \( \frac{1}{\lambda_{k-1}} > \max \{ L_i \} \geq L_i \), we immediately obtain the the sequence \( \{ F(x^k) \} \) is non-increasing.

2. Let \( n \) be a positive integer. Summing the above inequality from \( k = 1 \) to \( n \), we obtain
\[
\sum_{k=1}^{n} \left\{ \min \left\{ \frac{1}{\lambda_{k-1}} - L_i \right\} \sum_{i=1}^{m} D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \right\} \leq \sum_{k=1}^{n} \left\{ \sum_{i=1}^{m} \left[ \left( \frac{1}{\lambda_{k-1}} - L_i \right) D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \right] \right\} \leq F(x^0) - F(x^n) \leq F(x^0) - \nu(P) \leq \infty,
\]
and because there exists \( C > 0 \) such that \( C \leq \min_{1 \leq k \leq n, 1 \leq i \leq m} \left\{ \frac{1}{\lambda_{k-1}} - L_i \right\} \), we guarantee the first desired assertion. Further we deduce that the sequence \( \{ \sum_{i=1}^{m} D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \} \) converges to zero.

3. From the above inequality we also obtain
\[
n \min_{1 \leq k \leq n} \left[ \sum_{i=1}^{m} D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \right] \leq \sum_{k=1}^{n} \left[ \sum_{i=1}^{m} D_{h_i} \left( x_{k+1}^{i_1}, x_{i_1} ^{k_1} \right) \right] \leq \frac{1}{C} \left( F(x^0) - \nu(P) \right).
\]
After division by \( n \), we obtain the desired result.
4. Recall the optimality condition of the subproblem in each iteration, we have

\[ 0 \in \partial g_i(x^{k+1}_i) + \nabla f(x^{k+1}, x^{k}_i) + \frac{1}{\lambda_k} \left[ \nabla h_i(x^{k+1}_i) - \nabla h_i(x^{k}_i) \right]. \]

Then combining the above equation from \( i = 1 \) to \( i = m \), and by [29, Proposition 10.5] we obtain

\[ 0 \in \partial g(x^{k+1}) + \nabla f(x^{k+1}) + \begin{bmatrix} \nabla f(x^k) - \nabla f(x^{k+1}) \\ \vdots \\ \nabla f(x^{k+1}_{<i}, x^{k}_{\geq i}) - \nabla f(x^{k+1}) \\ \vdots \\ \nabla f(x^{k+1}_{<m}, x^{k}_{m}) - \nabla f(x^{k+1}) \end{bmatrix} + \begin{bmatrix} \nabla h_1(x^{k+1}_1) - \nabla h_1(x^{k}_1) \\ \vdots \\ \nabla h_1(x^{k+1}_1) - \nabla h_i(x^{k}_i) \\ \vdots \\ \nabla h_m(x^{k+1}_m) - \nabla h_m(x^{k}_m) \end{bmatrix}. \]

We know that \( \left\{ \sum_{i=1}^{m} D_{h_i} \left( (x^{k}_i, x^{k-1}_i) \right) \right\} \) converges to zero, and \( h_i \) is strictly convex, as a result \( \|x^k - x^{k-1}\| \to 0 \). With the continuity of \( \nabla f \) and \( \nabla h_i \), we know that

\[ \nabla f(x^{k+1}_{<i}, x^{k}_{\geq i}) - \nabla f(x^{k+1}) \to 0, \quad \nabla f(x^{k+1}_{<i}, x^{k}_{\geq i}) - \nabla f(x^{k+1}) \to 0. \]

Let \( \bar{x} \) be a limit point of \( \{x^k\} \), i.e., the sub-sequence \( \{x^k_i\} \) converges to \( \bar{x} \), then by the outer semi-continuity of \( \partial g \) we will have

\[ 0 \in \partial g(\bar{x}) + \nabla f(\bar{x}), \]

which indicates that \( \bar{x} \in X^* \).

All these four items are guaranteed. \( \square \)

Remark 4.4. If \( g_i \) is further assumed to be convex (but \( f \) is still nonconvex), then \( \lambda_k \) can be chosen in the interval \( \left( 0, \min \left\{ \frac{1+g(h_i)}{L_i} \right\} \right) \), which is larger than \( \left( 0, \frac{1}{\max \{L_i\}} \right) \) for the non-convex case. If \( f \) is additionally assumed to be convex too, then Item (4) in Theorem 4.3 can be improved and the sequence \( \{x^k\} \) is guaranteed to converge to a global minimizer of (1.1).

5 Improved convergence results

Theorem 4.3 shows that when the proposed Algorithm 1 is applied to solve the nonconvex and non-global Lipschitz continuity case of the problem (1.1), any accumulation point of the sequence is a stationary point in sense of the limiting subdifferential. This is not surprised from the variational analysis perspective. In this section, we show that the stationarity of accumulation points can be enhanced if \( \nabla f \) is globally Lipschitz continuous.

5.1 Strong stationarity

We first show that the stationarity in sense of the limiting subdifferential established in Theorem 4.3 can be improved. Indeed, according to Remark 2.1 in [18], the stationary point in sense of the limiting subdifferential may not exclude the existence of descent direction and thus it may be spurious. Let us recall an interesting example in [18]:

\[ \min_{x=(x_1, x_2)} \left( x_1 - 1 \right)^2 + x_2^3 + x_3^2 + I_S(x), \quad f(x) \text{ and } g(x). \]
where $S := \{ x \mid \| x \|_0 \leq 1, x \geq 0 \}$. Obviously, $f$ is nonconvex, differentiable and its gradient is globally Lipschitz continuous. For this example, it is trivial to see that the origin $(0, 0)$ is a stationary point in sense of the limiting subdifferential, while there exists a trivial descent direction, $(1, 0)$, that reduces the objective by increasing $x_1$. Hence, $(0, 0)$ is obviously not a local minimizer and it becomes interesting to ask if the stationarity in sense of limiting subdifferential in Theorem 4.3 can be further improved. We can verify that if the proposed Algorithm 1 is implemented with $(0, 0)$ as the initial iterate, then the sequence $\{ x^k \}$ indeed converges to $(1, 0)$, which is a local minimizer of this example. Hence, it is possible to consider stronger stationarity than the limiting subdifferential for Algorithm 1, and as we shall show, for some applications we can even claim the local optimality for the stationary point. It seems there are very few literatures along this line of research.

For simplicity, in this subsection let us just focus on the case where $h_i(x_i) = \frac{1}{2} \| x_i \|^2$, i.e., $f$ has a per-block Lipschitz continuous gradient. The following theorem shows some convergence result of Algorithm 1 with strong stationarity.

**Theorem 5.1.** Let $\{ x^k \}$ be the sequence generated by Algorithm 1 and $\bar{x}$ an limiting point of $\{ x^k \}$. Then $\bar{x}$ is a strongly stationary point, i.e.,

$$0 \in \nabla f(\bar{x}) + \partial^f g(\bar{x}).$$

**Proof.** For the sequence $\{ x^k \}$ generated by Algorithm 1 with $h_i(x_i) = \frac{1}{2} \| x_i \|^2$, we have

$$x^{k+1} = \arg \min_u \left\{ g(u) + \frac{1}{2\lambda_k} \| u - \bar{z}^k \|^2 \right\},$$

where $\bar{z}^k := x^k - \lambda_k \nabla f(x^k, x^{\geq i})$. Without loss of generality, we assume that $\lambda_k$ converges to $\lambda > 0$. According to Theorem 4.3, the sequence $\{ x^{k+1} - x^k \}$ converges to 0. Thus for any limiting point $\bar{x}$ of $\{ x^k \}$, suppose that $\{ x^i \}$ be the subsequence of $\{ x^k \}$ converging to $\bar{x}$, there must hold that $x^{i+1} \rightarrow \bar{x}$, and $\bar{z}^i \rightarrow \bar{x} - \lambda \nabla f(\bar{x})$ by the continuity of $\nabla f$. By assumption, we know that $g$ is prox-bounded (see Definition 1.23 in [29]). Therefore, according to Theorem 1.25 in [29], we have

$$\bar{x} = \arg \min_u \left\{ g(u) + \frac{1}{2\lambda} \| u - \bar{x} + \lambda \nabla f(\bar{x}) \|^2 \right\}.$$

Then, it follows immediately that

$$0 \in \nabla f(\bar{x}) + \partial^f g(\bar{x}),$$

i.e., $\bar{x}$ is a strongly stationary point. \qed

**Remark 5.2.** According to the inclusion (3.1), the strong stationarity established in Theorem 5.1 is better than the one in terms of limiting subdifferential in Theorem 4.3. Hence, the novelty of our result is revealed by comparing these two theorems. This improvement of the strong stationarity is achieved only with the additional assumption of the globally Lipschitz continuity of $\nabla f$. As shown in the proof of Theorem 5.1, specific features of Algorithm 1 are fully considered in the analysis.

**Remark 5.3.** It is worth mentioning that our analysis can be generalized to the convergence analysis for various other first-order methods. For instance, in [6], the authors discussed the proximal alternating linearized minimization (PALM) method for nonconvex and nonsmooth problems and they proved the global convergence of the generated sequence under the KL-property of the objective function. Following our just-presented analysis, we can improve the stationarity for the PALM in [6] as well. The details are omitted for succinctness.
5.2 Local optimality

The strong stationarity in Theorem 5.1 inspires us to further consider the local optimality, despite the nonconvexity of the model (1.1). This target is generally impossible. But, we show that it is doable for some special case such as when the function $g$ in (1.1) is the indicator function of a set $\Omega$, i.e., the problem (1.1) reduces to

$$\min_{x \in \Omega} f(x). \quad (5.2)$$

Let $\{x^i\}$ be the sequence generated by Algorithm 1 for (5.2). By virtue of Theorem 5.1, each limiting point $\bar{x}$ of $\{x^i\}$ satisfies

$$-\nabla f(\bar{x}) \in \tilde{N}_\Omega(\bar{x}) = T^o_\Omega(\bar{x}),$$

and thus

$$\langle \nabla f(\bar{x}), d \rangle \geq 0, \quad \forall d \in T_\Omega(\bar{x}).$$

The following corollary then demonstrates the local optimality of $\bar{x}$.

**Corollary 5.4.** Let $g_i(x_i)$ denote the indicator function of a possibly nonconvex set $X_i$, i.e.,

$$g_i(x_i) = \mathbb{1}_{X_i}(x_i);$$

$f$ be convex; and $\Omega = X_1 \times \cdots \times X_m$. If there exists $\delta > 0$ such that

$$[\Omega - \{\bar{x}\}] \cap B(0, \delta) \subset T_\Omega(\bar{x}), \quad (5.3)$$

where $B(0, \delta)$ denotes a ball centered at 0 with radius $\delta$ and $T_\Omega(\bar{x})$ denotes the tangent cone of $\Omega$ at $\bar{x}$, then $\bar{x}$ is a local minimizer of the problem (5.2).

**Proof.** For every $x \in B(\bar{x}, \delta) \cap \Omega$, by (5.3), we have $x - \bar{x} \in [\Omega - \{\bar{x}\}] \cap B(0, \delta) \subset T_\Omega(\bar{x})$. Then, according to Theorem 5.1, it follows that $-\nabla f(\bar{x}) \in T^o_\Omega(\bar{x})$, and thus $\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0$. Combining it with the convexity of $f$, we have

$$f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq f(\bar{x}),$$

and therefore $\bar{x}$ is a local minimizer. \qed

The next proposition reveals that the assumption (5.3) is not restrictive.

**Proposition 5.5.** If $\Omega$ is the union of finitely many convex sets, the assumption (5.3) in Corollary 5.4 is satisfied.

**Proof.** For $\Omega := \bigcup_{i \in I} \Omega_i$ where $I = \{1, \ldots, \ell\}$. Setting $I(\bar{x}) := \{i \in I \mid \bar{x} \in \Omega_i\}$, then for all $i \notin I(\bar{x})$, we have $\bar{x} \notin \Omega_i$. For each $i \notin I(\bar{x})$, since $\Omega_i$ is closed, then there exists $\delta_i > 0$ such that for all $x \in B(\bar{x}, \delta_i), x \notin \Omega_i$. Let $\delta := \min_{i \notin I(\bar{x})} \{\delta_i\}$, we have $x \notin \bigcup_{i \notin I(\bar{x})} \Omega_i$ for any $x \in B(\bar{x}, \delta)$. And since

$$\Omega = \left( \bigcup_{i \in I(\bar{x})} \Omega_i \right) \cup \left( \bigcup_{i \notin I(\bar{x})} \Omega_i \right),$$

then for any $x \in \Omega \cap B(\bar{x}, \delta)$, it holds $x \in \bigcup_{i \in I(\bar{x})} \Omega_i$. Given any $d \in [\Omega - \{\bar{x}\}] \cap B(0, \delta)$, we have $\bar{x} + d \in \Omega \cap B(\bar{x}, \delta)$ and thus $x + d \in \bigcup_{i \in I(\bar{x})} \Omega_i$. And therefore, there exists $i_d \in I(\bar{x})$ such that $x + d \in \Omega_{i_d}$. Then by the convexity of $\Omega_{i_d}$, the closed interval $[\bar{x}, \bar{x} + d]$ belongs to $\Omega_{i_d} \subseteq \Omega$, which is equivalent to for all $t \in [0, 1]$, $\bar{x} + td \in \Omega$. Hence $d \in T_\Omega(\bar{x})$. Thus we can conclude that $[\Omega - \{\bar{x}\}] \cap B(0, \delta) \subset T_\Omega(\bar{x})$. \qed
5.3 Application

In this subsection, we consider the linear regression model with cardinality constraints and show that all the assumptions in Corollary 5.4 are satisfied. The model is

\[
\min_x \frac{1}{2} \left\| Ax - b \right\|^2 + \underbrace{f(x)}_{g(x)} + \iota_C(x),
\]

where \( \iota \) is the indicator function of the set \( C := \{ x \in \mathbb{R}^n : \| x \|_0 \leq \ell \} \) with \( \ell \) a positive integer smaller than \( n \). This model is fundamental in various areas such as statistical learning, compressive sensing, sparse regression, sparse controller design. Note that \( f \) in (5.4) is convex, differentiable with a globally Lipschitz gradient, but \( g \) is nonconvex (see, e.g.,[2, 4]).

According to [19], the objective function in (5.4) satisfies the KL property and thus, due to [6], the generated sequence converges to a stationary point in sense of the limiting subdifferential. Let the index of nonzero entries of \( x \) be \( \mathcal{I}(x) := \{ i \in \{1, \ldots, n\} : x_i \neq 0 \} \). Then we have \( \text{card}(\mathcal{I}(x)) = \| x \|_0 \).

Let \( J_r := \{ J \in 2^{\{1, \ldots, n\}} : \text{card}(J) = \ell \} \) denote the subset of \( \{1, \ldots, m\} \) with the cardinality of \( r \). It is easy to see that the cardinality cone can be represented as the union of finitely many subspaces:

\[
\mathcal{C} = \bigcup_{J \in J_r} \mathcal{C}_J,
\]

where \( \mathcal{C}_J := \text{span}\{ e_j : j \in J \} \), and \( e_j \) is the unit vector in \( \mathbb{R}^n \) with the \( j \)-th component being 1 and others being 0. Hence, all assumptions in Corollary 5.4 are satisfied by the model (5.4). This means the sequence generated by Algorithm 1 is guaranteed to converge to not just a stationary point, but a local minimizer of the problem (5.4).

6 Numerical Experiments

We report some preliminary numerical results to verify the efficiency of the proposed Algorithm 1. It is worth mentioning that this work is in theoretical nature, and its focus is the convergence analysis. These numerical results are in auxiliary nature and thus no comparison with other methods is provided. All the numerical experiments are implemented in MATLAB R2015b on a Dell Laptop with 2.60GHz CPU and 8.00GB memory.

For succinctness, we just test the Poisson linear inverse problem with \( \ell_1 \) regularization. Given a matrix \( A \in \mathbb{R}^{m \times n} \) and \( b \in \mathbb{R}_+^n \), the Poisson linear inverse problem [9] is to reconstruct a signal or image \( x \in \mathbb{R}_+^n \) from a noisy measurements \( b \) such that \( Ax \simeq b \). A standard formulation for this problem is measuring the proximity of two nonnegative vectors \( Ax \) and \( b \) with the KL divergence, i.e.,

\[
\min_{x \in \mathbb{R}_+^n} \sum_{i=1}^n \left[ \frac{b_i}{(Ax)_i} \log \left( \frac{b_i}{(Ax)_i} \right) - b_i + (Ax)_i \right] + \mu \| x \|_1,
\]

where the \( \ell_1 \) term \( \| x \|_1 \) is for reflecting sparsity of the solution. This is obviously a convex minimization problem; \( f \) is differentiable but its gradient is not globally Lipschitz continuous. Thus, if some typical BCD type methods are implemented, there is no convergence guarantee.

We test the typical image de-blurring problem via the model (6.1). Given a real image \( x \in \mathbb{R}_+^n \) (we test the classical “Lena” image with \( n = 65536 \), and “Cameraman” image with \( n = 262144 \)),
the matrix $A \in \mathbb{R}^{n \times n}_+$ is set to be the blurring matrix associated with a spatially invariant point spread function, and the vector $b = Ax + e \in \mathbb{R}_+$ where $e \in \mathbb{R}^n_+$ denotes a Poisson noise vector. We have $m = 256$ for “Lena” image and $m = 512$ for “Cameraman” image, and the variable $x$ is separated as 256 and 512 blocks, respectively. Note that each block of variables has the dimension of 256 ($n_i = 256$) and 512 ($n_i = 512$), respectively. For $h_i$, as in [3], we choose the Burg’s entropy, i.e.,

$$h_i(x_i) = -\sum_{j=1}^{n_i} \log(x_i)_j, \quad \text{dom } h_i = \mathbb{R}_{++}^{n_i}.$$

The pbL condition of $(f, h_i)$ can be guaranteed with $L_i \geq \|b_i\|_1$ where $b_i$ denotes the $i$-th block of $b$ (The proof detail is included in the supplemental materials). According to [3], $\alpha(h_i) = 0$ and hence we choose $\lambda_k = \frac{0.99}{\max\{\|b_i\|_1\}} \in (0, \frac{1}{\max\{L_i\}})$.

For this application, the subproblem (4.2) is specified as

$$\min_{u \geq 0} \left\{ \mu u + \langle \nabla_i f(x_i, z_{\neq i}), u \rangle + \frac{1}{\lambda_k} \sum_{j=1}^{n_i} \left( \frac{u_j(x_i)}{(x_i)_j} - \log \frac{u_j(x_i)}{(x_i)_j} \right) \right\},$$

and thus for $j = 1, \ldots, n_i$, we have

$$\left[T^h_{\lambda_k}(x_i, z_{\neq i})\right]_j = \frac{(x_i)_j}{1 + \lambda_k (x_i)_j + \lambda_k \left(\nabla_i f(x_i, z_{\neq i})\right)_j (x_i)_j}.$$

We set $\mu = 0.1$ in our experiments.

We plot the numerical results of the proposed Algorithm 1 in Figure 1. The “Objective Value” denotes the objective function value $F(x^k)$ of the generated sequence, while the “Relative Error” denotes the relative error of the sequence $\{x^k\}$, i.e., $\|x^{k+1} - x^k\|/\|x^k\|$. These two plots in the left-hand-side are the results for “Lena” image and those in the right-hand-side for “Cameraman” image. These curves clearly demonstrate the convergence of the proposed algorithm; thus our theoretical analysis is numerically verified.

7 Conclusion and Future Work

This work focuses on the convergence analysis for a Bregman-distance-based block coordinate proximal gradient (BCPG) method for nonconvex composite minimization problems. We show its convergence without the globally Lipschitz continuity of the gradient of the differentiable function in the objective, and enhance the stationarity of the stationary point in terms of Fréchet (regular) subdifferential with the globally Lipschitz continuity. Our techniques can be immediately used to improve the convergence results for other popular methods such as the proximal alternating linearized minimization method in [6]. These results are new in the BCD literature, and may inspire more deeper results in the area. Meanwhile, we empirically observe the linear convergence rates for both the Poisson linear inverse problem with $\ell_1$ regularization (6.1) and the linear regression model with cardinality constraints (5.4). Existing results, including those based on the KL-property, seem not able to explain these linear convergence performances. This fact motivates us to further consider some theoretical conditions that can both be satisfied by some applications such as (5.4) and (6.1) and ensure the linear convergence of some BCD type methods, or more general first-order methods. We expect that some error bound conditions tailored for the specific structures of the algorithm under discussion may be helpful. We leave this topic for future research.
Figure 1: Convergence result for Poisson linear inverse problem with $\ell_1$ regularization

References


A phL Condition of \((f, h_i)\) in Poisson Linear Inverse Problem

\[
f(x) = \sum_{j=1}^{n} b_j \log \frac{b_j}{Ax_j} - b_j + (Ax)_j
\]

If \(L_i \geq \|b_i\|_1\) where \(b_i\) denotes the \(i\)-th block of \(b\), then we can guarantee the pbL condition of \((f, h_i)\).

Proof. Since \(f\) and \(h\) are both twice differentiable on \(\mathbb{R}^n_{++}\), the convexity of \(L_i h_i - f(\cdot, z_{\neq i})\) on \(\mathbb{R}^n_{++}\) is equivalent to

\[
L \left( \nabla^2 h_i(x_i)d_i, d_i \right) - \left( \nabla^2 f(x_i, z_{\neq i})d_i, d_i \right) \geq 0, \forall x_i \in \mathbb{R}^n_{++}, \forall d_i \in \mathbb{R}^n.
\]

With detailed computation of \(\nabla^2 h_i\), the above inequality implies that

\[
L \left( \nabla^2 h_i(x_i)d_i, d_i \right) = L \sum_{j=1}^{n_i} \frac{(d_i)_j^2}{(A)_j x_i + \sum_{\ell=1, \ell \neq i}^{m} (A)_\ell z_{\ell}} \left( (A)_i \right)_j,
\]

\[
\left( \nabla^2 f(x_i, z_{\neq i})d_i, d_i \right) = \sum_{j=1}^{m} (b_i)_j \left( \frac{(d_i)_j}{(A)_j x_i + \sum_{\ell=1, \ell \neq i}^{m} (A)_\ell z_{\ell}} \right)^2.
\]

By a simple application of Jensen’s inequality to the nonnegative convex function \(t^2\), it follows that for any \(u \in \mathbb{R}^n_+\) and \(x_i \in \mathbb{R}^n_{++}\),

\[
\frac{\langle u, d \rangle^2}{\langle u, x_i \rangle^2} \leq \sum_j \frac{u_j (x_i)_j}{\langle u, x_i \rangle} \left( \frac{(d_i)_j}{(x_i)_j} \right)^2 \leq \sum_{j=1}^{n_i} \frac{(d_i)_j^2}{(x_i)_j} \forall d_i \in \mathbb{R}^n.
\]

Further since every element of \(A\) is positive, we have

\[
\left( \nabla^2 f(x_i, z_{\neq i})d_i, d_i \right) \leq \sum_{j=1}^{m} (b_i)_j \left( \frac{(d_i)_j}{(A)_j x_i + \sum_{\ell=1, \ell \neq i}^{m} (A)_\ell z_{\ell}} \right)^2 \leq \sum_{j=1}^{m} (b_i)_j \sum_{j=1}^{n_i} \frac{(d_i)_j^2}{(x_i)_j^2} \forall x_i \in \mathbb{R}^n_{++}.
\]

As a result, if \(L_i \geq \left( \sum_{j=1}^{m} (b_i)_j \right) = \|b_i\|_1\), we guarantee the pbL condition of \((f, h_i)\). \(\square\)