Regret Analysis of Block Coordinate Gradient Methods for Online Convex Programming

Xiaoqin Hua[*] So Kadomoto[†] Nobuo Yamashita[‡]

May 12, 2015

Abstract. In this paper, we propose two block coordinate gradient (BCG) methods for an online convex optimization problem with a separable structure: the BCG method with the cyclic rule and the BCG method with the random rule. The proposed methods solve a low dimensional problem on each iteration, and hence they are efficient for large scale problems. For the proposed methods, under usual assumptions, we show that their regret bounds are $O(\sqrt{T})$, where T is the number of time steps. These results are shown to be natural extensions of that for the greedy projection method by Zinkevich.

Keywords. Cyclic rule, Random rule, Regret, Online convex optimization problem, Stochastic optimization problem.

1 Introduction

The online convex programming is a powerful learning model, which has attracted great attention in many large scale optimization fields, such as the machine learning [1], the network routing [1], and the investment decisions [7]. By this model, a decision maker makes a sequence of decisions for his/her practical problems, where their possible options are given as a convex set in advance. The precise definition of the online convex programming problem is recalled by Definition 2.1 in Section 2. Roughly speaking, its main characteristics include the following two aspects in contrast to the classical convex optimization problems.

- We minimize a sequence of dynamically generated loss functions $\{F^t(x),\, t = 1, 2, \cdots\}$

[*]School of Mathematics and Physics, Jiangsu University of Science and Technology, Zhenjiang 212003, CHINA. Current address: Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, JAPAN. E-mail: hua.xiaoqin.22r@st.kyoto-u.ac.jp

[†]Brownies inc.

[‡]Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, JAPAN. E-mail:nobuo@i.kyoto-u.ac.jp

in the online optimization problems, where t denotes the time step when new function is generated.

- We must make a decision at the time step t, denoted by x^t, before getting the true loss function $F^t(x)$.

Note that the online convex optimization problems are closely related to the stochastic convex optimization problems. For example, the stochastic gradient method is essentially same as the greedy projection method [26] for the online optimization problems.

In this paper, we consider an online convex optimization problem with a separable structure, whose loss function $F^t : \Omega \to \mathcal{R}$ at the time step t is given as follows.

$$F^t(x) := f^t(x) + \tau\psi(x), \quad t = 1, 2, \cdots, \tag{1.1}$$

where $f^t : \Omega \to \mathcal{R}$ is smooth and convex, $\Omega \subseteq \bigcap_{t=1}^{\infty} \mathrm{dom}\, F^t$ is a nonempty convex set, τ is a positive constant, and $\psi : \Omega \to (-\infty, \infty]$ is a proper, convex and lower semicontinuous (l.s.c.) function with the block separable structure, that is, $\psi(x) = \sum_{i=1}^{N} \psi_i(x_{\mathcal{J}^i})$, where $x_{\mathcal{J}^i} \in \mathcal{R}^{|\mathcal{J}^i|}$, $\psi_i : \mathcal{R}^{|\mathcal{J}^i|} \to \mathcal{R}$ and $n = \sum_{i=1}^{N} |\mathcal{J}^i|$. The loss function (1.1) appears in many practical problems, where function $f^t(x)$ represents a static function at the time step t, and function $\psi(x)$ is a regularization term. For example, in the simple online linear regression problem [2, 17], function $f^t(x)$ is used to estimate the relationship among variables. In the sequential investment problem [2], function $f^t(x)$ denotes the logistic wealth ratio, which is given by a logistic function. The most common variants of function $\psi(x)$ are l_1-regularization [12], l_2-regularization [24], and an indicator function with respect to closed convex separable constraints [3].

From the characteristics of the online convex optimization problems, we know that it is impossible to select a point x^t that exactly minimize the loss function $F^t(x)$ at the time step t, because we do not know the true loss function $F^t(x)$ until the decision x^t is determined. Instead, the researchers, who study the online optimization problems, focus on proposing an algorithm to generate decisions $\{x^t, t = 1, 2, \cdots\}$, with which, for given $T > 0$, the practical total loss $\sum_{t=1}^{T} F^t(x^t)$ is not much larger than the ideal total loss $\sum_{t=1}^{T} F^t(x^*)$, where x^* is an optimal solution in some sense, e.g., $x^* \in \mathrm{argmin}_{x\in\Omega} \frac{1}{T}\sum_{t=1}^{T} F^t(x)$ [26]. For convenience, we call the difference between these two values the "regret" [26], denoted by $R(T)$, i.e.,

$$R(T) = \sum_{t=1}^{T} F^t(x^t) - \sum_{t=1}^{T} F^t(x^*). \tag{1.2}$$

Hence, the goal for the online convex optimization problems is to construct an algorithm, with which the generating decisions make us to achieve a regret as low as possible. We say that an algorithm is a no internal regret algorithm if the regret $R(T)$ is an infinitesimal of higher order than T [7]. We also note that the "no internal regret" is closely related to the "ergodic convergence" in the stochastic optimization. See Corollaries 4.3 and 4.7 for details.

The applications of the online programming are mostly built on a large scale. Some researchers have studied the performances of the gradient methods for the online convex optimization problems [24, 26]. When function $\psi(x)$ in (1.1) is an indicator function, Zinkevich [26] proved that the projected gradient method for an online convex optimization problem has a regret $O(\sqrt{T})$. When function $\psi(x)$ in (1.1) is a general regularization function, Xiao [24] proposed a dual averaging method, which is first proposed by Nesterov for the classical convex optimization problems. He showed that the proposed dual average method achieves the same regret $O(\sqrt{T})$. However, both of these two methods are full gradient methods, i.e., they update all components of the variable x on each iteration. When the scale of the problem becomes very large, the evaluations for updating the gradients of each iteration would take much time.

Recently, the "block" type methods are becoming very popular, especially for the large scale problems [18, 20, 21, 25]. On each iteration, the "block" type methods choose a subset $J \subseteq \{1, 2, \cdots, n\}$, and then update the components $\{x_j, j \in J\}$ by solving a low dimensional subproblem. Here, the set J is called the "block". Compared to the full gradient methods, the block type methods can reduce the calculation time on each iteration. Quite recently, Xu and Yin [25] proposed a block coordinate stochastic gradient method with the cyclic rule for a regularized stochastic optimization problem, which is related to the online optimization problem. Under the Lipschitz continuity assumption, they showed that the proposed method converges with $O(\frac{1+\log T}{\sqrt{1+T}}N)$, where N is the number of blocks. Furthermore, as the number of blocks reduces to 1, i.e., $N = 1$, this upper bound reduces to $O(\frac{1+\log T}{\sqrt{1+T}})$, which is bigger than the average regret $\frac{R(T)}{T} = O(\frac{1}{\sqrt{T}})$ of the greedy projection method [26].

In this paper, using the idea of "block" type methods, we propose two block coordinate gradient (BCG) methods for an online convex optimization problem with (1.1). These two methods are different in the rules of choosing blocks. One is the BCG method with the cyclic rule (C-BCG), the other is the BCG method with the random rule (R-BCG). For the proposed methods, we make our research on the following two aspects. Firstly, we establish their regret bounds, respectively. In particular, we show that the C-BCG method has a regret $O(\sqrt{T})$. Additionally, we find that the regret bound of the C-BCG method is independent of the number of blocks N under some assumptions, although we solve N subproblems at each time step.

For the R-BCG method, we prove that the expectation of the regret of the R-BCG method is $O(\sqrt{T})$. When the total number of blocks reduces to one, and function $\psi(x)$ is an indicator function, the regret of the proposed two methods reduce to the same result in [26]. Hence, the proposed methods in this paper can be regarded as natural extensions of the greedy projection method [26]. Secondly, we extend the proposed two methods to the corresponding convex stochastic optimization problems. Although the C-BCG method proposed in this paper is essentially same as the block coordinate stochastic gradient method with the cyclic rule [25], by different analysis, we show that the ergodic convergence upper bounds of the proposed two methods are tighter than that in [25].

Note that, quite recently, Dang and Lan [6] and Wang and Banerjee [23] independently gave similar results on the "block" type methods with the random rule for the online or stochastic optimization problems. The results of the R-BCG method in this paper has been mainly accomplished by one of the authors of this paper, in his master thesis [11], which has been submitted in March, 2013.

The paper is organized as follows. In Section 2, we introduce the algorithms of the block coordinate gradient methods with the cyclic rule and the random rule, respectively. Then we introduce some basic assumptions and present revelent properties in Section 3. In Section 4, we investigate the regret analysis for the proposed two methods, respectively. Finally, we conclude this paper in Section 5.

Throughout this paper, we use the following notations. When h is differentiable, the vector ∇h denotes the gradient of h. When h is nondifferentiable and convex, the set ∂h denotes the subdifferential of h. For a given vector $x \in \mathcal{R}^n$, $\|x\|$ denotes its standard 2-norm. The vector x_J denotes the subvector of x, consisting of $x_j,\ j \in J$ while vector $x_{\bar{J}}$ denotes the subvector consisting of $x_j,\ j \notin J$. The notation $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. In addition, for a vector $x \in \mathcal{R}^n$ and a symmetric positive definite matrix $G \in \mathcal{R}^{n \times n}$, the G-norm is defined as $\|x\|_G = \sqrt{\langle x, Gx \rangle}$.

2 Block Coordinate Gradient (BCG) methods for the online convex optimization problem

In this paper, we propose two block coordinate gradient (BCG) methods for the online convex optimization problem with (1.1), one is the BCG method with the cyclic rule, the other is the BCG method with the random rule. The BCG method with the cyclic rule is essentially same as the method proposed in [25] for stochastic optimization problems.

For convenience, we start with recalling several important results in [26], including the definition of the online convex programming problem and the greedy projection method.

Definition 2.1. *An online convex programming problem consists of a feasible set $\Omega \subseteq \mathcal{R}^n$ and an infinite sequence $\{F^1, F^2, \cdots\}$, where $F^t : \Omega \to \mathcal{R}$ is a convex function for each t. At each time step t, an online convex programming algorithm selects a vector $x^t \in \Omega$. After the vector is selected, it receives the loss function F^t.*

The loss functions $\{F^t(x), t = 1, 2 \cdots\}$ considered in this paper are defined by (1.1).

The greedy projection method [26], which is also called the projected gradient method, is described as follows.

Greedy projection method:

Step 0: Choose an initial point $x^1 \in \Omega$ and set a sequence of constants $\lambda_1, \lambda_2, \cdots > 0$.

Step 1: Update the vector x^t according to

$$x^{t+1} = x^t + d^t\,,$$
$$d^t = P_\Omega(x^t - \lambda_t \nabla f^t(x^t)) - x^t,$$

where $P_\Omega(\cdot)$ denotes a projection onto the set Ω.

Note that constants $\{\lambda_t, t = 1, 2, \cdots\}$ in the above greedy projection method are called the "learning rates". Before proposing the block coordinate gradient methods, we present several basic assumptions. Throughout this paper, the variable x is assumed to be partitioned into N blocks, i.e., $x^T = (x^T_{\mathcal{J}^1}, \cdots, x^T_{\mathcal{J}^N})$, where $\mathcal{J}^i$, $i \in \{1, 2, \cdots, N\}$, is a subset of $\{1, 2, \cdots, N\}$ such that

$$\mathcal{J}^i \cap \mathcal{J}^j = \varnothing, \forall i, j \in \{1, 2, \cdots, N\}, i \neq j;$$
$$\bigcup_{i=1}^{N} \mathcal{J}^i = \{1, 2, \cdots, n\}.$$

We also assume that function $\psi(x)$ is block separable with respect to the blocks $\{\mathcal{J}^i, i = 1, \cdots, N\}$, i.e., $\psi(x) = \sum_{i=1}^{N} \psi_i(x_{\mathcal{J}^i})$. For the set Ω in (1.1), we suppose that $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$, where operator "$\times$" denotes the Cartesian product. For given ∇f^t at time step t, we consider the direction, which is defined by

$$d(x; J, t, \lambda) = \operatorname*{argmin}_{d \in \mathcal{R}^n} \left\{ \langle \nabla f^t(x), d\rangle + \frac{1}{2\lambda}\|d\|^2 + \tau\psi(x + d) \,\middle|\, d_{\bar{J}} = 0, x + d \in \Omega \right\}. \quad (2.1)$$

When $J = \{1, 2, \cdots, n\}$ and $\psi(x) = 0$, direction $d(x; J, t, \lambda)$ reduces to the direction d^t in Step 1 of the greedy projection method. Note that the direction $d(x; J, t, \lambda)$ given by (2.1) is well defined, since the minimizer of the corresponding optimization problem always exists and

is unique [19, Theorem 31.5]. Moreover, the direction $d(x; J, t, \lambda)$ is a descent direction for the loss function $F^t(x)$.

The rule to choose blocks is also important for the convergence. In this paper, we adopt the following two rules.

(a) Cyclic rule [13]: Choose blocks in $\{\mathcal{J}^1, \mathcal{J}^2, \cdots, \mathcal{J}^N\}$ cyclically.

(b) Random rule [18]: Let $\{p_1, \cdots, p_N\}$ be a set of probabilities such that $p_i > 0$, $i = 1, 2, \cdots, N$, and $\sum_{i=1}^{N} p_i = 1$. On each iteration, we choose a block $\mathcal{J}^i \in \{\mathcal{J}^1, \mathcal{J}^2, \cdots, \mathcal{J}^N\}$ with probability p_i.

If there exists an element $i \in \{1, 2, \cdots, N\}$ such that $p_i = 0$ in the random rule, it means that the block $\mathcal{J}^i$ will never be updated. This case has no meaning. Hence, the assumption $p_i > 0$, $i = 1, 2, \cdots, N$, is a natural one.

Next, we describe the frameworks of the BCG methods with the cyclic rule and the random rule, respectively.

Algorithm 1. The BCG method with the cyclic rule (C-BCG):

Step 0: Choose an initial point $x^1 \in \operatorname{int} \Omega$. Let $t := 1$.

Step 1: If some termination condition holds, then stop. Otherwise go to Step 2.

Step 2: Let $x^{t,0} := x^t$ and $i := 1$. Go to Step 2-1.

Step 2-1: Set the learning rate $\lambda_{t,i} \in (0, +\infty)$. Solve the subproblem (2.1) with $x = x^{t,i-1}$, $J = \mathcal{J}^i$, $\lambda = \lambda_{t,i}$ and get a direction $d^{t,i} = d(x^{t,i-1}; \mathcal{J}^i, t, \lambda_{t,i})$.

Step 2-2: Set $x^{t,i} := x^{t,i-1} + d^{t,i}$. If $i = N$, then go to Step 3. Otherwise, let $i := i + 1$ and go to Step 2-1.

Step 3: Set $x^{t+1} := x^{t,N}$. Let $t := t + 1$ and go to Step 1.

Algorithm 2. The BCG method with the random rule (R-BCG):

Step 0: Choose an initial point $x^1 \in \operatorname{int} \Omega$. Let $t := 1$.

Step 1: If some termination condition holds, then stop. Otherwise go to Step 2.

Step 2: Choose a block $\mathcal{J}^i$ by the random rule and set the learning rate $\lambda_{t,i} \in (0, +\infty)$. Solve the subproblem (2.1) with $x = x^t$, $J = \mathcal{J}^i$, $\lambda = \lambda_{t,i}$ and get a direction $d^t = d(x^t; \mathcal{J}^i, t, \lambda_{t,i})$.

Step 3: Set $x^{t+1} := x^t + d^t$. Let $t := t + 1$, and go to Step 1.

At each time step, the C-BCG method solves subproblem (2.1) N times, while the R-BCG method solves subproblem (2.1) once. When $N = 1$ and function $\psi(x)$ is an indicator function, both of these two methods reduce to the greedy projection method [26].

3 Basic assumptions

In this section, we introduce several basic assumptions and present relevant properties, which will be used in the subsequent sections.

Given a constant $T > 0$, we denote the set of all optimal solutions of the problem $\min\limits_{x\in\Omega} \sum\limits_{t=1}^{T} F^t(x)$ by $X^{*,[T]}$ in the rest of this paper.

For the loss functions $f^t(x)$ and $\psi(x)$, we make the following assumptions, where Assumptions 3.1 and 3.2 are also used in [26].

Assumption 3.1. *The feasible set Ω for loss functions $\{F^t,\ t = 1, 2, \cdots\}$ in (1.1) is nonempty and compact.*

For convenience, we define

$$D = \max_{x,y\in\Omega} \|x - y\|. \tag{3.1}$$

It follows from Assumption 3.1 that $D < \infty$.

Assumption 3.2. *There exists a positive constant G such that $\|\nabla f^t(x)\| \le G$ and $\|\partial\psi(x)\| \le G$ hold for all $t > 0$ and $x \in \Omega$.*

Note that when $f^t(x)$ is a linear function, that is, $f^t(x) = \langle a^t, x\rangle + b^t$ with some $a^t \in \mathcal{R}^n$, $b^t \in \mathcal{R}$, we have $\nabla f^t(x) = a^t$. Then, $\|\nabla f^t(x)\| \le G$ means that $\|a^t\| \le G$ holds for any $t > 0$. When $f^t(x)$ is a quadratic function with $f^t(x) = \frac{1}{2}x^\top A^t x$, where $\{A^t \in \mathcal{R}^{n\times n}, t = 1, 2, \cdots\}$ are all symmetric and positive semidefinite, we get $\nabla f^t(x) = A^t x$. From Assumption 3.1, $\|\nabla f^t(x)\| \le G$ is equivalent to that $\|A^t\| \le \frac{G}{D}$. For function $\psi(x)$, when $\psi(x) = \|x\|_1$, we have $\|\partial\psi(x)\| \le \sqrt{n}$. When $\psi(x) = \|x\|_2$, we get $\|\partial\psi(x)\| \le 1$.

It is worth mentioning that Assumptions 3.1 and 3.2 are a little restrict in this paper. In fact, we only need to assume that there exists a compact set $\tilde{\Omega} \subseteq \mathcal{R}^n$ such that the iterations $\{x^t\} \subseteq \tilde{\Omega}$, $X^{*,[T]} \subseteq \tilde{\Omega}$, and that $\|\nabla f^t(x^t)\|$, $\|\partial\psi(x^t)\| \le G$ for all $x^t \in \tilde{\Omega}$. For simplicity, we adopt Assumptions 3.1 and 3.2 in this paper, which are in accord with the assumptions in [26].

The following assumption is a Lipschitz continuity-like assumption, which is originally proposed in [10].

Assumption 3.3. *The gradient ∇f^t is block lower triangular Lipschitz continuous with respect to blocks $\{\mathcal{J}^i,\ i = 1, 2, \cdots, N\}$ for any $t > 0$, i.e., there exists a nonnegative constant M such that, for any $t > 0$,*

$$\|g^t(x, y) - \nabla f^t(x)\| \le M\|y - x\|, \forall x, y \in \Omega, \tag{3.2}$$

where $g^t : \mathcal{R}^{n+n} \to \mathcal{R}^n$ *with*

$$g^t_{\mathcal{J}^i}(x, y) = \nabla_{\mathcal{J}^i} f^t(y_{\mathcal{J}^1}, \cdots, y_{\mathcal{J}^{i-1}}, x_{\mathcal{J}^i} \cdots, x_{\mathcal{J}^N}), \quad i = 1, \cdots, N. \tag{3.3}$$

In the following paper, we use the notation g^t instead of $g^t(x^t, x^{t+1})$ when it is clear from the context. The next remark illustrates the rationality of Assumption 3.3.

Remark 3.1. *When* $N = 1$ *or function* $f^t(x)$ *is separable with respect to the blocks* $\{\mathcal{J}^i, i = 1, \cdots, N\}$*, we have* $g^t(x, y) = \nabla f^t(x)$*, which yields that* $M = 0$ *in (3.2). When* $N > 1$*, it is shown in [10] that* $M \leq 2\max\{L_{f^1}, \cdots, L_{f^N}\}$ *holds for many classes of functions* $f^t(x)$*. For example, the Hessian matrix* $\nabla^2 f^t(x)$ *is tridiagonal or row diagonal dominant, it holds that* $M \leq 2L_f$*. For a general function* $f^t(x)$*, Hua and Yamashita [10] showed that* $M \leq \sqrt{N}L_f$ *in [10]. However, until now, we have not found an example where* $N^\sigma L \leq M$ *for a positive constant* σ*. Hence, Assumption 3.3 is a reasonable assumption.*

Under Assumptions 3.1-3.3, we can show that the vector g^t is bounded for all $t > 0$.

Lemma 3.1. *Suppose that Assumptions 3.1-3.3 hold. Then we have*

$$\|g^t\|^2 = \sum_{i=1}^{N} \|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 \leq \bar{G}^2, \tag{3.4}$$

where $\bar{G} = MD + G$*. Moreover, when* $N = 1$ *or function* $f^t(x)$ *is separable with respect to the blocks* $\{\mathcal{J}^i, i = 1, \cdots, N\}$*, we have that* $\|g^t\| \leq G$*.*

Proof. Since we denote $g^t = g^t(x^t, x^{t+1})$, from the definition of $g^t(x, y)$ in (3.3), we have that

$$\|g^t\|^2 = \sum_{i=1}^{N} \|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2.$$

Moreover, from Assumptions 3.1-3.3, we have that

$$\|g^t\| \leq \|g^t - \nabla f^t(x^t)\| + \|\nabla f^t(x^t)\| \leq M\|x^{t+1} - x^t\| + G \leq MD + G.$$

Hence, the relation (3.4) holds.

When $N = 1$ or function $f^t(x)$ is separable with respect to the blocks $\{\mathcal{J}^i, i = 1, \cdots, N\}$, we have $\|g^t\| = \|\nabla f^t(x^t)\|$ from Remark 3.1, which together with Assumption 3.2 yields that $\|g^t\| \leq G$. □

For the learning rate λ in the proposed methods, we make the following assumption.

Assumption 3.4. *The learning rate* $\lambda_{t,i}$ *in the C-BCG method or the R-BCG method is given by* $\lambda_{t,i} = \frac{c\beta_i}{\sqrt{t}}$*,* $t = 1, 2, \cdots$*,* $i = 1, 2, \cdots, N$*, where* $c > 0$*, and* $\beta_i \in [\underline{\beta}, \bar{\beta}]$*,* $\bar{\beta} \geq \underline{\beta} > 0$*.*

The constants $\{\beta_i, i = 1, 2, \cdots, N\}$ in Assumption 3.4 act as scaling factors of the learning rates on different blocks. When $\beta_1 = \cdots = \beta_N = 1$ and $c = 1$, we have $\lambda_{t,i} = \frac{1}{\sqrt{t}}$, which reduces to the case in [26].

Next, we recall the regret of the greedy projection method, which is given in [26].

Theorem 3.2. *Suppose that Assumptions 3.1-3.2 hold. Let $\lambda_t = \frac{1}{\sqrt{t}}$ and $x^{*,[T]} \in X^{*,[T]}$. Then, the regret $R(T)$ of the greedy projection method satisfies*

$$R(T) \leq \frac{\sqrt{T}}{2} D^2 + \frac{(2\sqrt{T} - 1)}{2} G^2, \tag{3.5}$$

where constant D is defined by (3.1), and constant G is given in Assumption 3.2.

The following lemma is a well known result for the proximal like algorithm, which is given in [22, Property 1].

Lemma 3.2. *For any proper l.s.c. convex function $\varphi : \Omega \to (-\infty, \infty]$ and any $z \in \Omega$, if*

$$z_+ = \operatorname*{argmin}_{x \in \Omega} \left\{ \varphi(x) + \frac{1}{2\lambda} \|x - z\|^2 \right\},$$

then we have

$$\varphi(x) + \frac{1}{2\lambda} \|x - z\|^2 \geq \varphi(z_+) + \frac{1}{2\lambda} \|z_+ - z\|^2 + \frac{1}{2\lambda} \|x - z_+\|^2, \forall x \in \Omega.$$

4 Regret analysis

In this section, we give the regret analysis for the proposed two methods, the C-BCG method and the R-BCG method, respectively.

4.1 Regret of the BCG method with the cyclic rule

In this subsection, we present the regret analysis of the C-BCG method for the online convex optimization problem with (1.1). Throughout this subsection, the sequence $\{x^t\}$ denotes the sequence generated by the C-BCG method.

We first introduce several technical lemmas. The following lemma presents main characteristics of the sequence $\{x^t\}$, which can be verified easily, and hence, we omit its proof here.

Lemma 4.1. *For the sequence $\{x^t\}$, we have*

$$\begin{aligned}
&x^t_{\mathcal{J}^i} = x^{t,0}_{\mathcal{J}^i} = x^{t,j}_{\mathcal{J}^i}, \ \ \forall\, i = 1, 2, \cdots, N, \ 1 \leq j < i. \\
&x^{t+1}_{\mathcal{J}^i} = x^{t,N}_{\mathcal{J}^i} = x^{t,j}_{\mathcal{J}^i}, \ \ \forall\, i = 1, 2, \cdots, N, \ i \leq j \leq N. \\
&x^{t+1}_{\mathcal{J}^i} - x^t_{\mathcal{J}^i} = x^{t,N}_{\mathcal{J}^i} - x^{t,0}_{\mathcal{J}^i} = d^{t,i}_{\mathcal{J}^i}, \ \ \forall\, i = 1, 2, \cdots, N.
\end{aligned}$$

The following lemma states that each movement $\|x^{t+1}-x^t\|$ is closely related to the learning rate defined in Assumption 3.4.

Lemma 4.2. *Suppose that Assumptions 3.1-3.4 hold, and suppose that function $\psi(x)$ in (1.1) is an indicator function with respect to the set Ω. Then, for any $t > 0$, we have*

$$\|x^{t+1} - x^t\| \leq \frac{c\tilde{G}}{\sqrt{t}},$$

where $\tilde{G} = \bar{\beta}\sqrt{2\bar{G}^2 + 2\tau^2 G^2}$ and $\bar{G} = MD + G$.

Proof. Since function $\psi(x)$ in (1.1) is block separable, from (2.1), the subvector $d^{t,i}_{\mathcal{J}^i}$ can be rewritten as

$$d^{t,i}_{\mathcal{J}^i} = \underset{x^{t,i-1}_{\mathcal{J}^i}+d_{\mathcal{J}^i}\in\Omega_i}{\operatorname{argmin}} \left\{ \frac{1}{2\lambda_{t,i}}\|d_{\mathcal{J}^i} + \lambda_{t,i}\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 + \tau\psi_i(x^{t,i-1}_{\mathcal{J}^i} + d_{\mathcal{J}^i}) \right\}. \tag{4.1}$$

From the first order optimality condition, we have that

$$\langle \frac{1}{\lambda_{t,i}} d^{t,i}_{\mathcal{J}^i} + \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau\eta^{t,i}_{\mathcal{J}^i}, d_{\mathcal{J}^i} - d^{t,i}_{\mathcal{J}^i}\rangle \geq 0, \ \forall d_{\mathcal{J}^i} \text{ such that } x^{t,i-1}_{\mathcal{J}^i} + d_{\mathcal{J}^i} \in \Omega_i, \tag{4.2}$$

where $\eta^{t,i}_{\mathcal{J}^i} \in \partial\psi_i(x^{t,i}_{\mathcal{J}^i})$. Since $x^{t,i-1}_{\mathcal{J}^i} \in \Omega_i$, we let $d_{\mathcal{J}^i} = 0$ in (4.2) and get

$$\langle \frac{1}{\lambda_{t,i}} d^{t,i}_{\mathcal{J}^i} + \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau\eta^{t,i}_{\mathcal{J}^i}, -d^{t,i}_{\mathcal{J}^i}\rangle \geq 0, \tag{4.3}$$

which implies that

$$\|d^{t,i}_{\mathcal{J}^i}\|^2 \leq \lambda_{t,i}\langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau\eta^{t,i}_{\mathcal{J}^i}, -d^{t,i}_{\mathcal{J}^i}\rangle \leq \lambda_{t,i}\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) + \tau\eta^{t,i}_{\mathcal{J}^i}\|\|d^{t,i}_{\mathcal{J}^i}\|.$$

Dividing by $\|d^{t,i}_{\mathcal{J}^i}\|$ on both sides and squaring it, we get

$$\begin{aligned}\|d^{t,i}_{\mathcal{J}^i}\|^2 &\leq \lambda^2_{t,i}\left(\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\| + \tau\|\eta^{t,i}_{\mathcal{J}^i}\|\right)^2 \\ &\leq 2\frac{c^2\beta_i^2}{t}\left(\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 + \tau^2\|\eta^{t,i}_{\mathcal{J}^i}\|^2\right).\end{aligned}$$

Summing this inequality over i from 1 to N, we obtain

$$\|x^{t+1} - x^t\|^2 = \sum_{i=1}^N \|d^{t,i}_{\mathcal{J}^i}\|^2 \leq 2\frac{c^2\bar{\beta}^2}{t}\left(\sum_{i=1}^N \|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 + \tau^2\sum_{i=1}^N \|\eta^{t,i}_{\mathcal{J}^i}\|^2\right). \tag{4.4}$$

Since $(\eta^{t,1}_{\mathcal{J}^1}, \cdots, \eta^{t,N}_{\mathcal{J}^N}) \in \partial\psi(x^{t+1})$, it follows from Assumption 3.2 that $\sum_{i=1}^N \|\eta^{t,i}_{\mathcal{J}^i}\|^2 \leq G^2$, which together with Lemma 3.1 and (4.4) proves the desired result. □

The next result presents an estimator between $F^t(x^t)$ and $F^t(x)$, which plays a key role for the final regret analysis.

Lemma 4.3. *For any $t > 0$, we have*

$$F^t(x^t) - F^t(x) \leq S^{1,t} + S^{2,t}(x) + S^{3,t}(x),$$

where $S^{1,t}$, $S^{2,t}(x)$ and $S^{3,t}(x)$ are defined as follows.

$$S^{1,t} = \sum_{i=1}^{N} \left[-\frac{1}{2\lambda_{t,i}} \|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle \right]; \tag{4.5}$$

$$S^{2,t}(x) = \sum_{i=1}^{N} \frac{1}{2\lambda_{t,i}} \left[\|x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \|x_{\mathcal{J}^i} - x^{t,i}_{\mathcal{J}^i}\|^2 \right]; \tag{4.6}$$

$$S^{3,t}(x) = \sum_{i=1}^{N} \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) - \nabla_{\mathcal{J}^i} f^t(x^t), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle + \tau[\psi(x^t) - \psi(x^{t+1})]. \tag{4.7}$$

Proof. Using Lemma 3.2 with $\varphi(x) = \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle + \tau\psi_i(x_{\mathcal{J}^i})$, $\lambda = \lambda_{t,i}$, $z_+ = x^{t,i}_{\mathcal{J}^i}$, $z = x^{t,i-1}_{\mathcal{J}^i}$, $i \in \{1, \cdots, N\}$, we have

$$\begin{aligned} \tau\psi_i(x^{t,i}_{\mathcal{J}^i}) - \tau\psi_i(x_{\mathcal{J}^i}) \leq & \frac{1}{2\lambda_{t,i}} \|x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \frac{1}{2\lambda_{t,i}} \|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \frac{1}{2\lambda_{t,i}} \|x_{\mathcal{J}^i} - x^{t,i}_{\mathcal{J}^i}\|^2 \\ & + \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle - \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} \rangle. \end{aligned} \tag{4.8}$$

Moreover, from the convexity of function $f^t(x)$, we obtain

$$f^t(x^t) - f^t(x) \leq -\langle x - x^t, \nabla f^t(x^t) \rangle. \tag{4.9}$$

Then we have

$$\begin{aligned} & F^t(x^t) - F^t(x) \\ & = f^t(x^t) - f^t(x) + \tau[\psi(x^t) - \psi(x)] \\ & \leq -\langle x - x^t, \nabla f^t(x^t) \rangle + \tau[\psi(x^{t+1}) - \psi(x)] + \tau[\psi(x^t) - \psi(x^{t+1})] \\ & = \sum_{i=1}^{N} \left\{ -\langle x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}, \nabla_{\mathcal{J}^i} f^t(x^t) \rangle + \tau[\psi_i(x^{t,i}_{\mathcal{J}^i}) - \psi_i(x_{\mathcal{J}^i})] \right\} + \tau[\psi(x^t) - \psi(x^{t+1})], \end{aligned}$$

where the inequality follows from (4.9), and the last equality follows from Lemma 4.1. Combining with inequality (4.8), we obtain the desired inequality. □

Next, we establish upper bounds for $S^{1,t}$, $S^{2,t}(x)$ and $S^{3,t}(x)$, respectively.

Lemma 4.4. *Suppose that Assumptions 3.1-3.4 hold. Then, for any $t > 0$, we get*

$$S^{1,t} \leq \frac{c\bar{\beta}}{2\sqrt{t}} \bar{G}^2.$$

Proof. For any $t > 0$ and $i \in \{1, \cdots, N\}$, we have

$$\begin{aligned}
&-\frac{1}{2\lambda_{t,i}}\|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\|^2 - \langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}), x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\rangle \\
&= -\frac{1}{2\lambda_{t,i}}\left\|x^{t,i}_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i} + \lambda_{t,i}\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\right\|^2 + \frac{\lambda_{t,i}}{2}\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 \\
&\leq \frac{\lambda_{t,i}}{2}\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2.
\end{aligned}$$

Summing this inequality over i from 1 to N, we get

$$S^{1,t} \leq \sum_{i=1}^{N}\frac{\lambda_{t,i}}{2}\|\nabla_{\mathcal{J}^i} f^t(x^{t,i-1})\|^2 \leq \frac{c\bar{\beta}}{2\sqrt{t}}\|g^t\|^2 \leq \frac{c\bar{\beta}}{2\sqrt{t}}\bar{G}^2,$$

where the last two inequalities follow from Lemma 3.1. □

For convenience, we define a diagonal matrix $B \in \mathcal{R}^{n\times n}$ with

$$B_{jj} = \beta_i,\ \forall j \in \mathcal{J}^i, i = 1, 2, \cdots, N, \tag{4.10}$$

where $\{\beta_i,\ i = 1, 2, \cdots, N\}$ are constants given in Assumption 3.4. Since we assume $\beta_i \geq \underline{\beta} > 0$ for any $i \in \{1, \cdots, N\}$, matrix B is invertible.

For $S^{2,t}(x)$, it follows from (4.6), Assumption 3.4, and the definition of the norm $\|\cdot\|_{B^{-1}}$ that

$$S^{2,t}(x) = \frac{\sqrt{t}}{2c}\left[\|x - x^t\|^2_{B^{-1}} - \|x - x^{t+1}\|^2_{B^{-1}}\right]. \tag{4.11}$$

A bound for $S^{3,t}(x)$ is given by the following lemma.

Lemma 4.5. *Suppose that Assumptions 3.1-3.4 hold. Then, for any $t > 0$, we have*

$$S^{3,t}(x) \leq \frac{c}{\sqrt{t}}M\tilde{G}\|x - x^t\| + \tau[\psi(x^t) - \psi(x^{t+1})],$$

where $\tilde{G} = \bar{\beta}\sqrt{2\bar{G}^2 + 2\tau^2 G^2}$ and $\bar{G} = MD + G$.

Proof. In contrast to (4.7), we only need to show that $\sum_{i=1}^{N}\langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) - \nabla_{\mathcal{J}^i} f^t(x^t), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\rangle \leq \frac{c}{\sqrt{t}}M\tilde{G}\|x - x^t\|$ holds for any $t > 0$. In fact, we have

$$\begin{aligned}
&\sum_{i=1}^{N}\langle \nabla_{\mathcal{J}^i} f^t(x^{t,i-1}) - \nabla_{\mathcal{J}^i} f^t(x^t), x_{\mathcal{J}^i} - x^{t,i-1}_{\mathcal{J}^i}\rangle \\
&= \langle g^t - \nabla f^t(x^t), x - x^t\rangle \\
&\leq \|g^t - \nabla f^t(x^t)\|\|x - x^t\| \\
&\leq M\|x^{t+1} - x^t\|\|x - x^t\|
\end{aligned}$$

$$\leq M\frac{c}{\sqrt{t}}\tilde{G}\|x - x^t\|,$$

where the second inequality follows from Assumption 3.3, and the last inequality follows from Lemma 4.2. Thus, this completes the proof. □

Now we show the regret of the BCG method with the cyclic rule for the online convex optimization problem with (1.1).

Theorem 4.1. *Suppose that Assumptions 3.1-3.4 hold. Let $\{x^t\}$ be generated by the C-BCG method for the online optimization problem with (1.1). Then, for any $x^{*,[T]} \in X^{*,[T]}$, we have*

$$R(T) \leq \left(\frac{c\bar{\beta}\bar{G}^2}{2} + \frac{D^2}{2c\underline{\beta}} + M\tilde{G}Dc\right)(2\sqrt{T} - 1) + \tau DG,$$

where $\tilde{G} = \bar{\beta}\sqrt{2\bar{G}^2 + 2\tau^2 G^2}$ and $\bar{G} = MD + G$.

Proof. It follows from Lemma 4.3 that

$$R(T) = \sum_{t=1}^{T}\{F^t(x) - F^t(x^{*,[T]})\} = \sum_{t=1}^{T} S^{1,t} + \sum_{t=1}^{T} S^{2,t}(x^{*,[T]}) + \sum_{t=1}^{T} S^{3,t}(x^{*,[T]}).$$

Moreover, we have

$$\sum_{t=1}^{T}\frac{c}{\sqrt{t}} \leq c + c\int_{t=1}^{T}\frac{dt}{\sqrt{t}} \leq c + 2c\sqrt{T} - 2c = c(2\sqrt{T} - 1), \tag{4.12}$$

which, together with Lemma 4.4, yields that

$$\sum_{t=1}^{T} S^{1,t} \leq \sum_{t=1}^{T}\frac{c\bar{\beta}}{2\sqrt{t}}\bar{G}^2 = \frac{c\bar{\beta}\bar{G}^2}{2}(2\sqrt{T} - 1).$$

For $S^{2,t}(x^{*,[T]})$, it follows from (4.11) that

$$\begin{aligned}
\sum_{t=1}^{T} S^{2,t}(x^{*,[T]}) =& \frac{1}{2c}\|x^{*,[T]} - x^1\|_{B^{-1}}^2 - \frac{\sqrt{T}}{2c}\|x^{*,[T]} - x^{T+1}\|_{B^{-1}}^2 \\
&+ \sum_{t=2}^{T}\left(\frac{\sqrt{t}}{2c} - \frac{\sqrt{t-1}}{2c}\right)\|x^{*,[T]} - x^t\|_{B^{-1}}^2 \\
\leq& \frac{1}{2c\underline{\beta}}D^2 + \sum_{t=2}^{T}\left(\frac{\sqrt{t}}{2c} - \frac{\sqrt{t-1}}{2c}\right)\frac{1}{\underline{\beta}}D^2 \\
=& \frac{D^2}{2c\underline{\beta}}\sqrt{T}.
\end{aligned}$$

Let $\eta \in \partial\psi(x^1)$. Then it follows from Assumption 3.2 that $\|\eta\| \leq G$. For $S^{3,t}(x^{*,[T]})$, from Lemma 4.5, we have

$$\sum_{t=1}^{T} S^{3,t}(x^{*,[T]}) \leq \sum_{t=1}^{T}\left\{\frac{c}{\sqrt{t}}M\tilde{G}\|x^{*,[T]} - x^t\| + \tau[\psi(x^t) - \psi(x^{t+1})]\right\}$$

$$
\begin{aligned}
=&\sum_{t=1}^{T}\frac{c}{\sqrt{t}}M\tilde{G}\|x^{*,[T]}-x^t\|+\tau[\psi(x^1)-\psi(x^{T+1})]\\
\leq&M\tilde{G}D\sum_{t=1}^{T}\frac{c}{\sqrt{t}}+\tau\langle\eta,x^1-x^{T+1}\rangle\\
\leq&M\tilde{G}Dc(2\sqrt{T}-1)+\tau DG, \qquad (4.13)
\end{aligned}
$$

where the second inequality follows from Assumption 3.1 and the convexity of function $\psi(x)$, and the last inequality follows from (4.12) and Assumptions 3.1-3.2.

Hence, we get

$$
\begin{aligned}
R(T)&\leq\frac{c\bar{\beta}\bar{G}^2}{2}(2\sqrt{T}-1)+\frac{D^2}{2c\underline{\beta}}\sqrt{T}+M\tilde{G}Dc(2\sqrt{T}-1)+\tau DG\\
&\leq\left(\frac{c\bar{\beta}\bar{G}^2}{2}+\frac{D^2}{2c\underline{\beta}}+M\tilde{G}Dc\right)(2\sqrt{T}-1)+\tau DG,
\end{aligned}
$$

where the second inequality follows from the fact $\sqrt{T}\leq 2\sqrt{T}-1, T\geq 1$. □

Note that the regret bound of the C-BCG method in Theorem 4.1 is independent of the number of blocks N and the dimension n. Moreover, Theorem 4.1 implies that $\frac{R(T)}{T}\leq O(\frac{1}{\sqrt{T}})$, and hence, the C-BCG method is a no internal regret algorithm.

The next remark states that the regret bound of the C-BCG method in Theorem 4.1 is an extension of the bound of the greedy projection method [26].

Remark 4.2. *When function $\psi(x)=0$, the evaluation for $\sum_{t=1}^{T}S^{3,t}(x^{*,[T]})$ in (4.13) reduces to $\sum_{t=1}^{T}S^{3,t}(x^{*,[T]})\leq M\tilde{G}Dc(2\sqrt{T}-1)$. Moreover, if we let $N=1$, $\beta_1=1$, and $c=1$, from Lemma 3.1 and Remark 3.1, we have $\bar{G}=G$ and $M=0$. Hence, the regret of the C-BCG method reduces to $R(T)\leq\frac{G^2}{2}(2\sqrt{T}-1)+\frac{D^2}{2}\sqrt{T}$, which is the same as Theorem 3.2. Therefore, Theorem 4.1 is a natural extension of Theorem 3.2 for the greedy projection method.*

The result in Theorem 4.1 for the C-BCG method can be extended to the following convex stochastic optimization problem with a separable structure.

$$
\underset{x}{\text{minimize}}\,\tilde{F}(x):=E_z[f(x,z)]+\tau\psi(x), \qquad (4.14)
$$

where $z=(u,v)\in\mathcal{R}^{n+n}$ is an input-output pair of the data drawn from an unknown underlying distribution, $f(x,z)$ is the loss function of using u with parameter x to predict v, $E_z[f(x,z)]$ denotes the expected value of the loss function $f(x,z)$ with respect to the selection pair z. Function $\psi(x)$ in (4.14) is assumed to have the same block separable property as that in (1.1).

An usual way to solve stochastic optimization problem (4.14) is to approximate the expectation of the whole loss $E_z[f(x,z)]$ by using a finite set of independent observations $z_1, \cdots, z_t$, and solve the following problem instead.

$$\underset{x}{\text{minimize}} \frac{1}{T}\sum_{t=1}^{T} f(x,z_t) + \tau\psi(x). \tag{4.15}$$

Problem (4.15) can be regarded as the corresponding batch optimization problem for the online optimization problem with $F^t(x) = f(x,z_t)+\tau\psi(x)$. Let $x^* \in \underset{x}{\operatorname{argmin}} \tilde{F}(x)$. The corresponding regret can be written as follows.

$$R(T) = \sum_{t=1}^{T}\big\{f(x^t,z_t) + \tau\psi(x^t)\big\} - \sum_{t=1}^{T}\big\{f(x^*,z_t) + \tau\psi(x^*)\big\}. \tag{4.16}$$

Corollary 4.3. *Suppose that Assumptions 3.1-3.4 hold. Let $\{x^t\}$ be generated by the C-BCG method for the convex stochastic optimization problem (4.14), and let $\bar{x}^T = \dfrac{1}{T}\sum_{t=1}^{T} x^t$, $T \geq 1$. Then, for any $x^* \in \underset{x}{\operatorname{argmin}} \tilde{F}(x)$, we have*

$$E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq \left(\frac{c\bar{\beta}\bar{G}^2}{2} + \frac{D^2}{2c\underline{\beta}} + M\tilde{G}Dc\right)\frac{2\sqrt{T}-1}{T} + \frac{1}{T}\tau DG.$$

Proof. Let $z_{[T]} = \{z_1, \cdots, z_T\}$, where $\{z_i,\, 1 \leq i \leq T\}$ follow an independent and isotonical distribution. Note that the variables $\{x^t,\, 1 \leq t \leq T\}$ are dependent on the random variables $\{z_1, \cdots, z_{t-1}\}$, but independent on the random variables $\{z_t, \cdots, z_T\}$.

Hence, we have

$$E_{z_{[T]}}[f(x^t,z_t) + \tau\psi(x^t)] = E_{z_{[t-1]}}\big[E_{z_{[t]}}[f(x^t,z_t)] + \tau\psi(x^t)\big] = E_{z_{[t-1]}}[\tilde{F}(x^t)] = E_{z_{[T]}}[\tilde{F}(x^t)];$$
$$E_{z_{[T]}}[f(x^*,z_t) + \tau\psi(x^*)] = E_{z_{[t]}}[f(x^*,z_t)] + \tau\psi(x^*) = \tilde{F}(x^*).$$

Combining these two relations with (4.16), we get

$$0 \leq E_{z_{[T]}}[R(T)] = \sum_{t=1}^{T}\Big(E_{z_{[T]}}[\tilde{F}(x^t)] - \tilde{F}(x^*)\Big). \tag{4.17}$$

On the other hand, by the convexity of function $\tilde{F}(x)$, we have

$$\tilde{F}(\bar{x}^T) = \tilde{F}(\frac{1}{T}\sum_{t=1}^{T} x^t) \leq \frac{1}{T}\sum_{t=1}^{T}\tilde{F}(x^t). \tag{4.18}$$

Subtracting the optimal value $\tilde{F}(x^*)$ on both sides and taking the expectation with respect to the random variable $z \in z_{[T]}$, we get

$$E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\Big(E_{z_{[T]}}[\tilde{F}(x^t)] - \tilde{F}(x^*)\Big) = \frac{1}{T}E_{z_{[T]}}R(T).$$

From Theorem 4.1, we prove the desired result. □

Note that Corollary 4.3 implies that $E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq O(\frac{1}{\sqrt{T}})$, where the upper bound $O(\frac{1}{\sqrt{T}})$ is sharper than $O(\frac{1+\log T}{\sqrt{1+T}}N)$ given in [25].

4.2 Regret of the BCG method with the random rule

In this subsection, we give a regret bound for the R-BCG method for the online optimization problem with (1.1). Throughout this subsection, the sequence $\{x^t\}$ denotes the sequence generated by the R-BCG method. The results in this subsection is mainly based on the master thesis [11] by one of the authors, which is carried out independently of [6, 23].

Since we choose blocks in the R-BCG method randomly, the sequence $\{x^t\}$ consists of random variables. Therefore, we will evaluate the expected value of the regret.

For any block $\mathcal{J}^i$, $i \in \{1, \cdots, N\}$, we denote

$$\tilde{x}^{t+1}_{\mathcal{J}^i} = x^t_{\mathcal{J}^i} + d^t_{\mathcal{J}^i}. \tag{4.19}$$

For simplicity, we denote the vector $((\tilde{x}^{t+1}_{\mathcal{J}^i})^T, (x^t_{\overline{\mathcal{J}^i}})^T)^T$ by $(\tilde{x}^{t+1}_{\mathcal{J}^i}, x^t_{\overline{\mathcal{J}^i}})$ when it is clear from the context. Given a vector x^t, we denote the expectation of x^{t+1} with respect to x^t by $E[x^{t+1} \mid x^t]$, i.e.,

$$E[x^{t+1} \mid x^t] = \sum_{i=1}^{N} p_i(\tilde{x}^{t+1}_{\mathcal{J}^i}, x^t_{\overline{\mathcal{J}^i}}). \tag{4.20}$$

Then we have

$$E[x^{t+1}_{\mathcal{J}^i} \mid x^t] = p_i \tilde{x}^{t+1}_{\mathcal{J}^i} + (1-p_i)x^t_{\mathcal{J}^i},\ i = 1, 2, \cdots, N. \tag{4.21}$$

$$E[f^{t+1}(x^{t+1}) \mid x^t] = \sum_{i=1}^{N} p_i f^{t+1}(\tilde{x}^{t+1}_{\mathcal{J}^i}, x^t_{\overline{\mathcal{J}^i}}). \tag{4.22}$$

For any $t \geq 2$, we let

$$\xi_{[t-1]} = \{J^1, \ldots, J^{t-1}\},$$

where J^i, $i \in \{1, \cdots, t-1\}$, denotes the block we choose at the time step i. Obviously, $\xi_{[t-1]}$ is a set of random variables. We denote the expectation of $F(x^t)$ with respect to $\xi_{[t-1]}$ by $E_{\xi_{[t-1]}}[F(x^t)]$. Then, we have

$$E_{\xi_{[T]}}[F(x^t)] = E_{\xi_{[t-1]}}[F(x^t)], \quad \forall\, t > 0 \text{ such that } T \geq t-1. \tag{4.23}$$

Note that

$$E_{\xi_{[0]}}[F(x^1)] = F(x^1). \tag{4.24}$$

By an analogous argument to Lemma 4.3, we obtain the following important estimator for $F^t(x^t) - F^t(x)$, whose proof is omitted here.

Lemma 4.6. *For the sequence $\{x^t\}$, we have*

$$F^t(x^t) - F^t(x) \leq W^{1,t} + W^{2,t}(x) + W^{3,t},$$

where $W^{1,t}, W^{2,t}(x)$ and $W^{3,t}$ are defined as follows.

$$W^{1,t} = \sum_{i=1}^{N} \left[-\frac{1}{2\lambda_{t,i}} \|\tilde{x}^{t+1}_{\mathcal{J}^i} - x^t_{\mathcal{J}^i}\|^2 - \langle \nabla_{\mathcal{J}^i} f^t(x^t), \tilde{x}^{t+1}_{\mathcal{J}^i} - x^t_{\mathcal{J}^i} \rangle \right]; \tag{4.25}$$

$$W^{2,t}(x) = \sum_{i=1}^{N} \frac{1}{2\lambda_{t,i}} \left[\|x_{\mathcal{J}^i} - x^t_{\mathcal{J}^i}\|^2 - \|x_{\mathcal{J}^i} - \tilde{x}^{t+1}_{\mathcal{J}^i}\|^2 \right]; \tag{4.26}$$

$$W^{3,t} = \sum_{i=1}^{N} \left[\tau\psi_i(x^t_{\mathcal{J}^i}) - \tau\psi_i(\tilde{x}^{t+1}_{\mathcal{J}^i}) \right]. \tag{4.27}$$

Next, we establish the upper bounds for these three items $W^{1,t}, W^{2,t}(x)$ and $W^{3,t}$, respectively. The following lemma gives an upper bound for $W^{1,t}$, which can be proven by a similar way to Lemma 4.4. For simplicity, we omit the proof here.

Lemma 4.7. *Suppose that Assumptions 3.2 and 3.4 hold. Then, for any $t > 0$, we get*

$$W^{1,t} \leq \frac{c\bar{\beta}}{2\sqrt{t}} G^2.$$

For convenience, we define a diagonal matrix $P \in \mathcal{R}^{n\times n}$ with

$$P_{jj} = p_i, \ \forall j \in \mathcal{J}^i, i = 1, 2, \cdots, N. \tag{4.28}$$

We let

$$\underline{p} = \min\{p_1, \cdots, p_N\}.$$

Since we assume that $\underline{p} > 0$ in the random rule, matrix P is also invertible. For the term $W^{2,t}(x)$, it can be reformulated as follows.

Lemma 4.8. *For any $t > 0$, we get*

$$W^{2,t}(x) = \frac{c}{2\sqrt{t}} \left[\|x - x^t\|^2_{B^{-1}P^{-1}} - E[\|x - x^{t+1}\|^2_{B^{-1}P^{-1}} \mid x^t] \right].$$

Proof. By the definition of norm $\|\cdot\|_{B^{-1}P^{-1}}$, we have

$$\|x - x^{t+1}\|^2_{B^{-1}P^{-1}} = \sum_{i=1}^{N} \frac{1}{p_i\beta_i} \|x_{\mathcal{J}^i} - x^{t+1}_{\mathcal{J}^i}\|^2. \tag{4.29}$$

Moreover, it follows from the property of the expectation that

$$E[\|x_{\mathcal{J}^i} - x^{t+1}_{\mathcal{J}^i}\|^2_{B^{-1}P^{-1}} \mid x^t] = p_i \|x_{\mathcal{J}^i} - \tilde{x}^{t+1}_{\mathcal{J}^i}\|^2_{B^{-1}P^{-1}} + (1 - p_i)\|x_{\mathcal{J}^i} - x^t_{\mathcal{J}^i}\|^2_{B^{-1}P^{-1}}. \tag{4.30}$$

Hence, we get

$$\begin{aligned} E[\|x - x^{t+1}\|^2_{B^{-1}P^{-1}} \mid x^t] &= E[\sum_{i=1}^{N} \frac{1}{\beta_i p_i}\|x_{\mathcal{J}^i} - x^{t+1}_{\mathcal{J}^i}\|^2 \mid x^t] \\ &= \sum_{i=1}^{N} \frac{1}{\beta_i p_i}\left(p_i\|x_{\mathcal{J}^i} - \tilde{x}^{t+1}_{\mathcal{J}^i}\|^2 + (1-p_i)\|x_{\mathcal{J}^i} - x^t_{\mathcal{J}^i}\|^2\right) \\ &= \sum_{i=1}^{N} \frac{1}{\beta_i}\left(\|x_{\mathcal{J}^i} - \tilde{x}^{t+1}_{\mathcal{J}^i}\|^2 - \|x_{\mathcal{J}^i} - x^t_{\mathcal{J}^i}\|^2\right) + \|x - x^t\|^2_{B^{-1}P^{-1}}, \end{aligned}$$

where the first and last equalities follow from (4.29), and the second equality follows from (4.30). Combining with (4.26), we prove the desired result. □

The following lemma states the reformulation of $W^{3,t}$.

Lemma 4.9. *For any $t > 0$, we get*

$$W^{3,t} = \sum_{i=1}^{N} \frac{1}{p_i}\left[\tau\psi_i(x^t_{\mathcal{J}^i}) - \tau E[\psi_i(x^{t+1}_{\mathcal{J}^i}) \mid x^t]\right].$$

Proof. By (4.21), we have

$$E[\psi_i(x^{t+1}_{\mathcal{J}^i}) \mid x^t] = p_i\psi_i(\tilde{x}^{t+1}_{\mathcal{J}^i}) + (1-p_i)\psi_i(x^t_{\mathcal{J}^i}),$$

that is,

$$\tau\psi_i(x^t_{\mathcal{J}^i}) - \tau\psi_i(\tilde{x}^{t+1}_{\mathcal{J}^i}) = \frac{1}{p_i}\tau\psi_i(x^t_{\mathcal{J}^i}) - \frac{1}{p_i}\tau E[\psi_i(x^{t+1}_{\mathcal{J}^i}) \mid x^t].$$

Hence, this lemma holds. □

Now, we can show the regret of the R-BCG method for the online optimization problem with (1.1).

Theorem 4.4. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Let $\{x^t\}$ be generated by the R-BCG method for the online optimization problem with (1.1). Then, for any $x^{*,[T]} \in X^{*,[T]}$, we have*

$$E_{\xi_{[T-1]}}[R(T)] < \frac{\tau}{2\underline{p}}(D^2 + G^2) + (\frac{\bar{\beta}c\bar{G}^2}{2} + \frac{D^2}{2\underline{\beta}c\underline{p}})(2\sqrt{T} - 1).$$

Proof. It follows from Lemma 4.6 that

$$\begin{aligned} E_{\xi_{[T-1]}}[R(T)] &= E_{\xi_{[T-1]}}[\sum_{t=1}^{T}\{F^t(x) - F^t(x^{*,[T]})\}] \\ &\le E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{1,t}] + E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{2,t}(x^{*,[T]})] + E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{3,t}]. \end{aligned}$$

By similar arguments to the proof of Theorem 4.1, we can show that

$$E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{1,t}] = \frac{c\bar{\beta}G^2(2\sqrt{T}-1)}{2}; \tag{4.31}$$

$$E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{2,t}(x^{*,[T]})] \le \frac{D^2}{2\underline{\beta}c\underline{p}}\sqrt{T}. \tag{4.32}$$

For $E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{3,t}]$, it follows from Lemma 4.9 that

$$\begin{aligned}
E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{3,t}] &= \sum_{t=1}^{T}\left\{E_{\xi_{[T-1]}}\Big[\sum_{i=1}^{N}\frac{1}{p_i}\big[\tau\psi_i(x^t_{\mathcal{J}^i})\big]\Big] - E_{\xi_{[T-1]}}\Big[\sum_{i=1}^{N}\frac{1}{p_i}\Big[\tau E\big[\psi_i(x^{t+1}_{\mathcal{J}^i}) \mid x^t\big]\Big]\Big]\right\} \\
&= \sum_{t=1}^{T}\left\{E_{\xi_{[t-1]}}\Big[\sum_{i=1}^{N}\frac{1}{p_i}\big[\tau\psi_i(x^t_{\mathcal{J}^i})\big]\Big] - E_{\xi_{[t]}}\Big[\sum_{i=1}^{N}\frac{1}{p_i}\tau\psi_i(x^{t+1}_{\mathcal{J}^i})\Big]\right\} \\
&= E_{\xi_{[0]}}\Big[\sum_{i=1}^{N}\frac{1}{p_i}\big[\tau\psi_i(x^1_{\mathcal{J}^i})\big]\Big] - E_{\xi_{[T]}}\Big[\sum_{i=1}^{N}\frac{1}{p_i}\big[\tau\psi_i(x^{T+1}_{\mathcal{J}^i})\big]\Big] \\
&\le \max_{x\in\Omega}\Big\{\sum_{i=1}^{N}\frac{1}{p_i}\big[\tau\psi_i(x^1_{\mathcal{J}^i})\big] - \sum_{i=1}^{N}\frac{1}{p_i}\big[\tau\psi_i(x_{\mathcal{J}^i})\big]\Big\},
\end{aligned}$$

where the second equality follows from (4.23).

Let $\hat{x}\in\Omega$ be the vector such that the above maximum holds, and $\eta_{\mathcal{J}^i}\in\partial\psi_i(x^1_{\mathcal{J}^i})$. Then it follows from the convexity of function $\psi(x)$ that

$$\psi_i(x^1_{\mathcal{J}^i}) - \psi_i(\hat{x}_{\mathcal{J}^i}) \le \langle \eta_{\mathcal{J}^i}, x^1_{\mathcal{J}^i} - \hat{x}_{\mathcal{J}^i}\rangle \le \frac{1}{2}\Big(\|\eta_{\mathcal{J}^i}\|^2 + \|x^1_{\mathcal{J}^i} - \hat{x}_{\mathcal{J}^i}\|^2\Big).$$

Hence, we obtain

$$\begin{aligned}
E_{\xi_{[T-1]}}[R(T)] &\le \frac{c\bar{\beta}G^2}{2}(2\sqrt{T}-1) + \frac{D^2}{2\underline{\beta}c\underline{p}}\sqrt{T} + \sum_{i=1}^{N}\frac{\tau}{2p_i}\left[\|\eta_{\mathcal{J}^i}\|^2 + \|x^1_{\mathcal{J}^i} - \hat{x}_{\mathcal{J}^i}\|^2\right] \\
&< \frac{\tau}{2\underline{p}}(G^2+D^2) + (\frac{\bar{\beta}cG^2}{2} + \frac{D^2}{2\underline{\beta}c\underline{p}})(2\sqrt{T}-1),
\end{aligned}$$

where the last inequality follows from the fact $\sqrt{T} \le 2\sqrt{T}-1$ and Assumptions 3.1-3.2. □

Theorem 4.4 implies that the C-BCG method is a no internal regret algorithm for the online optimization problem with (1.1).

The next remark shows that the regret bound of the C-BCG method is also an extension of the greedy projection method [26].

Remark 4.5. *When function $\psi(x)=0$, we have that $\sum_{i=1}^{N}\frac{1}{p_i}\left[\tau\psi_i(x^1_{\mathcal{J}^i})\right] = \sum_{i=1}^{N}\frac{1}{p_i}\left[\tau\psi_i(x^{T+1}_{\mathcal{J}^i})\right] = 0$. Hence, it holds that $E_{\xi_{[T-1]}}[\sum_{t=1}^{T} W^{3,t}] = 0$. Moreover, when $N=1$, we have $\underline{p}=1$. If*

we let $\beta_1 = 1$ and $c = 1$, then the regret bound in Theorem 4.4 reduces to $E_{\xi_{[T-1]}}[R(T)] < \frac{G^2}{2}(2\sqrt{T}-1) + \frac{D^2}{2}\sqrt{T}$, which is the same as Theorem 3.2. Hence, Theorem 4.4 is an extension of Theorem 3.2 for the greedy projection method.

If we set the probability $p_i = \frac{1}{N}$, $i = 1, \cdots, N$, for the R-BCG method, the upper bound in Theorem 3.2 can be reduced as follows, which implies that the R-BCG method has a regret of $O(N\sqrt{T})$.

Corollary 4.6. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Let $\{x^t\}$ be generated by the R-BCG method with $p_i = \frac{1}{N}$, $i = 1, \cdots, N$, for the online optimization problem with (1.1). Then, for any $x^{*,[T]} \in X^{*,[T]}$, we have*

$$E_{\xi_{[T-1]}}[R(T)] < \frac{N\tau}{2}(G^2 + D^2) + (\frac{\bar{\beta}c\bar{G}^2}{2} + \frac{ND^2}{2\underline{\beta}c})(2\sqrt{T}-1),$$

i.e. the R-BCG method has a regret of $O(N\sqrt{T})$.

Generally, it holds that $\underline{p} \geq \frac{1}{N}$, which yields that the upper bound in Theorem 4.4 is bigger than the bound in Corollary 4.6.

We also can extend the C-BCG method to the stochastic optimization problem (4.14) by the analogous analysis to Corollary 4.3. For simplicity, we omit its proof here.

Corollary 4.7. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Let $\{x^k\}$ be generated by the R-BCG method for the stochastic optimization problem (4.14), $\bar{x}^T = \frac{1}{T}\sum_{k=1}^{T} x^k$, $T \geq 1$, and $p_i = \frac{1}{N}$, $i = 1, \cdots, N$. Then, for any $x^* \in \operatorname{argmin}_x \tilde{F}(x)$, we have*

$$E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq \frac{N\tau}{2T}(G^2 + D^2) + (\frac{\bar{\beta}cG^2}{2} + \frac{D^2N}{2\underline{\beta}c})\frac{2\sqrt{T}-1}{T},$$

that is, $E_{z_{[T]}}[\tilde{F}(\bar{x}^T)] - \tilde{F}(x^*) \leq O(\frac{N}{\sqrt{T}})$.

From Corollary 4.7, we know that the upper bound $O(\frac{N}{\sqrt{T}})$ of the C-BCG method for the stochastic optimization problem (4.14) is also sharper than $O(\frac{1+\log T}{\sqrt{1+T}}N)$ given in [25].

5 Conclusion

In this paper, we propose two block coordinate gradient (BCG) methods for the online convex programming problem with (1.1). One is the BCG method with the cyclic rule, the other is the BCG method with the random rule. We show that both of these two methods have a regret $O(\sqrt{T})$, which are the same as [26]. Moreover, we extend our results to the regularized

stochastic optimization problem, and show that the results in this paper are tighter than that in [25].

In [9], an extension of the BCG method, called the "block coordinate proximal gradient (BCPG) methods with variable Bregman functions", has been studied, where the quadratic term $\frac{1}{2}\|d\|^2$ in (2.1) is replaced by the Bregman distance $B_\eta(x, x+d)$. It is shown in [9] that the BCPG methods have the same convergence rate with the block coordinate gradient descent (BCGD) method for the classical separable optimization problems. Hence, it may be possible to obtain a similar convergence of the BCPG methods with variable Bregman functions for the corresponding online or stochastic optimization problems as the results in this paper.

References

[1] N. Bansal, A. Blum, S. Chawla, and A. Meyerson, Online oblivious routing, Proceedings of the fifteenth annual ACM symposium on Parallel algorithms and architectures, pp. 44-49, 2003.

[2] S. Bubeck, Introduction to online optimization, Lecture Notes, 2011.

[3] M. Collins, A. Globerson, T. Koo, X. Carreras, and P. L. Bartlett, Exponentiated gradient algorithms for conditional random fields and Max-Margin Markov networks, The Journal of Machine Learning Research, Vol. 9, pp. 1775-1822, 2008.

[4] T. Cover, Universal portfolios, Mathematical Finance, Vol.1, pp.1-19, 1991.

[5] V. Dani, T.P. Hayes, and S.M. Kakade, The price of bandit information for online optimization, In Advances in Neural Information Processing Systems, Vol. 22, pp. 345-352, MIT Press, 2008.

[6] C. D. Dang and G. Lan, Stochastic block mirror descent methods for nonsmooth and stochastic optimization, arXiv:1309.2249, September 2013.

[7] D.P. Foster and R. Vohra, Regret in the on-line decision problem, Games and Economic Behavior, Vol. 29, pp. 7-35, 1999.

[8] E. Hazan, A. Agarwal, and S. Kale, Logarithmic regret algorithms for online convex optimization, Machine Learning, Vol. 69, pp.169-192, 2007.

[9] X.Q. Hua and N. Yamashita, Block coordinate proximal gradient methods with variable Bregman functions for nonsmooth separable optimization, Department of Applied Mathe-

matics and Physics, Graduate School of Informatics, Kyoto University, Technical Report, 2014.

[10] X.Q. Hua and N. Yamashita, Iteration complexity of a block coordinate gradient descent method for convex optimization, to appear in SIAM Journal on Optimization.

[11] S. Katomoto, A randomized block-coordinate descent method for online convex optimization problem, Master thesis, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, March, 2013. http://www-optima.amp.i.kyoto-u.ac.jp/papers/master /2013_master_kadomoto.pdf.

[12] J. Langford, L. Li, and T. Zhang, Sparse online learning via truncated gradient, In Advances in Neural Information Processing Systems 22, 2008.

[13] D.G. Luenberger, Linear and nonlinear programming, Addision-Wesley, 1984.

[14] Y. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems, SIAM Journal on Optimization, 22(2012), pp. 341-362.

[15] Y. Nesterov, Gradient methods for minimizing composite objectiv e function. CORE Discussion Paper 1007/76, Catholic University of Louvain, Belgium, 2007.

[16] Y. Nesterov, Introductory lectures on convex optimization: a basic course, 2004.

[17] F. Orabona, K. Crammer, and N. Cesa-Bianchi, A generalized online mirror descent with applications to classification and regression, Technical report, Unimi, arXiv:1304.2994, 2013.

[18] P. Richtárik and M. Takáč, Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function, Mathematical Programming. Ser. A, Vol. 144, pp.1-38, 2014.

[19] R.T. Rockafellar, Convex analysis, Princeton University Press, Princeton, New Jersey, 1970.

[20] P. Tseng, Approximation accuracy, gradient methods, and error bound for structured convex optimization, Mathematical Programming, Vol. 125, pp. 263-295, 2010.

[21] P. Tseng, Convegence of a block coordinate descent method for nondifferentiable minimization, Journal of Optimization and Applications, Vol.109, pp. 475-494, 2001.

[22] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, Manuscript, 2008.

[23] H. Wang and A. Banerjee, Randomized block coordinate descent for online and stochastic optimization, arXiv:1407.0107, July 2014.

[24] L. Xiao, Dual averaging methods for regularized stochastic learning and online optimization, Journal of Machine Learning Research, Vol. 11, pp. 2543 - 2596, 2010.

[25] Y. Xu and W. Yin, Block stochastic gradient iteration for convex and nonconvex optimization, arXiv:1408.2597, August 2014.

[26] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, In Proceedings 20th International Conference on Machine Learning, pp. 928-936, 2003.