SIAM J. OPTIM.
Vol. 23, No. 4, pp. 2037–2060


# ON THE CONVERGENCE OF BLOCK COORDINATE DESCENT TYPE METHODS*

AMIR BECK† AND LUBA TETRUASHVILI†

**Abstract.** In this paper we study smooth convex programming problems where the decision variables vector is split into several blocks of variables. We analyze the block coordinate gradient projection method in which each iteration consists of performing a gradient projection step with respect to a certain block taken in a cyclic order. Global sublinear rate of convergence of this method is established and it is shown that it can be accelerated when the problem is unconstrained. In the unconstrained setting we also prove a sublinear rate of convergence result for the so-called alternating minimization method when the number of blocks is two. When the objective function is also assumed to be strongly convex, linear rate of convergence is established.

**Key words.** block descent methods, alternating minimization, rate of convergence, convex optimization

**AMS subject classifications.** 90C06, 90C25, 65K05

**DOI.** 10.1137/120887679

## 1. Introduction.

One of the first variable decomposition methods for solving general minimization problems is the so-called alternating minimization method [5, 14], which is based on successive global minimization with respect to each component vector in a cyclic order. This fundamental method appears in the literature under various names such as the block-nonlinear Gauss–Seidel method or the block coordinate descent method (see, e.g., [4]). The convergence of the method was extensively studied in the literature under various assumptions. For example, Auslender studied in [1] the convergence of the method under a strong convexity assumption, but without assuming differentiability. In [4] Bertsekas showed that if the minimum with respect to each block of variables is unique, then any accumulation point of the sequence generated by the method is also a stationary point. Grippo and Sciandrone showed in [7] convergence results of the sequence generated by the method under different sets of assumptions such as strict quasi convexity with respect to each block. Luo and Tseng proved in [9] that under the assumptions of strong convexity with respect to each block, existence of a local error bound of the objective function, and proper separation of isocost surfaces, linear rate of convergence can be established.

Another closely related method, which will be the main focus of this paper, is the block coordinate gradient projection (BCGP) method in which at each subiteration, the exact minimization with respect to a certain block of variables is replaced with an employment of a single step of the gradient projection method (a step toward the gradient followed by an orthogonal projection). This method has a clear advantage over alternating minimization when exact minimization with respect to each of the

*Received by the editors August 9, 2012; accepted for publication (in revised form) August 12, 2013; published electronically October 16, 2013.
http://www.siam.org/journals/siopt/23-4/88767.html
†Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 3200, Israel (becka@ie.technion.ac.il, lubate@tx.technion.ac.il). The research of Amir Beck was partially supported by the Israel Science Foundation under grant ISF 253/12.

component blocks is not an easy task. In [8] Luo and Tseng studied minimization problems of the form

$$\min_{\mathbf{x}}\{g(\mathbf{E}\mathbf{x}) + \langle \mathbf{b}, \mathbf{x}\rangle : \mathbf{x} \in X\},$$

where, loosely speaking, the main assumptions were that $g$ is strictly convex and twice continuously differentiable over the domain of $g$ and $\nabla^2 g(\mathbf{E}\mathbf{x}^*)$ is positive definite for every optimal solution $\mathbf{x}^*$. Under these assumptions, it was shown that the BCGP method with each block consisting of a single variable has a linear rate of convergence.

Recently, Nesterov studied in [13] a randomized version of the method in the convex setting in which the selection of the block on which a gradient projection step is performed is not done by a deterministic rule (such as the cyclic rule), but rather via a predescribed distribution. For the first time, Nesterov was able to establish global nonasymptotic rates of convergence in the convex case without any strict convexity, strong convexity, uniqueness, or error bound assumptions. Specifically, it was shown that the rate of convergence of the expectation sequence of the randomized method is sublinear under the assumption of Lipschitz continuity of the gradient and linear under a strong convexity assumption. In addition, an accelerated $O(1/k^2)$ was devised in the unconstrained setting. Probabilistic results on the convergence of the function values were also provided.

Recently, there has been a wide interest in randomized methods. In [17] Richtarik and Takac generalized Nesterov's results to the case when a separable nonsmooth term is added. The authors also demonstrated in [16] the ability of various block coordinate decent methods to solve large-scale truss topology design problems. In [19] Shalev-Shwartz and Tewari considered the stochastic coordinate descent method applied to the $l_1$-regularized loss minimization problem. It was shown in [19] that the number of iterations required to obtain an expected $\varepsilon$-accurate solution is $O(1/\varepsilon)$.

Despite the apparent large amount of results in the stochastic case, there are only few results on global rate of convergence in the deterministic case. Under an assumption on the isotonicity of the gradient of $f$, Saha and Tewari were able to prove in [18] an $O(1/k)$ rate of convergence of the sequence of function values of the cyclic coordinate descent and the cyclic coordinate minimization methods. The complexity of a greedy approach was studied by Dhillon, Ravikumar, and Tewari in [6]. However, as was pointed out by Nesterov in [13], it seems that there are no global rate of convergence results for the cyclic block coordinate gradient descent (BCGD) method under general convexity assumptions in the deterministic case. The main goal of this paper is to rectify this situation and establish several new results on the rate of convergence of the BCGP method and other related schemes for convex programming problems.

In section 2 we lay out the basic setting in the unconstrained case and present the BCGD method in which each iteration consists of making a gradient step with respect to each block in a cyclic order. The convergence analysis of the method is given in section 3, where sublinear $O(1/k)$ rate of convergence ($k$ being the iteration index) is established under the assumption that the objective function is convex with Lipschitz continuous gradient. We show by a numerical example that even though the multiplicative constant obtained in the BCGD method is worse than the one obtained by the randomized approach in [13], the deterministic BCGD method can even have a better empirical performance than the randomized counterpart. When in addition the objective function is strongly convex, the linear rate of convergence is proved. We then show in section 4 that we can incorporate the BCGD into an optimal scheme

resulting in an accelerated method with an $O(1/k^2)$ rate of convergence. In section 5 we make a detour to discuss the alternating minimization method in the case where the variables are split into two blocks. We establish a sublinear rate of convergence and prove a linear rate of convergence when an additional strong convexity assumption is made. Finally, we discuss the constrained problem where each of the block vectors is constrained to be in a given convex closed set. In this case each gradient step is followed by an orthogonal projection operator and the method is consequently called the *block coordinate gradient projection method*. We show that a sublinear rate of convergence can be established in this case too.

Throughout the paper we use the following notation. Vectors are denoted by boldface lowercase letters, e.g., $\mathbf{y}$, and matrices by boldface uppercase letters e.g., $\mathbf{A}$. The $i$th component of a vector $\mathbf{y}$ is written as $y_i$. The identity matrix of order $n$ is denoted by $\mathbf{I}_n$. In this paper, all norms are the usual $l_2$ norms.

**2. Problem formulation and basic setting.** Consider the general optimization model

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}, \tag{2.1}$$

where we assume the following basic assumptions:

- $f$ is a continuously differentiable convex function whose gradient is Lipschitz over $\mathbb{R}^n$.
- The optimal set of (2.1) is nonempty and is denoted by $X^*$ and the corresponding optimal value is denoted by $f^*$.

We will assume that the vector $\mathbf{x}$ of decision variables has the following partition:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}(1) \\ \mathbf{x}(2) \\ \vdots \\ \mathbf{x}(p) \end{pmatrix}, \tag{2.2}$$

where $\mathbf{x}(i) \in \mathbb{R}^{n_i}$ with $n_1, n_2, \ldots, n_p$ being $p$ positive integer numbers satisfying $n_1 + n_2 + \cdots + n_p = n$. We use the notation of [13] and define the matrices $\mathbf{U}_i \in \mathbb{R}^{n \times n_i}$, $i = 1, \ldots, p$, for which

$$(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p) = \mathbf{I}_n.$$

Then in our notation $\mathbf{x}(i) = \mathbf{U}_i^T \mathbf{x}$ for every $\mathbf{x} \in \mathbb{R}^n$ and $i = 1, \ldots, p$, and we also have that if $\mathbf{x}$ is given as in (2.2), then $\mathbf{x} = \sum_{i=1}^p \mathbf{U}_i \mathbf{x}(i)$. In addition, we will also define the vector of partial derivatives corresponding to the variables in the vector $\mathbf{x}(i)$ as

$$\nabla_i f(\mathbf{x}) \equiv \mathbf{U}_i^T \nabla f(\mathbf{x}), \quad i = 1, 2, \ldots, p.$$

We assume that the gradient of $f$ is block-coordinatewise Lipschitz continuous and that the Lipschitz constant corresponding to block $i$ is $L_i$, meaning that

$$\|\nabla_i f(\mathbf{x} + \mathbf{U}_i \mathbf{h}_i) - \nabla_i f(\mathbf{x})\| \le L_i \|\mathbf{h}_i\| \quad \text{for every } \mathbf{h}_i \in \mathbb{R}^{n_i}. \tag{2.3}$$

The constants $L_1, L_2, \ldots, L_p$ will also be called *the block Lipschitz constants*. The gradient of $f$, $\nabla f$, is of course also Lipschitz continuous, and we denote its "global" Lipschitz constant by $L$. That is,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L \|\mathbf{x} - \mathbf{y}\| \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

We are now ready to define the BCGD method. In this method, at each iteration we perform a gradient step with a constant stepsize with respect to a different block of variables taken in a cyclic order.

**The BCGD method.**

**Input:** $\bar{L}_j (j = 1, \ldots, p)$—upper bounds on the Lipschitz constant $L_j$ ($\bar{L}_j \geq L_j$).

**Initialization:** $\mathbf{x}_0 \in \mathbb{R}^n$.
**General step (k=0,1,...):** Set $\mathbf{x}_k^0 = \mathbf{x}_k$ and define recursively

$$\mathbf{x}_k^i = \mathbf{x}_k^{i-1} - \frac{1}{\bar{L}_i} \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), i = 1, \ldots, p. \tag{2.4}$$

Set $\mathbf{x}_{k+1} = \mathbf{x}_k^p$.

We will be especially interested in two constant stepsize strategies:

- Exact constant stepsize rule. In this setting

$$\bar{L}_j = L_j, \quad j = 1, \ldots, p.$$

- Conservative constant stepsize rule. Here the upper estimates on the Lipschitz constants are all chosen to be equal to the global Lipschitz constant, that is,

$$\bar{L}_j = L, \quad j = 1, \ldots, p.$$

We will assume that the level set

$$S = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$$

is compact and we denote (similarly to [13])

$$R(\mathbf{x}_0) \equiv \max_{\mathbf{x} \in \mathbb{R}^n} \max_{\mathbf{x}^* \in X^*} \{\|\mathbf{x} - \mathbf{x}^*\| : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}.$$

In particular, by the monotonicity of $\{f(\mathbf{x}_k)\}_{k \geq 0}$,

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq R(\mathbf{x}_0) \text{ for every } k = 0, 1, \ldots.$$

**3. Convergence analysis of the BCGD method.** The efficiency estimates of the BCGD method depend on several constants which are now defined. The maximal and minimal block Lipschitz constants are denoted by

$$L_{\max} = \max_{j=1,\ldots,p} L_j, \tag{3.1}$$

$$L_{\min} = \min_{j=1,\ldots,p} L_j. \tag{3.2}$$

We use similar notation for the maximal and minimal upper estimates on the block Lipschitz constants:

$$\bar{L}_{\max} = \max_{j=1,\ldots,p} \bar{L}_j, \tag{3.3}$$

$$\bar{L}_{\min} = \min_{j=1,\ldots,p} \bar{L}_j. \tag{3.4}$$

It is known (see [13, Lemma 2]) that the Lipschitz constants $L_1, L_2, \ldots, L_p, L$ satisfy the relation

$$L \le \sum_{i=1}^{p} L_i,$$

which immediately implies that

$$L \le pL_{\max} \tag{3.5}$$

as well as

$$L \le p\bar{L}_{\max}. \tag{3.6}$$

In addition, the ratio

$$\kappa = \frac{L_{\max}}{L_{\min}}, \tag{3.7}$$

which will be called the *scalability factor*, will play an important role in the rate of convergence analysis. If the scalability factor is "large," then the problem is poorly scaled. We will also use quite frequently in our analysis the well-known descent lemma which is now recalled for the sake of completeness and whose proof can be found, for example, in [4].

LEMMA 3.1 (descent lemma [4, Proposition A.24]). *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function whose gradient $\nabla g$ is Lipschitz with constant $M$. Then*

$$g(\mathbf{y}) \le g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{M}{2}\|\mathbf{x} - \mathbf{y}\|^2 \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

As a direct consequence we have the following "block" version of the lemma.

LEMMA 3.2 (block descent lemma). *Suppose that $f$ is a continuously differentiable function over $\mathbb{R}^n$ satisfying (2.3) and assume that $i \in \{1, 2, \ldots, p\}$. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ be two vectors which differ only in the $i$th block, that is, there exists an $\mathbf{h} \in \mathbb{R}^{n_i}$ such that $\mathbf{v} - \mathbf{u} = \mathbf{U}_i\mathbf{h}$. Then*

$$f(\mathbf{v}) \le f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle + \frac{L_i}{2}\|\mathbf{u} - \mathbf{v}\|^2.$$

We will now show that subject to the underlying assumptions on the objective function, we can prove a sublinear rate of convergence of the sequence of function values.

**3.1. Sublinear rate of convergence.** The next technical lemma is essential in establishing the sublinear rate of convergence.

LEMMA 3.3. *Let $\{\mathbf{x}_k\}_{k\ge 0}$ be the sequence generated by the BCGD method. Then for every $k = 0, 1, 2, \ldots$*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \ge \frac{1}{4\bar{L}_{\max}\left(1 + pL^2/\bar{L}_{\min}^2\right)}\|\nabla f(\mathbf{x}_k)\|^2, \tag{3.8}$$

*where $\bar{L}_{\max}$ and $\bar{L}_{\min}$ are given in (3.3) and (3.4), respectively.*

*Proof.* By the block descent lemma (Lemma 3.2) and the fact that $\bar{L}_i \ge L_i, i = 1, \ldots, p$, we have that for all $i = 1, \ldots, p$

$$f(\mathbf{x}_k^i) \le f(\mathbf{x}_k^{i-1}) + \langle \nabla f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^i - \mathbf{x}_k^{i-1} \rangle + \frac{\bar{L}_i}{2}\|\mathbf{x}_k^i - \mathbf{x}_k^{i-1}\|^2. \tag{3.9}$$

Plugging the recursion relation (2.4) into (3.9) yields

$$f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}_k^i) \geq \frac{1}{2\bar{L}_i}\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2, \quad i = 1, 2, \ldots, p.$$

Summing over all the inequalities we get

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2}\sum_{i=1}^{p}\frac{1}{\bar{L}_i}\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2 \geq \frac{1}{2\max\limits_{j=1,\ldots,p}\{\bar{L}_j\}}\sum_{i=1}^{p}\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2. \tag{3.10}$$

By (2.4), and recalling that $\mathbf{x}_k^0 = \mathbf{x}_k$, it follows that for every $i = 0, \ldots, p$

$$\mathbf{x}_k = \mathbf{x}_k^i + \sum_{j=1}^{i}\frac{1}{\bar{L}_j}\mathbf{U}_j\nabla_j f(\mathbf{x}_k^{j-1}),$$

and therefore, for every $i = 1, 2, \ldots, p$,

$$\begin{aligned}\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_k^i)\|^2 &\leq L^2\|\mathbf{x}_k - \mathbf{x}_k^i\|^2 \\ &= L^2\left\|\sum_{j=1}^{i}\frac{1}{\bar{L}_j}\mathbf{U}_j\nabla_j f(\mathbf{x}_k^{j-1})\right\|^2 \\ &\leq \frac{L^2}{\min\limits_{j=1,\ldots,p}\{\bar{L}_j^2\}}\sum_{j=1}^{i}\|\nabla_j f(\mathbf{x}_k^{j-1})\|^2.\end{aligned}$$

Thus, for every $i = 1, \ldots, p$,

$$\begin{aligned}\|\nabla_i f(\mathbf{x}_k)\|^2 &\leq (\|\nabla_i f(\mathbf{x}_k) - \nabla_i f(\mathbf{x}_k^{i-1})\| + \|\nabla_i f(\mathbf{x}_k^{i-1})\|)^2 \\ &\leq 2\|\nabla_i f(\mathbf{x}_k) - \nabla_i f(\mathbf{x}_k^{i-1})\|^2 + 2\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2 \\ &\leq 2\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_k^{i-1})\|^2 + 2\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2 \\ &\leq 2\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2 + 2\frac{L^2}{\min\limits_{j=1,\ldots,p}\{\bar{L}_j^2\}}\sum_{j=1}^{i-1}\|\nabla_j f(\mathbf{x}_k^{j-1})\|^2.\end{aligned}$$

Summing over $i = 1, \ldots, p$ we obtain that

$$\begin{aligned}\sum_{i=1}^{p}\|\nabla_i f(\mathbf{x}_k)\|^2 &\leq 2\sum_{i=1}^{p}\left(1 + (p-i)\frac{L^2}{\min\limits_{j=1,\ldots,p}\{\bar{L}_j^2\}}\right)\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2 \\ &\leq 2\left(1 + p\frac{L^2}{\min\limits_{j=1,\ldots,p}\{\bar{L}_j^2\}}\right)\sum_{i=1}^{p}\|\nabla_i f(\mathbf{x}_k^{i-1})\|^2. \qquad (3.11)\end{aligned}$$

Combining (3.10) and (3.11), the desired result (3.8) follows. ∎

*Remark* 3.1. Note that a direct result of Lemma 3.3 is that the BCGD method generates a nonincreasing sequence of function values $\{f(\mathbf{x}_k)\}_{k\geq 0}$.

*Remark* 3.2. Note that the convexity assumption on $f$ was not used in the proof of Lemma 3.3 and is thus valid for any $C^{1,1}$ function.

An almost direct consequence of the latter lemma is the following key relation between consecutive objective function values.

LEMMA 3.4. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the BCGD method. Then*

(3.12)
$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{4\bar{L}_{\max}\left(1+pL^2/\bar{L}_{\min}^2\right)R^2(\mathbf{x}_0)}(f(\mathbf{x}_k)-f^*)^2, \quad k=0,1,\ldots,$$

*where $\bar{L}_{\max}$ and $\bar{L}_{\min}$ are given in* (3.3) *and* (3.4)*, respectively.*

*Proof.* By the convexity of $f$ we have for every $\mathbf{x}^* \in X^*$ that

$$f(\mathbf{x}_k) - f^* \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle, \quad k=0,1,\ldots,$$

and hence, by the Cauchy–Schwartz inequality, it follows that

$$f(\mathbf{x}_k) - f^* \leq \|\nabla f(\mathbf{x}_k)\| \cdot \|\mathbf{x}_k - \mathbf{x}^*\| \leq R(\mathbf{x}_0)\|\nabla f(\mathbf{x}_k)\|, \quad k=0,1,\ldots,$$

which combined with (3.8) implies (3.12). ∎

To derive the rate of convergence of the function values sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ to the optimal value $f^*$, we will use the following simple and well-known lemma on convergence of nonnegative scalar sequences.

LEMMA 3.5. *Let $\{A_k\}_{k\geq 0}$ be a nonnegative sequence of real numbers satisfying*

$$A_k - A_{k+1} \geq \gamma A_k^2, \quad k=0,1,\ldots,$$

*and*

$$A_0 \leq \frac{1}{m\gamma}$$

*for some positive $\gamma$ and $m$. Then*

(3.13)
$$A_k \leq \frac{1}{\gamma}\cdot\frac{1}{k+m}, \quad k=0,1,\ldots,$$

*and in particular*

$$A_k \leq \frac{1}{\gamma}\cdot\frac{1}{k}, \quad k=1,2,\ldots.$$

*Proof.* For every $k=1,2,\ldots$ we have

$$\frac{1}{A_k} - \frac{1}{A_{k-1}} = \frac{A_{k-1}-A_k}{A_{k-1}A_k} \geq \gamma\frac{A_{k-1}^2}{A_{k-1}A_k} = \gamma\frac{A_{k-1}}{A_k} \geq \gamma,$$

where the last inequality follows from the monotonicity of $\{A_k\}$. We thus conclude that

$$\frac{1}{A_k} \geq \frac{1}{A_0} + \gamma k \geq \gamma(k+m)$$

and hence (3.13) follows. ∎

Combining Lemmata 3.4 and 3.5 we obtain the following result on the sublinear rate of convergence of the objective function sequence generated by the BCGD method.

THEOREM 3.6 (sublinear rate of convergence of the BCGD method). *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the BCGD method. Then*

(3.14)
$$f(\mathbf{x}_k) - f^* \leq 4\bar{L}_{\max}(1+pL^2/\bar{L}_{\min}^2)R^2(\mathbf{x}_0)\frac{1}{k+(8/p)}, \quad k=0,1,\ldots,$$

*where $\bar{L}_{\max}$ and $\bar{L}_{\min}$ are given in* (3.3) *and* (3.4)*, respectively.*

*Proof.* By the descent lemma we have that

$$f(\mathbf{x}_0) - f^* \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \frac{L}{2}R^2(\mathbf{x}_0).$$

Therefore,

$$f(\mathbf{x}_0) - f^* \leq \frac{L}{2}R^2(\mathbf{x}_0) \overset{(*)}{\leq} \frac{p\bar{L}_{\max}}{2}R^2(\mathbf{x}_0) \leq \frac{p}{8}\left(4\bar{L}_{\max}(1+pL^2/\bar{L}_{\min}^2)R^2(\mathbf{x}_0)\right),$$

where the inequality (*) follows from (3.6). On the other hand, by (3.12) it follows that

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{4\bar{L}_{\max}\left(1+pL^2/\bar{L}_{\min}^2\right)R^2(\mathbf{x}_0)}(f(\mathbf{x}_k) - f^*)^2, \quad k = 0, 1, \ldots.$$

Invoking Lemma 3.5 with $\gamma = \frac{1}{4\bar{L}_{\max}(1+pL^2/\bar{L}_{\min}^2)R^2(\mathbf{x}_0)}$ and $m = 8/p$, the desired result (3.14) follows. ∎

We can deduce some immediate consequences from Theorem 3.6 for the exact and conservative constant stepsize rules.

COROLLARY 3.7 (convergence of BCGD with exact constant stepsize rule). *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the BCGD method with exact constant stepsize rule ($\bar{L}_j = L_j, j = 1, \ldots, p$). Then*

$$f(\mathbf{x}_k) - f^* \leq 4L_{\max}(1+p^3\kappa^2)R^2(\mathbf{x}_0)\frac{1}{k+(8/p)}, \quad k = 0, 1, \ldots. \tag{3.15}$$

*Proof.* In the exact constant stepsize setting $\bar{L}_{\max} = L_{\max}, \bar{L}_{\min} = L_{\min}$, which along with the inequality $L \leq pL_{\max}$ and (3.14) implies (3.15). ∎

COROLLARY 3.8 (convergence of BCGD with conservative constant stepsize rule). *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the BCGD method with a conservative constant stepsize rule ($\bar{L}_j = L, j = 1, \ldots, p$). Then*

$$f(\mathbf{x}_k) - f^* \leq \frac{4L(1+p)R^2(\mathbf{x}_0)}{k+(8/p)}, \quad k = 0, 1, \ldots. \tag{3.16}$$

*Proof.* The proof follows by simple substitution of the relations $\bar{L}_{\min} = \bar{L}_{\max} = L$ in (3.14). ∎

*Remark* 3.3. When $p = 1$, the BCGD method with conservative stepsize amounts to the standard gradient descent method with constant stepsize $\frac{1}{L}$. The complexity bound obtained in Theorem 3.6 is

$$f(\mathbf{x}_k) - f^* \leq \frac{8L\|\mathbf{x}^* - \mathbf{x}_0\|^2}{k+8}. \tag{3.17}$$

Note that we were able to replace the expression $R^2(\mathbf{x}_0)$ by $\|\mathbf{x}^* - \mathbf{x}_0\|^2$ (for some $\mathbf{x}^* \in X^*$) due to the well-known Fejér monotonicity property of the sequence generated by the gradient descent method, namely, that for every $\mathbf{x}^* \in X^*$

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \|\mathbf{x}_k - \mathbf{x}^*\|, \quad k = 0, 1, \ldots.$$

The result (3.17) is very similar to the efficiency estimates derived in the literature for the gradient descent method. Specifically, in [10, p. 70] the bound

$$f(\mathbf{x}_k) - f^* \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k+4}, k = 0, 1, \ldots,$$

was established, while in [3, Theorem 1.1] the bound

$$f(\mathbf{x}_k) - f^* \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k}, k = 1, 2, \ldots,$$

was derived.

*Remark* 3.4. In [18] Saha and Tewari studied the case in which the gradient of the objective function[1] $\nabla f$, in addition to being Lipschitz continuous with constant $L$, also satisfies an isotonicity assumption which means that the operator $\mathbf{x} \mapsto \mathbf{x} - \frac{\nabla f(\mathbf{x})}{L}$ is isotone. Under this assumption the authors studied the cyclic coordinate descent method which is the BCGD method with a conservative constant stepsize rule and $p = n$. It was shown in [18] that the sequence of function values satisfies

$$f(\mathbf{x}_k) - f^* \leq \frac{L\|\mathbf{x}^* - \mathbf{x}_0\|^2}{2k}.$$

The multiplicative constant in the latter result is better from the constant obtained in (3.16) since it does not depend on $p$. However, the efficiency estimate in (3.16) is derived without the isotonicity assumption, and under general convexity assumptions.

**3.2. Comparison to randomized block coordinate descent.** The complexity of randomized BCGD methods were initially studied by Nesterov in [13]. The input for the randomized BCGD method is a probability distribution vector, that is, a vector $\mathbf{q} \in \Delta_n$ ($\Delta_n$ being the unit simplex). At iteration $k (k \geq 0)$, an index is generated randomly according to the probability vector $\mathbf{q}$, meaning that the probability that the generated index is $i$ is $q_i$. The next iterate is then defined as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{\bar{L}_i}\mathbf{U}_i \nabla_i f(\mathbf{x}_k).$$

In [13] the distribution vector was chosen as

$$q_i = \frac{\bar{L}_i^\alpha}{\sum_{i=1}^n \bar{L}_j^\alpha}, \quad i = 1, 2, \ldots, n,$$

where $\alpha \in [0, 1]$ is a parameter. Of course, when $\alpha = 0$ the blocks are chosen by a uniform distribution and when $\alpha = 1$ the blocks are chosen in a probability proportional to the size of the corresponding Lipschitz constant. Nesterov called the method RCDM($\alpha$,$\mathbf{x}_0$) since it depends on the parameter $\alpha$ and the initial point $\mathbf{x}_0$. The BCGD method is of course different from the RCDM method by the simple fact that it is deterministic rather than stochastic. Since the sequence generated by the RCDM method is a sequence of random variables, the efficiency estimate result either bounds the difference of the *expectation* of the function values $f(\mathbf{x}_k)$ and $f^*$ or bounds the actual difference $f(\mathbf{x}_k) - f^*$ with a certain probability. The efficiency estimate obtained in [13] for the sequence generated by RCDM(1,$\mathbf{x}_0$) is given by

$$\mathbb{E}(f(\mathbf{x}_k)) - f^* \leq \frac{2\sum_{j=1}^p \bar{L}_j}{k+4} R^2(\mathbf{x}_0). \tag{3.18}$$

[1] The model in [18] is actually more involved since it also involves an $l_1$ regularization term.

The multiplicative constant in the above result is $M_{\text{RCDM}} \equiv 2\big(\sum_{j=1}^{p} \bar{L}_j\big)R^2(\mathbf{x}_0)$ is smaller than the multiplicative constant in the result of Theorem 3.6 which is $M_{\text{BCGD}} \equiv 4\bar{L}_{\max}(1 + pL^2/\bar{L}_{\min}^2)R^2(\mathbf{x}_0)$. Indeed,

$$M_{\text{RCDM}} \leq 2p\bar{L}_{\max}R^2(\mathbf{x}_0) \leq 4p\bar{L}_{\max}R^2(\mathbf{x}_0) \leq 4\bar{L}_{\max}(1+pL^2/\bar{L}_{\min}^2)R^2(\mathbf{x}_0) = M_{\text{RCDM}}.$$

In addition, note that each iteration of the RCDM method requires only an update of a single block, while each iteration of the BCGD methods involves the update of all the blocks. Therefore, the efficiency estimate of the RCDM method is better than the one of the BCGD method. However, two remarks are in order in this respect. First of all, the comparison is made between two different types of results. The result (3.18) for the RCDM method does not *guarantee* that the distance of the function values from the optimal values are bounded by a sequence converging to 0 in an $O(1/k)$ rate, since it only relates to the expectation. Second, in practice, it does not seem that the RCDM method has a clear advantage over the BCGD method. We illustrate this phenomena with a numerical example.

*Example* 3.1. Consider the least squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2,$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{b} \in \mathbb{R}^n$. $\mathbf{A}$ is a nonsingular matrix, so obviously the optimal solution of the problem is the vector $\mathbf{A}^{-1}\mathbf{b}$ and the optimal value is $f^* = 0$. We consider the partition of variables to $p$ blocks, each with $n/p$ variables (we assume that $p$ divides $n$). We will also use the notation

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_p \end{pmatrix},$$

where $\mathbf{A}_i$ is the submatrix of $\mathbf{A}$ comprising the columns corresponding to the $i$th block, that is, columns $(i-1)p+1, (i-1)p+2, \ldots, ip$. We will compare four algorithms:

- the BCGD method with the exact constant stepsize rule, $\bar{L}_j = L_j = \lambda_{\max}(\mathbf{A}_i^T\mathbf{A}_i)$;
- the method RCDM(1,$\mathbf{x}_0$) from [13] with stepsizes $\frac{1}{L_j}$;
- the method RCDM(0,$\mathbf{x}_0$) from [13] with stepsizes $\frac{1}{L_j}$;
- the gradient method with stepsize $\frac{1}{L}$ with $L = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$.

We begin by describing one run of the four methods for $n = 100,\ p = 5$. In this run the components of $\mathbf{A}$ and $\mathbf{b}$ were independently generated from a standard normal distribution. To make a fair comparison, we count every $p$ iterations of the RCDM method (each working on one block) as only *one* iteration, so that the computational effort at each iteration of the four methods is exactly the same—equivalent to the computation of the entire gradient. The function values of the sequence generated by each of the methods after 10, 100, 500, and 1000 iterations, and with the zeros vector as the initial point, is given in Table 3.1.

Obviously, in this run the BCGD method performs better than the two randomized methods, as well as from the gradient method. It can be seen that RCDM(1,$\mathbf{x}_0$) is second to the BCGD method and performs better than RCDM(0,$\mathbf{x}_0$) in which the blocks are picked by a uniform distribution.

TABLE 3.1
*Results of the four methods for a single realization.*

| Iteration number | BCGD | RCDM(1,$\mathbf{x}_0$) | RCDM(0,$\mathbf{x}_0$) | Gradient |
|---|---|---|---|---|
| 10 | 4.1404 | 4.4959 | 5.4491 | 5.9135 |
| 100 | 0.8720 | 1.1311 | 1.0743 | 1.8180 |
| 500 | 0.1732 | 0.2105 | 0.2101 | 0.4240 |
| 1000 | 0.0829 | 0.0963 | 0.1008 | 0.2031 |

We now wish to test the performance of the method for various choices of $\mathbf{A}, \mathbf{b}$, and $p$ and for a large amount of runs. We consider three choices of $p$: 2, 5, and 20. In addition, we consider two settings related to the generation of $\mathbf{A}$. In the *scaled setting* the components of $\mathbf{A}$ were randomly and independently generated from a standard normal distribution. In the *unscaled setting*, after generating $\mathbf{A}$ as in the scaled case, for each $i = 1, 2, \ldots, p$ the columns of $\mathbf{A}$ corresponding to the $i$th block are multiplied by the number $i$. This way, the Lipschitz constants $L_i$ are significantly different from each other. For each value of $p$ and choice of setting (scaled/unscaled), 100 problems were generated and 1000 iterations of each of the methods were employed with the initial vector chosen as the vector of all zeros. Of the 600 problems that were tested, the BCGD method got the best (i.e., lower) results in 594 runs, showing a clear advantage to the BCGD method. To quantify the advantage of the BCGD method over the other methods, the BCGD method was taken as a reference method, and in each run we computed the relative difference between the value obtained after 1000 iterations by each of the three methods and the value obtained after 1000 iterations by the BCGD method. That is, if for a certain run the sequences generated by the BCGD, RCDM(1,$\mathbf{x}_0$), RCDM(0,$\mathbf{x}_0$), and gradient methods are $\{\mathbf{x}_k\}, \{\mathbf{y}_k\}, \{\mathbf{z}_k\}, \{\mathbf{w}_k\}$, then the following three numbers were computed:

$$\frac{f(\mathbf{y}_{1000}) - f(\mathbf{x}_{1000})}{f(\mathbf{x}_{1000})}, \frac{f(\mathbf{z}_{1000}) - f(\mathbf{x}_{1000})}{f(\mathbf{x}_{1000})}, \frac{f(\mathbf{w}_{1000}) - f(\mathbf{x}_{1000})}{f(\mathbf{x}_{1000})}.$$

Table 3.2 presents for each choice of $p$ and setting the quantities $\text{rel}_1$, $\text{rel}_2$, $\text{rel}_3$ which are the averages over 100 runs of the relative difference of the RCDM(1,$\mathbf{x}_0$), RCDM(0,$\mathbf{x}_0$), and gradient methods. For example, the fact that the value of $\text{rel}_1$ for $p = 2$ and the scaled setting is 0.06 means that the values obtained after 1000 iterations by the RCDM(1,$\mathbf{x}_0$) method were higher in average by 6% than the value obtained after 1000 iterations by the BCGD method.

Clearly, the worst method is the gradient method. In addition, it can be clearly observed from the data in the table that the advantage of the BCGD over the other

TABLE 3.2
*Average over* 100 *realizations of the relative differences between RCDM(*1,$\mathbf{x}_0$*), RCDM(*0,$\mathbf{x}_0$*), and the gradient methods and the BCGD method.*

| $p$ | Setting | $\text{rel}_1$ | $\text{rel}_2$ | $\text{rel}_3$ |
|---|---|---|---|---|
| 2 | scaled | 0.060 | 0.063 | 0.310 |
| 2 | unscaled | 0.383 | 0.056 | 0.898 |
| 5 | scaled | 0.167 | 0.174 | 0.998 |
| 5 | unscaled | 1.408 | 0.1436 | 3.620 |
| 20 | scaled | 0.374 | 0.366 | 2.013 |
| 20 | unscaled | 7.889 | 0.383 | 15.985 |

methods becomes more significant in the unscaled setting and as the number of blocks ($p$) becomes larger.

**3.3. Strongly convex functions: Linear rate of convergence.** The gradient method converges in a linear rate when the objective function is assumed to be strongly convex and continuously differentiable with Lipschitz gradient; see, e.g., [10, 15]. It is therefore not a surprise that if we further assume that $f$ is strongly convex, then the linear rate of convergence of the BCGD method can be proved. Indeed, let us assume in this subsection that $f$ is strongly convex with parameter $\sigma > 0$, i.e.,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}\|^2 \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{3.19}$$

The linear rate of convergence result is established in the next theorem.

THEOREM 3.9 (linear rate of convergence under strong convexity). *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the BCGD method as applied to the function $f$ which is also assumed to be strongly convex with parameter $\sigma > 0$. Then*

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\sigma}{2\bar{L}_{\max}(1 + pL^2/\bar{L}_{\min}^2)}\right)^k (f(\mathbf{x}_0) - f^*), \quad k = 0, 1, \ldots. \tag{3.20}$$

*Proof.* By minimizing both sides of (3.19) with respect to $\mathbf{y}$, we obtain that

$$f(\mathbf{x}) - f^* \leq \frac{1}{2\sigma}\|\nabla f(\mathbf{x})\|^2 \text{ for every } \mathbf{x} \in \mathbb{R}^n. \tag{3.21}$$

Combining (3.21) with inequality (3.8) yields for every $k = 1, 2, \ldots$

$$\begin{aligned}(f(\mathbf{x}_{k-1}) - f^*) - (f(\mathbf{x}_k) - f^*) &= f(\mathbf{x}_{k-1}) - f(\mathbf{x}_k) \\ &\geq \frac{1}{4\bar{L}_{\max}(1 + pL^2/\bar{L}_{\min})}\|\nabla f(\mathbf{x}_{k-1})\|^2 \\ &\geq \frac{\sigma}{2\bar{L}_{\max}(1 + pL^2/\bar{L}_{\min})}(f(\mathbf{x}_{k-1}) - f^*),\end{aligned}$$

which immediately implies the desired result. ∎

*Remark* 3.5. Using the same arguments as in Corollaries 3.7 and 3.8 we deduce that in the exact constant stepsize setting (3.20) amounts to

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\sigma}{2L_{\max}(1 + p^2\kappa^2)}\right)^k (f(\mathbf{x}_0) - f^*), \quad k = 0, 1, \ldots, \tag{3.22}$$

and in the conservative constant stepsize rule

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\sigma}{2L(1 + p)}\right)^k (f(\mathbf{x}_0) - f^*), \quad k = 0, 1, \ldots. \tag{3.23}$$

**3.4. Dynamic stepsize rule.** In many scenarios, the block Lipschitz constants $L_1, \ldots, L_p$ are not known and thus it is also important to incorporate into the

optimization scheme a procedure for estimating these constants. The following version of the BCGD method incorporates a simple backtracking procedure and is almost as simple as the constant stepsizes versions.

**The BCGD method with backtracking.**

**Input:** $\bar{L}_j^0$ $(j = 1, \dots, p)$—initial estimates of the block Lipschitz constants.
$\eta > 1$—a constant.

**Initialization:** $\mathbf{x}_0 \in \mathbb{R}^n$.
**General step (k=0,1,...):** Set $\mathbf{x}_k^0 = \mathbf{x}_k$ and define recursively

$$\mathbf{x}_k^i = \mathbf{x}_k^{i-1} - \frac{1}{\eta^{\ell_i} \bar{L}_i^0} \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), i = 1, \dots, p, \tag{3.24}$$

where $\ell_i$ is the smallest nonnegative integer number for which

$$f(\mathbf{x}_k^{i-1}) - f\left(\mathbf{x}_k^{i-1} - \frac{1}{\eta^{\ell_i} \bar{L}_i^0} \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1})\right) \geq \frac{1}{2\eta^{\ell_i} \bar{L}_i^0} \|\nabla_i f(\mathbf{x}_k^{i-1})\|^2.$$

Set $\mathbf{x}_{k+1} = \mathbf{x}_k^p$.

Note that in principal the initial estimates $\bar{L}_j^0$ can be chosen arbitrarily and are not assumed to be upper bounds on $L_j$. It is not difficult to prove the sublinear rate of convergence of this variant of the BCGD method. The arguments are very similar to those used in the constant stepsize setting and we thus state the result without proof.

THEOREM 3.10. *Let $\{\mathbf{x}_k\}$ be the sequence generated by the BCGD method with backtracking stepsize rule. Then*

$$f(\mathbf{x}_k) - f^* \leq 4\eta L_{\max}(1 + pL^2/(\bar{L}_{\min}^0)^2) R^2(\mathbf{x}_0) \frac{1}{k + (8/p)}, \quad k = 0, 1, \dots,$$

*where $L_{\max}$ is given in* (3.1) *and $\bar{L}_{\min}^0 = \min_{j=1,\dots,p} \bar{L}_j^0$.*

*If $f$ is strongly convex with parameter $\sigma > 0$, then*

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\sigma}{2\eta L_{\max}(1 + pL^2/(\bar{L}_{\min}^0)^2)}\right)^k (f(\mathbf{x}_0) - f^*), \quad k = 0, 1, \dots.$$

**4. Acceleration of the BCGD method.** It is well known that the rate of convergence of gradient methods can be accelerated by using a multistep strategy. This was first shown by Nesterov in [11] and was later generalized for nonsmooth convex composite models in [12, 2, 3]. A general scheme of an optimal method can be found in the book of Nesterov [10] and is given here.

**Optimal scheme.**

**Input:** $M > 0$.

**Initialization:** $\mathbf{x}_0 \in \mathbb{R}^n, \mathbf{v}_0 = \mathbf{x}_0, \gamma_0 = M$.
**General step (k=0,1,...):**

- Compute $\alpha_k \in (0,1)$ which is a solution of the quadratic equation

$$M\alpha_k^2 = (1-\alpha_k)\gamma_k.$$

  Set

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k.$$

- $$\mathbf{y}_k = \alpha_k \mathbf{v}_k + (1-\alpha_k)\mathbf{x}_k.$$

- **Core step:** Find $\mathbf{x}_{k+1}$ satisfying $f(\mathbf{x}_{k+1}) \leq f(\mathbf{y}_k) - \frac{1}{2M}\|\nabla f(\mathbf{y}_k)\|^2$.

- Set $\mathbf{v}_{k+1} = \mathbf{v}_k - \frac{\alpha_k}{\gamma_{k+1}}\nabla f(\mathbf{y}_k)$.

What is missing in the description of the optimal gradient scheme is the value of the parameter $M$ and the description of the core step. The $O(1/k^2)$ convergence rate of the optimal scheme is recalled in the following theorem.

THEOREM 4.1 (see [10, Theorem 2.2.2]). *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the optimal scheme. Then*

$$f(\mathbf{x}_k) - f^* \leq \frac{4M\|\mathbf{x}^* - \mathbf{x}_0\|^2}{(k+2)^2}, \quad k = 0, 1, \ldots.$$

We will now show how to incorporate a step of the BCGD method in order to devise an accelerated version of this method.

**Core step for the accelerated BCGD method.**
**(input: $\mathbf{y}_k$ output: $\mathbf{x}_{k+1}$)**
Set $\mathbf{y}_k^0 = \mathbf{y}_k$ and define recursively

$$\mathbf{y}_k^i = \mathbf{y}_k^{i-1} - \frac{1}{\bar{L}_i}\mathbf{U}_i\nabla_i f(\mathbf{y}_k^{i-1}), \quad i = 1, \ldots, p,$$

$\mathbf{x}_{k+1} = \mathbf{y}_k^p$.

Here, we assume that for every $i = 1, \ldots, p$, $\bar{L}_i$ is an upper bound on the block Lipschitz constants $L_i$. By Lemma 3.3 the required inequality

$$f(\mathbf{y}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2M}\|\nabla f(\mathbf{y}_k)\|^2$$

holds with

$$M = 2\bar{L}_{\max}(1 + pL^2/\bar{L}_{\min}^2),$$

and the convergence result is given by the following theorem.

THEOREM 4.2. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the optimal scheme with the BCGD core step. Then*

$$f(\mathbf{x}_k) - f^* \leq \frac{8\bar{L}_{\max}(1 + pL^2/\bar{L}_{\min}^2)\|\mathbf{x}^* - \mathbf{x}_0\|^2}{(k+2)^2}, \quad k = 0, 1, \ldots.$$

**5. The alternating minimization method.** A closely related method to the BCGD method is the alternating minimization method in which at each iteration the objective function is minimized with respect to a different block taken in a cyclic order. We will analyze the method when $p = 2$, that is, when the decision variables vector $\mathbf{x}$ is split into two parts,

$$\mathbf{x} = \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = (\mathbf{y}; \mathbf{z}), \quad \mathbf{y} \in \mathbb{R}^{n_1}, \mathbf{z} \in \mathbb{R}^{n_2}.$$

In addition to our basic assumptions given in section 2, we will also assume that the solution of the problem

$$\min_{\mathbf{y}\in\mathbb{R}^{n_1}} f(\mathbf{y}; \bar{\mathbf{z}})$$

exists for any vector $\bar{\mathbf{z}} \in \mathbb{R}^{n_2}$ and that similarly the solution of the problem

$$\min_{\mathbf{z}\in\mathbb{R}^{n_2}} f(\bar{\mathbf{y}}; \mathbf{z})$$

exists for any vector $\bar{\mathbf{y}} \in \mathbb{R}^{n_1}$. The alternating minimization method is detailed below.

**The alternating minimization method.**

**Initialization:** $\mathbf{x}_0 = (\mathbf{y}_0; \mathbf{z}_0)$, where $\mathbf{y}_0 \in \mathbb{R}^{n_1}, \mathbf{z}_0 \in \mathbb{R}^{n_2}$.
**General step (k=0,1,...):** Compute

$$\mathbf{y}_{k+1} \in \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^{n_1}} f(\mathbf{y}; \mathbf{z}_k), \tag{5.1}$$

$$\mathbf{z}_{k+1} \in \operatorname*{argmin}_{\mathbf{z}\in\mathbb{R}^{n_2}} f(\mathbf{y}_{k+1}; \mathbf{z}). \tag{5.2}$$

Set $\mathbf{x}_{k+1} = (\mathbf{y}_{k+1}; \mathbf{z}_{k+1})$.

The alternating minimization method is also known as the *nonlinear Gauss–Seidel* method or the *block coordinate descent* method (see, e.g., [4]). By its definition, the method produces a nonincreasing sequence of function values. In addition to the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ generated by the method, we will also be interested in the "sequence in between" defined by

$$\mathbf{x}_{k+\frac{1}{2}} \equiv (\mathbf{y}_{k+1}; \mathbf{z}_k), \quad k = 0, 1, \ldots,$$

and we have

$$f(\mathbf{x}_0) \geq f(\mathbf{x}_{\frac{1}{2}}) \geq f(\mathbf{x}_1) \geq f(\mathbf{x}_{\frac{3}{2}}) \geq f(\mathbf{x}_2) \geq \cdots.$$

The convergence of the alternating minimization method is based on the following technical lemma.

LEMMA 5.1. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ and $\{\mathbf{x}_{k+\frac{1}{2}}\}_{k\geq 0}$ be the sequences generated by the alternating minimization method. Then for any $k = 0, 1, \ldots$*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+\frac{1}{2}}) \geq \frac{1}{2L_1}\|\nabla f(\mathbf{x}_k)\|^2, \tag{5.3}$$

$$f(\mathbf{x}_{k+\frac{1}{2}}) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2L_2}\|\nabla f(\mathbf{x}_{k+\frac{1}{2}})\|^2. \tag{5.4}$$

*Proof.* By the definition of $\mathbf{y}_{k+1}$ given in (5.1) we have

$$f(\mathbf{y}_{k+1}; \mathbf{z}_k) \leq f\left(\mathbf{y}_k - \frac{1}{L_1}\nabla_1 f(\mathbf{x}_k); \mathbf{z}_k\right).$$

Therefore,

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_{k+\frac{1}{2}}) &= f(\mathbf{y}_k; \mathbf{z}_k) - f(\mathbf{y}_{k+1}; \mathbf{z}_k) \\ &\geq f(\mathbf{y}_k; \mathbf{z}_k) - f\left(\mathbf{y}_k - \frac{1}{L_1}\nabla_1 f(\mathbf{x}_k); \mathbf{z}_k\right) \\ &\geq \frac{1}{2L_1}\|\nabla_1 f(\mathbf{x}_k)\|^2, \end{aligned}$$

where the last inequality follows from the block descent lemma. The result (5.3) is a direct consequence of the fact that by the definition of the method we have $\nabla_2 f(\mathbf{x}_k) = 0$. A similar argument shows the validity of (5.4). ∎

The next theorem establishes the convergence rate of the alternating minimization method.

THEOREM 5.2. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the alternating minimization method. Then*

$$f(\mathbf{x}_k) - f^* \leq \frac{2\min\{L_1, L_2\}R^2(\mathbf{x}_0)}{k-1}, \quad k = 2, 3, \ldots. \tag{5.5}$$

*If $f$ is strongly convex with parameter $\sigma$, then*

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\sigma}{\min\{L_1, L_2\}}\right)^{k-1}(f(\mathbf{x}_0) - f^*), \quad k = 1, 2, \ldots. \tag{5.6}$$

*Proof.* Since $f$ is convex we have

$$f(\mathbf{x}_k) - f^* \leq R(\mathbf{x}_0)\|\nabla f(\mathbf{x}_k)\|. \tag{5.7}$$

On the other hand, by (5.3) and (5.4), we have for every $k = 0, 1, \ldots$

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq f(\mathbf{x}_k) - f(\mathbf{x}_{k+\frac{1}{2}}) \geq \frac{1}{2L_1}\|\nabla f(\mathbf{x}_k)\|^2, \tag{5.8}$$

$$f(\mathbf{x}_{k+\frac{1}{2}}) - f(\mathbf{x}_{k+\frac{3}{2}}) \geq f(\mathbf{x}_{k+\frac{1}{2}}) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2L_2}\|\nabla f(\mathbf{x}_{k+\frac{1}{2}})\|^2. \tag{5.9}$$

Combining (5.8) with (5.7) we obtain that

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{(f(\mathbf{x}_k) - f^*)^2}{2L_1 R^2(\mathbf{x}_0)}. \tag{5.10}$$

In addition,

$$f(\mathbf{x}_0) - f^* \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \frac{L}{2}R^2(\mathbf{x}_0).$$

Invoking Lemma 3.5 with $A_k = f(\mathbf{x}_k) - f^*$ and $\gamma = \frac{1}{2L_1R^2(\mathbf{x}_0)}$, we obtain that for every $k = 2, 3, \ldots$

$$f(\mathbf{x}_k) - f^* \leq \frac{2L_1R^2(\mathbf{x}_0)}{k} \leq \frac{2L_1R^2(\mathbf{x}_0)}{k-1}. \tag{5.11}$$

Combining (5.9) with (5.7) we obtain that for every $k = 0, 1, \ldots$

$$f(\mathbf{x}_{k+\frac{1}{2}}) - f(\mathbf{x}_{k+\frac{3}{2}}) \geq \frac{(f(\mathbf{x}_{k+\frac{1}{2}}) - f^*)^2}{2L_2R^2(\mathbf{x}_0)}. \tag{5.12}$$

Invoking Lemma 3.5 with $A_k = f(\mathbf{x}_{k+\frac{1}{2}}) - f^*$ and $\gamma = \frac{1}{2L_2R^2(\mathbf{x}_0)}$, we obtain that for every $k = 1, 2, \ldots$

$$f(\mathbf{x}_{k+\frac{1}{2}}) - f^* \leq \frac{2L_2R^2(\mathbf{x}_0)}{k}.$$

Finally, for every $k = 2, 3, \ldots$

$$f(\mathbf{x}_k) - f^* \leq f(\mathbf{x}_{k-\frac{1}{2}}) - f^* \leq \frac{2L_2R^2(\mathbf{x}_0)}{k-1},$$

which combined with (5.11) implies the result (5.5).

If $f$ is strongly convex with parameter $\sigma > 0$, then by (3.21) it follows that

$$\begin{aligned} f(\mathbf{x}_k) - f^* &\leq \frac{1}{2\sigma}\|\nabla f(\mathbf{x}_k)\|^2 \\ &\overset{(5.3)}{\leq} \frac{L_1}{\sigma}(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) = \frac{L_1}{\sigma}[(f(\mathbf{x}_k) - f^*) - (f(\mathbf{x}_{k+1}) - f^*)], \end{aligned}$$

which implies that

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\sigma}{L_1}\right)^k (f(\mathbf{x}_0) - f^*) \leq \left(1 - \frac{\sigma}{L_1}\right)^{k-1} (f(\mathbf{x}_0) - f^*). \tag{5.13}$$

Similarly,

$$\begin{aligned} f(\mathbf{x}_{k+\frac{1}{2}}) - f^* &\leq \frac{1}{2\sigma}\|\nabla f(\mathbf{x}_{k+\frac{1}{2}})\|^2 \\ &\overset{(5.4)}{\leq} \frac{L_2}{\sigma}(f(\mathbf{x}_{k+\frac{1}{2}}) - f(\mathbf{x}_{k+\frac{3}{2}})) = \frac{L_2}{\sigma}[(f(\mathbf{x}_{k+\frac{1}{2}}) - f^*) - (f(\mathbf{x}_{k+\frac{3}{2}}) - f^*)], \end{aligned}$$

and hence for all $k = 1, 2, \ldots$

$$f(\mathbf{x}_k) - f^* \leq f(\mathbf{x}_{k-\frac{1}{2}}) - f^* \leq \left(1 - \frac{\sigma}{L_2}\right)^{k-1} (f(\mathbf{x}_{\frac{1}{2}}) - f^*) \leq \left(1 - \frac{\sigma}{L_2}\right)^{k-1} (f(\mathbf{x}_0) - f^*),$$

which combined with (5.13) implies (5.6). ∎

The convergence result for the alternating minimization method given in Theorem 5.2 is better than the result for the BCGD method (with $p = 2$) given in Theorem 3.6. Let us compare the efficiency estimates given in Theorems 3.6 and 5.2 for the convex case. (The comparison for the strongly convex case is practically the same.) The multiplicative coefficients are of course ordered as follows:

$$4\bar{L}_{\max}(1 + pL^2/\bar{L}_{\min}) > 2\min\{L_1, L_2\};$$

this means that the efficiency estimate of the alternating minimization method is better than the one of the BCGD method. More importantly, note that in terms of Lipschitz constants, the efficiency estimate of the alternating minimization method proportionally depends on $\min\{L_1, L_2\}$, which means that if we have a poorly scaled problem in which one of the Lipschitz constants is much larger than the other, then the alternating minimization method practically ignores this scalability issue and is as strong as its strongest link. The opposite situation exists for the BCGD method, in which the efficiency estimate depends on $\bar{L}_{\max}$ which means that that the largest Lipschitz constant is the dominant constant in the efficiency estimate and the method is in a sense as strong as its weakest link. Of course, the fact that the alternating minimization method is faster than the BCGD method is not a surprise since the BCGD method only performs gradient steps with respect to the individual blocks, while the alternating minimization method performs an exact minimization with respect to each of the blocks.

**6. The constrained setting.** In the previous sections we considered the unconstrained setting in which the minimization is performed over the entire space $\mathbb{R}^n$. Unfortunately, it seems that the previous analysis cannot be directly generalized to the constrained setting. We will show, by using a different line of analysis, that the appropriate modification of the BCGD method to the constrained setting also possesses a sublinear rate of convergence. The constrained problem we consider is

$$\min\{f(\mathbf{x}) : \mathbf{x} \in X\}. \tag{6.1}$$

The objective function $f$ satisfies all the assumptions made in section 2 (convex, continuously differentiable with Lipschitz gradient; the block Lipschitz constants of the gradient are $L_i$ and the global Lipschitz constant is $L$). The feasible set $X \subseteq \mathbb{R}^n$ is the closed convex set defined by the Cartesian product

$$X \equiv X_1 \times X_2 \times \cdots \times X_p,$$

where $X_i \subseteq \mathbb{R}^{n_i}$ is a closed convex set for every $i = 1, 2, \ldots, p$. The definition of $R(\mathbf{x}_0)$ in the constrained setting is generalized as follows:

$$R(\mathbf{x}_0) \equiv \max_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{x}^*\in X^*} \{\|\mathbf{x} - \mathbf{x}^*\| : f(\mathbf{x}) \le f(\mathbf{x}_0), \mathbf{x} \in X\}.$$

The orthogonal projection of a vector $\mathbf{y}$ onto a given closed convex set $S \subseteq \mathbb{R}^k$ is denoted by $P_S(\mathbf{y})$, and we recall the basic property of the projection mapping—for a given $\mathbf{y} \in \mathbb{R}^k$, the following relation holds:

$$\langle \mathbf{y} - P_S(\mathbf{y}), \mathbf{x} - P_S(\mathbf{y})\rangle \le 0 \quad \text{for every } \mathbf{x} \in S. \tag{6.2}$$

In the constrained version of the BCGD method we employ an orthogonal projection after each gradient step, giving rise to the BCGP method, which is given below.

**The BCGP method.**

**Input:** $\bar{L}_j (j = 1, \ldots, p)$—upper bounds on the Lipschitz constant $L_j$ ($\bar{L}_j \geq L_j$).

**Initialization:** $\mathbf{x}_0 \in X$.
**General step (k=0,1,...):** Set $\mathbf{x}_k^0 = \mathbf{x}_k$ and define recursively for $i = 1, 2, \ldots, p$:

$$j = 1, 2, \ldots, p \quad \mathbf{x}_k^i(j) = \begin{cases} \mathbf{x}_k^{i-1}(j), & j \neq i, \\ P_{X_i}\left(\mathbf{x}_k^{i-1}(i) - \frac{1}{\bar{L}_i}\nabla_i f(\mathbf{x}_k^{i-1})\right), & j = i. \end{cases}$$

Set $\mathbf{x}_{k+1} = \mathbf{x}_k^p$.

To establish the sublinear rate of convergence of the BCGP method, we will now prove a result on the difference of consecutive function values $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$.

LEMMA 6.1. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the BCGP method. Then*

(6.3)
$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{\bar{L}_{\min}}{2p([\bar{L}_{\max} + 2L]R(\mathbf{x}_0) + M)^2}(f(\mathbf{x}_{k+1}) - f^*)^2, \quad k = 0, 1, \ldots,$$

*where*

(6.4)
$$M = \max_{\mathbf{x}^* \in X^*} \|\nabla f(\mathbf{x}^*)\|.$$

*Proof.* Since

$$\mathbf{x}_k^i(i) = P_{X_i}\left(\mathbf{x}_k^{i-1}(i) - \frac{1}{\bar{L}_i}\nabla_i f(\mathbf{x}_k^{i-1})\right),$$

it follows by the characterization (6.2) of the projection operator that

(6.5)
$$\left\langle \mathbf{x}_k^{i-1}(i) - \frac{1}{\bar{L}_i}\nabla_i f(\mathbf{x}_k^{i-1}) - \mathbf{x}_k^i(i), \mathbf{z} - \mathbf{x}_k^i(i) \right\rangle \leq 0$$

for every $\mathbf{z} \in X_i$. Plugging $\mathbf{z} = \mathbf{x}_k^{i-1}(i)$ into the latter inequality yields

(6.6)
$$\left\langle \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^{i-1}(i) - \mathbf{x}_k^i(i) \right\rangle \geq \bar{L}_i \|\mathbf{x}_k^i(i) - \mathbf{x}_k^{i-1}(i)\|^2, \quad i = 1, 2, \ldots, p.$$

By the block descent lemma,

$$f(\mathbf{x}_k^i) \leq f(\mathbf{x}_k^{i-1}) + \langle \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^i(i) - \mathbf{x}_k^{i-1}(i) \rangle + \frac{\bar{L}_i}{2}\|\mathbf{x}_k^i(i) - \mathbf{x}_k^{i-1}(i)\|^2, \quad i = 1, 2, \ldots, p,$$

which combined with (6.6) implies

$$f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}_k^i) \geq \frac{\bar{L}_i}{2}\|\mathbf{x}_k^{i-1} - \mathbf{x}_k^i\|^2, \quad i = 1, 2, \ldots, p.$$

Summing the above inequality for $i = 1, 2, \ldots, p$ and using the fact that $\bar{L}_i \geq \bar{L}_{\min}$, it follows that

(6.7)
$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{\bar{L}_{\min}}{2}\sum_{i=1}^{p}\|\mathbf{x}_k^i - \mathbf{x}_k^{i-1}\|^2 = \frac{\bar{L}_{\min}}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2,$$

where the last equality follows from the fact that

$$\mathbf{x}_k^i(i) - \mathbf{x}_k^{i-1}(i) = \mathbf{x}_{k+1}(i) - \mathbf{x}_k(i), \quad i = 1, 2, \ldots, p.$$

Let $i \in \{1, 2, \ldots, p\}$ and $\mathbf{x}^* \in X^*$. A direct consequence of (6.6) is the inequality

$$\langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^i - \mathbf{x}_k^{i-1} \rangle \leq 0, \tag{6.8}$$

from which it follows that

$$\langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}^* - \mathbf{x}_k^{i-1} \rangle \leq \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}^* - \mathbf{x}_k^i \rangle = \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle. \tag{6.9}$$

By the Lipschitz continuity of the gradient $\nabla f$ we have

$$\begin{aligned} \|\mathbf{U}_i \nabla_i f(\mathbf{x}_{k+1}) - \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1})\| &\leq \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k^{i-1})\| \\ &\leq L\|\mathbf{x}_{k+1} - \mathbf{x}_k^{i-1}\| \\ &\leq L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|, \end{aligned}$$

and hence, using the Cauchy–Schwartz inequality,

$$\begin{aligned} \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle &= \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle \\ &\quad + \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}) - \mathbf{U}_i \nabla_i f(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle \\ &\leq \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle \\ &\quad + L\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \|\mathbf{x}^*(i) - \mathbf{x}_{k+1}(i)\|, \end{aligned}$$

which combined with (6.9) yields

$$\langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}^* - \mathbf{x}_k^{i-1} \rangle \leq \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle + L\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \|\mathbf{x}^*(i) - \mathbf{x}_{k+1}(i)\|. \tag{6.10}$$

We will now use the following simple identity, which follows from the Cauchy–Schwartz inequality: if $a_1, a_2, \ldots, a_m$ are nonnegative integers, then

$$\sum_{i=1}^m a_i \leq \sqrt{m} \sqrt{\sum_{i=1}^m a_i^2}. \tag{6.11}$$

Using the above inequality with $m = p$ and $a_i = \|\mathbf{x}^*(i) - \mathbf{x}_{k+1}(i)\|$ and summing up (6.10) we obtain

$$\begin{aligned} \sum_{i=1}^p \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}^* - \mathbf{x}_k^{i-1} \rangle &\leq \sum_{i=1}^p \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle \\ &\quad + L\sqrt{p}\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \|\mathbf{x}^* - \mathbf{x}_{k+1}\|. \end{aligned} \tag{6.12}$$

Since $f$ is convex,

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f^* &\leq \langle \nabla f(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_* \rangle = \sum_{i=1}^p \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_* \rangle \\ &\overset{(6.12)}{\leq} \sum_{i=1}^p \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^{i-1} - \mathbf{x}_* \rangle + L\sqrt{p}\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \|\mathbf{x}^* - \mathbf{x}_{k+1}\|. \end{aligned} \tag{6.13}$$

Now, substituting $\mathbf{z} = \mathbf{x}^*$ in (6.5) results in the inequality

$$\langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^i - \mathbf{x}^* \rangle \le \bar{L}_i \langle \mathbf{x}_k^{i-1}(i) - \mathbf{x}_k^i(i), \mathbf{x}_k^i(i) - \mathbf{x}^*(i) \rangle.$$

Plugging the latter inequality into (6.13) implies

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f^* &\le \sum_{i=1}^{p} \langle \mathbf{U}_i \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^{i-1} - \mathbf{x}_k^i + \mathbf{x}_k^i - \mathbf{x}_* \rangle \\ &\quad + L\sqrt{p} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \|\mathbf{x}^* - \mathbf{x}_{k+1}\| \\ &\le \sum_{i=1}^{p} \bar{L}_i \|\mathbf{x}_k^{i-1}(i) - \mathbf{x}_k^i(i)\| \cdot \|\mathbf{x}_k^i - \mathbf{x}^*\| + \sum_{i=1}^{p} \|\nabla_i f(\mathbf{x}_k^{i-1})\| \\ &\quad \cdot \|\mathbf{x}_k^{i-1}(i) - \mathbf{x}_k^i(i)\| + L\sqrt{p} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \cdot \|\mathbf{x}^* - \mathbf{x}_{k+1}\|. \end{aligned} \tag{6.14}$$

Now, the following inequality is a consequence of (6.11) with $m = p, a_i = \|\mathbf{x}_k(i) - \mathbf{x}_{k+1}(i)\|$:

$$\sum_{i=1}^{p} \|\mathbf{x}_k^{i-1}(i) - \mathbf{x}_k^i(i)\| = \sum_{i=1}^{p} \|\mathbf{x}_k(i) - \mathbf{x}_{k+1}(i)\| \le \sqrt{p} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|.$$

Using the above inequality, along with the identity

$$\mathbf{x}_k^{i-1}(i) - \mathbf{x}_k^i(i) = \mathbf{x}_k(i) - \mathbf{x}_{k+1}(i),$$

we have the following inequalities:

$$\begin{aligned} \sum_{i=1}^{p} \bar{L}_i \|\mathbf{x}_k^{i-1} - \mathbf{x}_k^i\| \cdot \|\mathbf{x}_k^i - \mathbf{x}^*\| &\le \bar{L}_{\max} \sqrt{p} R(\mathbf{x}_0) \|\mathbf{x}_k - \mathbf{x}_{k+1}\|, \\ \sum_{i=1}^{p} \|\nabla_i f(\mathbf{x}_k^{i-1})\| \cdot \|\mathbf{x}_k^{i-1}(i) - \mathbf{x}_k^i(i)\| &\le \sum_{i=1}^{p} (\|\nabla f(\mathbf{x}_k^{i-1}) - \nabla f(\mathbf{x}^*)\| \\ &\quad + \|\nabla f(\mathbf{x}^*)\|) \cdot \|\mathbf{x}_k(i) - \mathbf{x}_{k+1}(i)\| \\ &\le (LR(\mathbf{x}_0) + M)\sqrt{p} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|, \\ L\sqrt{p} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \cdot \|\mathbf{x}^* - \mathbf{x}_{k+1}\| &\le L\sqrt{p} R(\mathbf{x}_0) \|\mathbf{x}_k - \mathbf{x}_{k+1}\|, \end{aligned}$$

which together with (6.14) and (6.7) yields

$$(f(\mathbf{x}_{k+1}) - f^*)^2 \le \frac{2p([\bar{L}_{\max} + 2L]R(\mathbf{x}_0) + M)^2}{\bar{L}_{\min}} (f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})),$$

establishing the desired result. □

The result of Lemma 6.1 bears a resemblance to the result of Lemma 3.4, but note that here the right-hand side depends on $f(\mathbf{x}_{k+1})$ rather than on $f(\mathbf{x}_k)$, which requires a different result on the rate of convergence of sequences.

LEMMA 6.2. *Let $\{A_k\}_{k \ge 0}$ be a nonnegative sequence of real numbers satisfying*

$$A_k - A_{k+1} \ge \gamma A_{k+1}^2, \quad k = 0, 1, \ldots, \tag{6.15}$$

*and*

$$A_1 \le \frac{1.5}{\gamma}, \ A_2 \le \frac{1.5}{2\gamma}$$

*for some positive $\gamma$. Then*

$$A_k \leq \frac{1.5}{\gamma}\frac{1}{k}, \quad k = 1, 2, \ldots. \tag{6.16}$$

*Proof.* We will prove the lemma by induction with respect to $k$. For $k = 1, 2$ we have

$$A_1 \leq \frac{1.5}{\gamma}, \; A_2 \leq \frac{1.5}{2\gamma}.$$

Let us assume that for some $k \geq 2$

$$A_k \leq \frac{1.5}{\gamma}\frac{1}{k}. \tag{6.17}$$

Combining (6.15) and the induction assumption (6.17), we get

$$\gamma A_{k+1}^2 + A_{k+1} \leq A_k \leq \frac{1.5}{\gamma} \cdot \frac{1}{k}.$$

Thus,

$$A_{k+1} \leq \frac{\sqrt{1 + \frac{4 \cdot 1.5}{k}} - 1}{2\gamma} \leq \frac{1.5}{\gamma} \cdot \frac{1}{k+1}. \qquad \square$$

Combining Lemmata 6.1 and 6.2 we arrive at the main result which establishes the sublinear rate of convergence of the BCGP method.

THEOREM 6.3 (sublinear rate of convergence of the BCGP). *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the BCGP method. Then for $k = 1, 2, \ldots$*

$$f(\mathbf{x}_k) - f^* \leq \frac{3p([\bar{L}_{\max} + 2L]R(\mathbf{x}_0) + M)^2}{\bar{L}_{\min}} \cdot \frac{1}{k}, \tag{6.18}$$

*where $M$ is given in* (6.4).

*Proof.* The result follows by invoking Lemma 6.2 with

$$A_k = f(\mathbf{x}_k) - f^* \text{ and } \gamma = \frac{\bar{L}_{\min}}{2p([\bar{L}_{\max} + 2L]R(\mathbf{x}_0) + M)^2}.$$

What is left to show is that

$$A_1 = f(\mathbf{x}_1) - f^* \leq \frac{1.5}{\gamma}, A_2 = f(\mathbf{x}_2) - f^* \leq \frac{1.5}{2\gamma} \tag{6.19}$$

and indeed, by the descent lemma we have for every $k$ ($\mathbf{x}^*$ is an arbitrary vector in $X^*$)

$$f(\mathbf{x}_k) - f^* \leq \langle \nabla f(\mathbf{x}^*), \mathbf{x}_k - \mathbf{x}^* \rangle + \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq MR(\mathbf{x}_0) + \frac{L}{2}R^2(\mathbf{x}_0),$$

where the last inequality uses (6.4) and the definition of $R(\mathbf{x}_0)$. Now, we also have

$$\begin{aligned}
\frac{1}{\gamma} &= \frac{2p([\bar{L}_{\max} + 2L]R(\mathbf{x}_0) + M)^2}{\bar{L}_{\min}} \geq \frac{2p(4\bar{L}_{\max}LR^2(\mathbf{x}_0) + 2\bar{L}_{\max}MR(\mathbf{x}_0))}{\bar{L}_{\min}} \\
&\overset{\bar{L}_{\max} \geq \bar{L}_{\min}}{\geq} 8pLR^2(\mathbf{x}_0) + 4pMR(\mathbf{x}_0) \\
&\geq 4p\left(MR(\mathbf{x}_0) + \frac{L}{2}R^2(\mathbf{x}_0)\right) \\
&\geq 4p(f(\mathbf{x}_k) - f^*),
\end{aligned}$$

showing the validity of (6.19) and, consequently, also of the desired result (6.18). □

Note that in the unconstrained case ($X = \mathbb{R}^n$), the constant $M$ defined in (6.4) is zero, and thus the efficiency bound reduces to

$$f(\mathbf{x}_k) - f^* \leq \frac{3p(\bar{L}_{\max} + 2L)^2 R^2(\mathbf{x}_0)}{\bar{L}_{\min}} \cdot \frac{1}{k}, \tag{6.20}$$

establishing once again the sublinear rate of convergence of the BCGD method. Despite for this, the analysis of section 3 is necessary for several reasons: first of all, the bound (6.20) might be worse than the bound given in Theorem 3.6. For example, in the conservative constant stepsize rule, the efficiency bound (6.20) amounts to

$$f(\mathbf{x}_k) - f^* \leq \frac{27pLR^2(\mathbf{x}_0)}{k}, \quad k = 1, 2, \ldots,$$

which is clearly worse than the bound obtained under the analysis of section 3:

$$f(\mathbf{x}_k) - f^* \leq \frac{4L(1+p)R^2(\mathbf{x}_0)}{k + 8/p}, \quad k = 0, 1, \ldots.$$

(The multiplicative constants are dominant over the additive constants in the denominator.) In addition, the analysis in this section, as opposed to the analysis of section 3, cannot be used in order to show two main results on the BCGD method: the linear rate of convergence under the strong convexity assumption and the proof that the BCGD method can be used as a core step of the optimal scheme.

**7. Concluding remarks.** In this paper we considered two block coordinate descent type methods. The first method is the block coordinate gradient projection method. We were able to show that when the blocks are taken in a cyclic order, the sequence of function values converges to the optimal value in an $O(1/k)$ rate, and when acceleration is incorporated in the unconstrained case, an $O(1/k^2)$ rate of convergence can be proved. Despite the fact that the analyzed method assumes that at each iteration the blocks are taken by the same order $i = 1, 2, \ldots, p$, the analysis of the method is actually also valid when at each iteration, a different order of the iterations is taken. The second method considered in the paper is the alternating minimization method when the number of blocks is two. For this method, we were also able to show an $O(1/k)$ rate of convergence result of the function values and that the constant depends on the *minimal* Lipschitz constant of the two blocks, which has the nice interpretation that the convergence rate is "optimistic"—it only depends on the "better behaved" problem. Finally, we would like to note that there are still several open issues that can be the basis of future research on the block coordinate descent method. For example, it is still unclear how to derive a rate of convergence result for the alternating minimization method when the number of blocks is greater than two. In addition, improving the constants in the efficiency estimates derived for the BCGP method is an important theoretical challenge.

## REFERENCES

[1] A. Auslender, *Optimisation*, Masson, Paris, 1976.

[2] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[3] A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery problems*, in Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar, eds., Cambridge University Press, Cambridge, UK, 2009, pp. 139–162.
[4] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
[5] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[6] I. S. Dhillon, P. D. Ravikumar, and A. Tewari, *Nearest Neighbor Based Greedy Coordinate Descent*, in Proceedings of NIPS, 2011, pp. 2160–2168.
[7] L. Grippo and M. Sciandrone, *Globally convergent block-coordinate techniques for unconstrained optimization*, Optim. Methods Soft., 10 (1999), pp. 587–637.
[8] T. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35.
[9] T. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 46 (1993), pp. 157–178.
[10] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer, Boston, 2004.
[11] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate* $O(1/k^2)$, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547.
[12] Y. E. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, http://www.ecore.be/DPs/dp1191313936.pdf (2007).
[13] Y. E. Nesterov, *Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems*, CORE Discussion paper 2010/2, 2010.
[14] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[15] B. T. Polyak, *Introduction to Optimization*, Transl. Ser. Math. Engrg., Optimization Software, Inc., New York, 1987.
[16] P. Richtarik and M. Takac, *Efficient Serial and Parallel Coordinate Descent Methods for Huge-Scale Truss Topology Design*, Oper. Res. Proc. 2011, Springer, 2012, pp. 27–32.
[17] P. Richtarik and M. Takac, *Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function*, Math. Prog. Ser. A, (2012), pulished online.
[18] A. Saha and A. Tewari, *On the non-asymptotic convergence of cyclic coordinate descent methods*, SIAM J. Optim., 23 (2013), pp. 576–601.
[19] S. Shalev-Shwartz and A. Tewari, *Stochastic methods for* $\ell_1$*-regularized loss minimization*, J. Mach. Learn. Res., 12 (2011), pp. 1865–1892.