2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

A random coordinate descent algorithm for large-scale sparse nonconvex optimization

Andrei Patrascu and Ion Necoara*

Abstract— **In this paper we develop a random coordinate descent method suitable for solving large-scale sparse nonconvex optimization problems with composite objective function. Under the typical assumptions of nonconvexity of the smooth part of the objective function and separability and convexity of the nonsmooth part (e.g. l_1 regularization, box indicator functions or others), we derive an algorithm with a very simple and cheap iteration. We prove sublinear convergence rate for our method to a stationary point. Numerical results show that our algorithm performs favourably in comparison to other algorithms on large-scale sparse nonconvex problems, e.g. the eigenvalue complementarity problem arising in different areas such as stability of dynamical systems, distributed control and resonance frequency of mechanical structures with friction.**

I. INTRODUCTION

In this paper we are concerned with large-scale linearly constrained optimization problems with composite objective functions described by the sum of a smooth nonconvex function and a nonsmooth convex function. For problems of moderate size there exist efficient algorithms such as interior-point methods, Quasi-Newton methods, and projected gradient methods with good complexity relative to the dimension. However, in the case of large-scale problems, the computation of the gradient or Hessian may be prohibitive as the iteration complexity of projected gradient or interior-point methods are of order $\mathcal{O}(n^2)$ or $\mathcal{O}(n^3)$, respectively, where n is the variable dimension. To obtain a lower iteration complexity as $\mathcal{O}(n)$ or even $\mathcal{O}(1)$, an appropriate way to approach these problems is through coordinate descent methods. Recent complexity results on random coordinate descent methods for solving smooth convex problems with separable constraints were obtained by Nesterov in [8]. In [13] an extension to convex optimization problems with composite objective functions is presented. For linearly constrained optimization problems with (composite) convex objective function, extensive complexity analysis of random coordinate descent methods can be found in [6], [7]. Further, in a series of papers [18], [19] Tseng developed a greedy (Gauss-Southwell) variant of coordinate descent method for solving linearly constrained optimization problems with composite objective function obtaining complexity estimates for the convex case and asymptotic linear convergence for the nonconvex case under error bound assumption.

The goal of this paper is to develop an efficient random coordinate descent method suitable for large-scale sparse nonconvex problems with composite objective function and linear coupled constraints. Our main result is the sublinear convergence rate in expectation of random coordinate descent method to some stationary point under the standard assumptions for composite optimization: convexity and separability of the nonsmooth part of the objective function and Lipschitz continuity of the gradient of the nonconvex part. Up to our knowledge, there is no complexity analysis of random coordinate descent algorithms for solving nonconvex optimization problems with linear coupled constraints.

It is well known that most analysis and design problems arising in robust and nonlinear control can be formulated as nonconvex optimization problems with polynomial objective functions and constraints [2], [11] (e.g robust stability analysis for characteristic polynomials, simultaneous stabilization of linear systems, pole assignment by static output feedback). Very often however, the main computational effort involved in solving most of the aforementioned control problems is the spectrum computation of symmetric/nonsymmetric matrices. Thus, an important application of our optimization model is the Eigenvalue Complementarity Problem (EiCP). The (EiCP) problem is often formulated as the maximization of the generalized Rayleigh quotient onto the standard simplex or as difference convex programming [10]. For reasons described later we analyze the properties of the logarithmic formulation of (EiCP) problem and use it to perform numerical experiments. We compare the practical behavior of our method with the method developed in [17] for the difference convex formulation of the (EiCP) problem. Our general optimization model can also be applied in other areas such as distributed computer systems [4], and traffic equilibrium [1].

The structure of the paper is as follows. We introduce the optimization model and the assumptions in Section II. Presentation of the coordinate descent algorithm and our main convergence result are given in Section III. In Section IV we provide numerical experiments for the practical behavior of our method on solving random (EiCP) problems.

II. PRELIMINARIES

We consider the space $\mathbb{R}^n$ composed by column vectors. For $x, y \in \mathbb{R}^n$ denote the scalar product by $\langle x, y\rangle = x^T y$. We use the same notation $\langle\cdot,\cdot\rangle$ for scalar products in spaces of different dimensions. For some norm $\|\cdot\|$ in $\mathbb{R}^n$, its dual norm is defined by $\|y\|^* = \max_{\|x\|=1} \langle y, x\rangle$. We consider the following decomposition of the variable dimension and of the identity matrix I_n: $n = \sum_{i=1}^{N} n_i$ and $I_n = [U_1 \cdots U_N]$, where

A. Patrascu and I. Necoara are with Automatic Control and Systems Engineering Department, University Politehnica Bucharest, 060042 Bucharest, Romania {andrei.patrascu, ion.necoara}@acse.pub.ro

978-3-952-41734-8/©2013 EUCA

$U_i \in \mathbb{R}^{n \times n_i}$. For brevity we use the following notation: for all $x \in \mathbb{R}^n$ and $i, j = 1, \cdots, N$, we denote:

$$x_i = U_i^T x \in \mathbb{R}^{n_i}, \quad \nabla_i f(x) = U_i^T \nabla f(x) \in \mathbb{R}^{n_i}$$
$$x_{ij} = \left[x_i^T \ x_j^T\right]^T, \quad \nabla_{ij} f(x) = \left[\nabla_i f(x)^T \ \nabla_j f(x)^T\right]^T.$$

In other words, x_i and $\nabla_i f(x)$ is the ith block component of the vector x and of the gradient $\nabla f(x)$, respectively. For simplicity of the exposition we use a context dependent notation as follows: let $x \in \mathbb{R}^n$, then x_{ij} denotes $[x_i^T x_j^T]^T \in \mathbb{R}^{n_i+n_j}$ in appropriate context, but in the case of operations with vectors from extended space $\mathbb{R}^n$, i.e. $y + x_{ij}$, we understand $y + U_i x_i + U_j x_j$. For any vector $x \in \mathbb{R}^n$, $\text{supp}(x)$ denotes the set of indices corresponding to nonzero components. Given a vector $a \in \mathbb{R}^n$, the subspace $\text{Null}(a^T)$ denotes the orthogonal subspace of a from $\mathbb{R}^n$.

We consider in this paper linearly constrained optimization problems with composite objective functions as follows:

$$F^* = \min_{x \in \mathbb{R}^n} F(x) \quad (:= f(x) + h(x)) \qquad (1)$$
$$\text{s.t. } a^T x = b,$$

where $a \in \mathbb{R}^n$ is a nonzero vector, f is a smooth nonconvex function and h is a nonsmooth convex separable function. Note that the (EiCP) problem is a particular case of this model (see Section IV). For brevity we use the notation $S = \{x \in \mathbb{R}^n : \ a^T x = b\}$. The following assumptions are typical for coordinate descent optimization with composite objective function, as considered in (1) (see e.g. [6], [7], [18]):

Assumption 1: .

(i) The function f has block-coordinate Lipschitz continuous gradient, i.e. there is $L_{ij} > 0$ such that

$$\|\nabla_{ij} f(x + s_{ij}) - \nabla_{ij} f(x)\| \leq L_{ij} \|s_{ij}\|,$$

for all $s_{ij} \in \mathbb{R}^{n_i+n_j}, x \in \mathbb{R}^n$ and $i \neq j = 1, \cdots, N$.

(ii) The function h is proper, convex, lower semicontinous and coordinatewise separable:

$$h(x) = \sum_{i=1}^{n} h_i(x_i) \ \forall x \in \mathbb{R}^n.$$

A well-known consequence of Assumption 1 (i) is [9]:

$$|f(x + s_{ij}) - f(x) - \langle \nabla_{ij} f(x), s_{ij} \rangle| \leq \frac{L_{ij}}{2} \|s_{ij}\|^2,$$

for all $s_{ij} \in \mathbb{R}^{n_i+n_j}$ and $x \in \mathbb{R}^n$. Based on this quadratic approximation of function f, for any $x \in \mathbb{R}^n$, if we denote $y = x + U_i s_i + U_j s_j$, then we get the following inequality:

$$F(y) \leq f(x) + \langle \nabla_{ij} f(x), s_{ij} \rangle + \frac{L_{ij}}{2} \|s_{ij}\|^2 + h(y), \qquad (2)$$

for all $s_{ij} \in \mathbb{R}^{n_i+n_j}$ and $x \in \mathbb{R}^n$. The first-order necessary optimality conditions of the optimization problem (1) are described as follows: a vector x^* is a stationary point of problem (1) if there exists a $\lambda^* \in \mathbb{R}$ such that

$$0 \in \nabla f(x^*) + \partial h(x^*) + \lambda^* a \text{ and } a^T x^* = b, \qquad (3)$$

where $\partial h(x)$ denotes the subdifferential of h at x.

III. Random coordinate descent method

In this section we present our random coordinate descent algorithm for solving the large-scale nonconvex optimization problem (1) that has many similarities with the algorithm from [7] developed for the convex case. Let the pair (i, j) be a random variable with a given probability distribution $p_{i_k j_k} = \Pr\big((i, j) = (i_k, j_k)\big)$, where we assume that $p_{ii} = 0$ for all i, and let $x^0 \in S$ be an initial feasible point. Considering the decomposition of the space $\mathbb{R}^n$ defined in subsection II, we define the following random coordinate descent algorithm for problem (1):

Algorithm (CRCD) (Coupled RCD)

1) Choose randomly a pair of (blocks) coordinates (i_k, j_k) with probability $p_{i_k j_k}$

2) Update: $x^{k+1} = x^k + U_{i_k} d_{i_k} + U_{j_k} d_{j_k}$,

where the direction $d_{ij} = [d_i^T \ d_j^T]^T$ is chosen as follows:

$$d_{ij} = \arg\min_{s_{ij}} \Big[f(x^k) + \langle \nabla_{ij} f(x^k), s_{ij} \rangle + \frac{L_{ij}}{2} \|s_{ij}\|^2 + h(x^k + s_{ij})\Big] \qquad (4)$$
$$\text{s.t. } \quad a_i^T s_i + a_j^T s_j = 0.$$

Note that the search direction d_{ij} in our algorithm is obtained by minimizing the right hand side of the Lipschitz relation (2). The reader should note that for problems with sufficiently sparse data and simple separable functions h (e.g. indicator function for box sets, l_1 regularization, etc) the computation of the ith component of the gradient and the direction d_{ij} requires $\mathcal{O}(n_i + n_j)$ operations (see [6], [19]). Moreover, in the scalar case, i.e. when $N = n$, the search direction d_{ij} can be computed in closed form, provided that h is simple, e.g. indicator function for box sets or l_1 regularization.

We assume that for every instance (i, j) we have $p_{ij} = p_{ji}$, resulting in $N(N-1)/2$ different pairs (i, j). In the sequel, we use notation ξ^k for the entire history of random pair choices and ϕ^k for the expectation of the objective function w.r.t. ξ^k:

$$\xi^k = \{(i_0, j_0), \cdots, (i_{k-1}, j_{k-1})\} \text{ and } \phi^k = E_{\xi^k}[F(x^k)].$$

Let us define the local subspace w.r.t pair (i, j) as: $S_{ij} = \{x \in S : \ x_l = 0 \ \ \forall l \neq i, j\}$. In order to provide the convergence rate of our algorithm, we have to introduce some definitions and auxiliary results.

Definition 1: Let $d, d' \in \mathbb{R}^n$, then the vector d' is *conformal* to d if: $\text{supp}(d') \subseteq \text{supp}(d)$ and $d'_j d_j \geq 0$ for all $j = 1, \cdots, n$.

We also introduce the notion of elementary vectors for a given linear subspace S.

Definition 2: An elementary vector d of $\text{Null}(a^T)$ is a vector $d \in \text{Null}(a^T)$ for which there is no nonzero $d' \in \text{Null}(a^T)$ conformal to d with $\text{supp}(d') \neq \text{supp}(d)$.

We now present some results for elementary vectors and conformal realization, whose proofs can be found in [14],

[15], [19]. A particular case of Exercise 10.6 in [15] and an interesting result in [14] provide us the following lemma:

Lemma 1: [14], [15] Given $d \in \text{Null}(a^T)$, if d is an elementary vector, then $|\text{supp}(d)| \leq 2$. Otherwise, d has a conformal realization:

$$d = d^1 + \cdots + d^s,$$

where $s \geq 2$ and $d^i \in \text{Null}(a^T)$ are elementary vectors conformal to d for all $i = 1, \cdots, s$.

An important property of convex and separable functions is given by the following lemma [19]:

Lemma 2: [19] Let h be component-wise separable and convex. For any $x, x + d \in \text{dom}h$, let d be expressed as $d = d^1 + \cdots + d^s$, for some $s \geq 1$ and some nonzero $d^t \in \mathbb{R}^n$ conformal to d for all $t = 1, \cdots, s$. Then,

$$h(x + d) - h(x) \geq \sum_{t=1}^{s} \left(h(x + d^t) - h(x) \right).$$

Let us consider $L \geq \max_{i,j} L_{ij}$. For any $x \in \mathbb{R}^n$ fixed, we introduce the following function:

$$\psi_L(s; x) = f(x) + \langle \nabla f(x), s \rangle + \frac{L}{2}\|s\|^2 + h(x + s)$$

and the following mapping associated to it:

$$d_L(x) = \arg\min_{s \in S} f(x) + \langle \nabla f(x), s \rangle + \frac{L}{2}\|s\|^2 + h(x+s). \quad (5)$$

Note that $\psi_L(s; x)$ is a L-strongly convex function in the variable s and thus the following inequality holds:

$$\psi_L(d; x) \geq \psi_L(d_L(x); x) + \frac{L}{2}\|d_L(x) - d\|^2 \quad \forall d \in \mathbb{R}^n. \quad (6)$$

The main properties of mapping $d_L(x)$ are given in the following lemma:

Lemma 3: If Assumption 1 holds and the sequence x^k is generated by Algorithm (CRCD) using a uniform distribution, then the following statements are valid:

(a) If x^k is convergent, then $d_L(x^k) \to 0$ as $k \to \infty$.

(b) A feasible point x^* is a stationary point for problem (1) if and only if $d_L(x^*) = 0$.

(c) The limit point of the sequence x^k is a stationary point for problem (1).

Proof: (a) Given a feasible point $\bar{x}$, if the sequence x^k is convergent to $\bar{x}$, then $\|x^{k+1} - x^k\| \to 0$ and thus $\|d_{i_k j_k}\| \to 0$. Since the pair (i_k, j_k) is a random variable, from Portmanteau lemma it follows that if $\|d_{i_k j_k}\| \to 0$, then $E_{i_k j_k}[\|d_{i_k j_k}\|] \to 0$. For brevity we use (i, j) and $E[s_{ij}]$ instead of (i_k, j_k) and $E_{ij}[s_{ij}]$, respectively, for any random pair (i, j) and $s_{ij} \in \mathbb{R}^{n_i + n_j}$. Also denote d_{ij} the search direction given by (4) at iteration k. From the definition of the function ψ we derive:

$$E[\psi_{L_{ij}}(d_{ij}; x^k)] \leq f(x^k) + \frac{2}{N(N-1)} \Big[\sum_{i,j} \langle \nabla_{ij} f(x^k), s_{ij} \rangle + \sum_{i,j} \frac{L_{ij}}{2}\|s_{ij}\|^2 + \sum_{i,j} h(x^k + s_{ij}) \Big]$$

for all $s_{ij} \in S_{ij}$. Using Lemma 1, we choose s_{ij} such that the corresponding extended vectors s^{ij} (with all zero entries excepting s_i and s_j on positions i and j) satisfies $d_L(x^k) = \sum_{ij} s^{ij}$ and from Lemma 2 it follows that

$$E[\psi_{L_{ij}}(d_{ij}; x^k)] \leq f(x^k) + \frac{2}{N(N-1)} \Big[\langle \nabla f(x^k), \sum_{i,j} s^{ij} \rangle + \frac{L}{2}\|\sum_{i,j} s^{ij}\|^2 + h(x + \sum_{i,j} s^{ij}) + \left(\frac{N(N-1)}{2} - 1 \right) h(x^k) \Big] =$$
$$\left(1 - \frac{2}{N(N-1)}\right) F(x^k) + \frac{2}{N(N-1)} \Big[f(x^k) + \langle \nabla f(x^k), d_L(x^k) \rangle + \frac{L}{2}\|d_L(x^k)\|^2 + h(x^k + d_L(x^k)) \Big] =$$
$$\left(1 - \frac{2}{N(N-1)}\right) F(x^k) + \frac{2}{N(N-1)} \psi_L(d_L(x^k); x^k).$$

We obtain a sequence which bounds from below $\psi_L(d_L(x^k); x^k)$ as follows:

$$\frac{N(N-1)}{2} E[\psi_{L_{ij}}(d_{ij}; x^k)] + \left(1 - \frac{N(N-1)}{2}\right) F(x^k) \leq \psi_L(d_L(x^k); x^k).$$

On the other hand, using Jensen inequality we derive another sequence which bounds $\psi_L(d_L(x^k); x^k)$ from above:

$$\psi_L(d_L(x^k); x^k)) =$$
$$\min_{s \in S} f(x^k) + \langle \nabla f(x^k), s \rangle + \frac{L}{2}\|s\|^2 + h(x^k + s) =$$
$$\min_{s_{ij} \in S_{ij}} \Big[f(x^k) + \langle \nabla f(x^k), E[s_{ij}] \rangle + \frac{L}{2}\|E[s_{ij}]\|^2 + h(x^k + E[s_{ij}]) \Big] \leq$$
$$\min_{s_{ij} \in S_{ij}} f(x^k) + \langle \nabla f(x^k), E[s_{ij}] \rangle + \frac{L}{2} E[\|s_{ij}\|^2] + E[h(x^k + s_{ij})] \leq E[\psi_L(d_{ij}; x^k)].$$

Assumption 1 (ii) and Portmanteau lemma allow us to claim that if $\|d_{ij}\| \to 0$ as $k \to \infty$, then the approximation $E[\psi_L(d_{ij}; x^k)]$ converges to $F(\bar{x})$ as $k \to \infty$. We conclude that both lower and upper bound sequences converges to $F(\bar{x})$, hence $\psi_L(d_L(x^k); x^k)$ converges to $F(\bar{x})$ as $k \to \infty$.

A trivial case of strong convexity relation (6) leads to:

$$\psi_L(0; x^k) \geq \psi_L(d_L(x^k); x^k) + \frac{L}{2}\|d_L(x^k)\|^2.$$

Note that $\psi_L(0; x^k) = F(x^k)$ and since both sequences $\psi_L(0; x^k)$ and $\psi_L(d_L(x^k); x^k)$ converge to $F(\bar{x})$ as $k \to \infty$, from strong convexity it follows that the sequence $d_L(x^k)$ converges to 0 as x^k tends to $\bar{x}$.

(b) Considering the optimality conditions for (5), it can be easily shown that if $d_L(x^*) = 0$ implies that x^* is a stationary point for (1). We prove the converse implication by contradiction. First, assume that x^* is a stationary point for (1) and there is a nonzero solution $d_L(x^*)$ of (5). Then, there exist $\lambda, \mu \in \mathbb{R}$ and $g(x^*) \in \partial h(x^*)$, $g(x^* + d_L(x^*)) \in \partial h(x^* + d_L(x^*))$, respectively, such that the optimality conditions for optimization problems (1) and (5) can be written as:

$$\begin{cases}\nabla f(x^*) + g(x^*) + \lambda a = 0\\ \nabla f(x^*) + Ld_L(x^* + g(x^* + d_L(x^*)) + \mu a = 0.\end{cases}$$

Taking the difference of the two relations above we get:

$$g(x^* + d_L(x^*)) - g(x^*) + Ld_L(x^*) + (\mu - \lambda)a = 0.$$

Considering the inner product with $d_L(x^*)$ we get:

$$L\|d_L(x^*)\|^2 + \langle g(x^* + d_L(x^*)) - g(x^*), d_L(x^*)\rangle = 0.$$

From convexity of h we see that both terms of the sum are nonnegative, thus the equality contradicts our hypothesis.

(c) As we proved in (a), if the sequence x^k generated by Algorithm (CRCD) converges to $\bar{x}$, then the sequence $d_L(x^k)$ converges to 0. Using the definition of $d_L(x^k)$ we have:

$$f(x)+\langle\nabla f(x^k), d_L(x^k)\rangle+\frac{L}{2}\|d_L(x^k)\|^2+h(x^k+d_L(x^k))\le$$
$$f(x) + \langle\nabla f(x^k), s\rangle + \frac{L}{2}\|s\|^2 + h(x^k + s) \quad \forall s \in S.$$

Taking $k \to \infty$ and using Assumption 1(ii) we get:

$$F(\bar{x}) \le f(\bar{x}) + \langle\nabla f(\bar{x}), s\rangle + \frac{L}{2}\|s\|^2 + h(\bar{x} + s).$$

This shows that $s = 0$ attains the minimum in (5) for $\bar{x}$, thus $d_L(\bar{x}) = 0$ and from (b) yields that $\bar{x}$ is a stationary point. ■

For clear and complete convergence results of the sequence generated by Algorithm (CRCD), see the technical report [12]. From the previous lemma we note that the mapping $d_L(x)$ appears to have an optimality residual role. We now present the main convergence result of our paper.

Theorem 1: Under the assumptions of Lemma 3 the following estimate on the expected convergence rate holds:

$$\min_{0\le i\le k-1}\|E_{\xi^i}[d_L(x^i)]\|^2 \le \frac{N^2\left(F(x^0) - F^*\right)}{Lk}.$$

Proof: Given a current feasible point x, denote $x^+ = x+U_id_i+U_jd_j$ as the next iterate, where direction (d_i, d_j) is given by Algorithm (CRCD) for some randomly chosen pair (i, j). For simplicity of the exposition we use the notation (ϕ, ϕ^+, ξ) instead of $(\phi^k, \phi^{k+1}, \xi^{k-1})$. Based on Lipschitz relation (2) we derive:

$$F(x^+) \le f(x) + \langle\nabla_{ij} f(x), d_{ij}\rangle + \frac{L_{ij}}{2}\|d_{ij}\|^2 + h(x^+) \le$$
$$f(x)+\langle\nabla_{ij}f(x), s_{ij}\rangle+\frac{L_{ij}}{2}\|s_{ij}\|^2+h(x + s_{ij})\ \forall s_{ij} \in S_{ij}.$$

Taking expectation in both sides, we get:

$$E_{ij}[F(x^+)] \le$$
$$E_{ij}[f(x)+\langle\nabla_{ij}f(x), s_{ij}\rangle+\frac{L_{ij}}{2}\|s_{ij}\|^2 + h(x + s_{ij})] =$$
$$f(x) + \frac{2}{N(N-1)}\Big[\sum_{i,j}\langle\nabla_{ij}f(x), s_{ij}\rangle+$$
$$\sum_{i,j}\frac{L_{ij}}{2}\|s_{ij}\|^2 + \sum_{i,j}h(x + s_{ij})\Big] \quad \forall s_{ij} \in S_{ij}.$$

From Lemma 1 it follows that any $d \in S$ has a conformal realization defined by $d = \sum_t s^t$, where the vectors $s^t \in S$ are elementary vectors conformal to d. Therefore, observing that every vector s^t has nonzero components in at most two blocks, then any vector $d \in S$ can be generated by $d = \sum_{i,j} s_{ij}$, where $s_{ij} \in S_{ij}$ and their extensions in $\mathbb{R}^n$ have at most two nonzero blocks and are conformal to d. We can apply Lemma 2 for coordinate-wise separable functions $\|\cdot\|^2$ and $h(\cdot)$ and we obtain:

$$E_{ij}[F(x^+)] \le f(x) + \frac{2}{N(N-1)}\Big(\langle\nabla f(x), \sum_t s^t\rangle+$$
$$\frac{L}{2}\|\sum_t s^t\|^2 + h(x + \sum_t s^t) + \left(\frac{N(N-1)}{2} - 1\right)h(x)\Big) =$$
$$\left(1 - \frac{2}{N(N-1)}\right)F(x) + \frac{2}{N(N-1)}\Big[f(x)+$$
$$\langle\nabla f(x), d\rangle + \frac{L}{2}\|d\|^2 + h(x + d)\Big], \qquad (7)$$

for any $d \in S$. Taking expectation w.r.t ξ in (7) for $d = d_L(x)$, we can derive:

$$\phi-\phi^+ \ge E_\xi[\psi(0; x)] - \left(1 - \frac{2}{N(N-1)}\right)E_\xi[\psi(0; x)]$$
$$- \frac{2}{N(N-1)}E_\xi[\psi(d_L(x); x)] =$$
$$\frac{2}{N(N-1)}\left(E_\xi[\psi(0; x)] - E_\xi[\psi(d_L(x); x)]\right) \ge$$
$$\frac{L}{N(N-1)}E_\xi[\|d_L(x)\|^2] \ge \frac{L}{N(N-1)}\|E_\xi[d_L(x)]\|^2,$$

where we used the strong convexity property (6) of function ψ. Now, considering the iteration index k and summing up with respect to the entire history we get:

$$\frac{L}{N(N-1)}\left(\sum_{i=0}^{k}\|E_{\xi^i}[d_L(x^i)]\|^2\right) \le F(x^0) - F^*.$$

This inequality leads us to the above result. ■

IV. Numerical Experiments

In this section we analyze the practical efficiency of the Algorithm (CRCD) derived above. As we have seen from the theoretical results, the performance of coordinate descent methods is strongly correlated to the problem dimension. We will test our algorithm on a well-known application: eigenvalue complementarity problems with sparse data. The eigenvalues of a matrix $A \in \mathbb{R}^{n\times n}$ can be seen equivalently as the roots of the characteristic polynomial det$(A - \lambda I_n)$. It is well-known that the eigenvalues can have an important role in systems and control theory, e.g. to describe expected long-time behavior of a dynamical system or to represent only intermediate values of a computational method in robust control. In some circumstances, the optimization approach for eigenvalue computation is better than the algebraic one. A classical optimization problem formulation involves the

Rayleigh quotient as the objective function of some nonconvex optimization problem [5]. The Eigenvalue Complementarity Problem (EiCP) is a generalization of the classical eigenvalue problem, which can be formulated as follows:

(EiCP) Problem:

Given matrices A and B, find $\lambda \in \mathbb{R}$ and $x \neq 0$ such that

$$\begin{cases} w = (\lambda B - A)x, \\ w \geq 0, \ x \geq 0, \ w^T x = 0. \end{cases}$$

It is interesting to note that any non-negative and irreducible matrix has an unique (including the multiplicity) complementarity eigenvalue, which is its Perron root [16]. Therefore, for the case when A is non-negative and irreducible and $B = I_n$, the solution of (EiCP) is an eigenvector corresponding to the largest eigenvalue. Finding the maximal eigenvalue for some matrix has many applications in engineering, system theory, graph theory and mechanics.

Moreover, if matrices A and B are symmetric, then we have symmetric (EiCP). In [3] and [10] it has been shown that Symmetric (EiCP) has multiple equivalent convex/nonconvex optimization formulations. We describe further the main nonconvex formulation concerning the symmetric (EiCP) problem. The numerical experiments from [17] show the advantage of the nonconvex formulations of the large-scale (EiCP) problem in comparison with the convex formulation. Taking in account that, for $A, B \succeq 0$, the convex formulation of (EiCP) is given by a QCQP problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & x^T A x \\ \text{s.t.} \ & x^T B x \leq 1, \end{aligned}$$

we remark that the quadratic constraints are relatively hard to satisfy in comparison with the standard simplex set, even for a first-order method. In [3], [10], [17] we can find recent results on convergence analysis of gradient methods for solving the logarithmic Rayleigh quotient formulation:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x) \quad \left(:= \ln \frac{x^T B x}{x^T A x}\right) \\ \text{s.t.} \ & e^T x = 1, \ x \geq 0. \end{aligned} \tag{8}$$

and a relevant motivation for the numerical efficiency of this approach. In order to have well-defined objective function, in most of the aforementioned papers the positive-definiteness of matrices A and B has been assumed. In this paper, we consider the class of non-negative matrices, i.e. $A, B \geq 0$, with positive diagonal elements, i.e. $A_{ii} \neq 0$ and $B_{ii} \neq 0$ for all $i = 1, \cdots, n$. For this class of matrices the problem (8) is also well-defined on the simplex. In order to apply our algorithm (CRCD) on the logarithmic formulation of the (EiCP) problem, we have to compute the Lipschitz constants L_{ij}. For brevity we introduce the notations $\Delta_n = \{x \in \mathbb{R}^n : e^T x = 1, x \geq 0\}$ for the standard simplex and the function $g_A(x) = \ln x^T A x$. For a given matrix A, we denote $A_{ij} = [A_i^T \ A_j^T]^T$ the pair (i, j) of block-rows of matrix A and $(A_{ij})_{ij}$ the pair (i, j) of block-columns from A_{ij}.

Lemma 4: Given non-negative matrix $A \in \mathbb{R}^{n \times n}$ such that $A_{ii} \neq 0$ for all $i = 1, \cdots, n$, then the function $g_A(x) = \ln x^T A x$ has 2 block-coordinate Lipschitz gradient on the standard simplex, i.e

$$\|\nabla_{ij} g_A(x+h_{ij}) - \nabla_{ij} g_A(x)\| \leq L_{ij} \|h_{ij}\| \ \ \forall x, x+h_{ij} \in \Delta_n,$$

where an upper bound on Lipschitz constant L_{ij} is given by

$$L_{ij} \leq \frac{2N}{\min\limits_{1 \leq i \leq N} A_{ii}} \|(A_{ij})_{ij}\|.$$

Proof: The Hessian of the function $g_A(x)$ is given by $\nabla^2 g_A(x) = \frac{2A}{x^T A x} - \frac{4(Ax)(Ax)^T}{(x^T A x)^2}$. Note that $(\nabla^2_{ij} g_A(x))_{ij} = \frac{2(A_{ij})_{ij}}{x^T A x} - \frac{4(Ax)_{ij}(Ax)^T_{ij}}{(x^T A x)^2}$. With the same arguments as in [17] we claim that

$$\|(\nabla^2_{ij} g_A(x))_{ij}\| \leq \|\frac{2(A_{ij})_{ij}}{x^T A x}\|.$$

From the Mean Value theorem we have:

$$\begin{aligned} & \|\nabla_{ij} g_A(x+h_{ij}) - \nabla_{ij} g_A(x)\| = \\ & \left\| \left(\int_0^1 (\nabla^2_{ij} g_A(x + \tau h_{ij}))_{ij} \mathrm{d}\tau \right) h_{ij} \right\| \leq \\ & \int_0^1 \|(\nabla^2_{ij} g_A(x + \tau h_{ij}))_{ij}\| \cdot \mathrm{d}\tau \cdot \|h_{ij}\| \leq \\ & \|\frac{2(A_{ij})_{ij}}{x^T A x}\| \cdot \|h_{ij}\| \ \ \forall x, x + h_{ij} \in \Delta_n. \end{aligned}$$

Observing that $\min\limits_{x \in \Delta_n} x^T A x > 0$ we obtain

$$\min_{x \in \Delta_n} x^T A x \geq \min_{x \in \Delta_n} \left(\min_{1 \leq i \leq N} A_{ii} \right) \|x\|^2 = \frac{1}{N} \min_{1 \leq i \leq N} A_{ii}.$$

and the above result can be easily derived. ■

Using the previous lemma, we can derive the block-coordinatewise Lipschitz constants for the objective function of the logarithmic formulation of the (EiCP). In the notations introduced before, the logarithmic formulation is given by

$$\min_{x \in \Delta_n} f(x) := g_B(x) - g_A(x).$$

Therefore, the local Lipschitz constants L_{ij} of function f are estimated as $L_{ij} \geq \frac{2N}{\min\limits_{1 \leq i \leq N} B_{ii}} \|(B_{ij})_{ij}\| + \frac{2N}{\min\limits_{1 \leq i \leq N} A_{ii}} \|(A_{ij})_{ij}\|$.

In [17] a variant of Difference of Convex functions (DC) algorithm is analyzed. The authors in [17] transformed both previously mentioned formulations of (EiCP) to equivalent (DC) formulations and solve these problems using (DC) Algorithm. Further, we present a comparison between (CRCD) Algorithm and (DC) Algorithm from [17]. For completeness, we also present the (DC) Algorithm for logarithmic formulation of (EiCP). Given $x^0 \in \mathbb{R}^n$, for $k \geq 0$ do:

```
Algorithm DC [17]
1. Set $y^k = \left(\mu I_n + \frac{2A}{\langle x^k, Ax^k \rangle} - \frac{2B}{\langle x^k, Bx^k \rangle}\right) x^k$,
2. Solve the quadratic program
   $x^{k+1} = \arg\min \left\{ \frac{\mu}{2}\|x\|^2 - \langle x, y_k \rangle : \langle e, x \rangle = 1, x \geq 0 \right\}$,
```

where μ is a parameter chosen in a preliminary stage of the algorithm such that the function $\frac{1}{2}\mu\|x\|^2 + \ln(x^T Ax)$ is convex. Note that (DC) Algorithm has a relatively cheap iteration, since the most computations are performed at Step 2 where matrix vector multiplication has to be computed and a projection onto simplex needs to be done. Note that in the case when at least one matrix A and B is dense, the computation of the sequence y^k is involved, typically $\mathcal{O}(n^2)$ operations. However, when these matrices are sparse the computation can be done efficiently. There are efficient algorithms for computing the projection onto simplex, e.g. Block Pivotal Principal Pivoting Algorithm described in [3], whose iteration complexity is of order $\mathcal{O}(n)$. In practical application, the value of parameter μ is crucial for the rate of convergence of (DC) Algorithm. The authors in [17] provide an approximation of μ that can be computed easily when the matrix A from (8) is positive definite. However, for indefinite matrices (as the case of non-negative irreducible matrices considered in this paper) one requires the solution of certain NP-hard problem to obtain the good approximation.

For numerical experiments we implemented both methods in C code and tested the algorithms on a PC with Intel Xeon E5410 CPU and 8 Gb RAM memory (without using any kind of parallelism) for large-scale sparse (EiCPs). We generated random sparse symmetric non-negative and irreducible matrices of dimension n and in both algorithms we start from random initial points. Each line of the generated matrices has at most 20 nonzero entries. The stopping criterion in both algorithms is $|f(x^k) - f(x^{k+1})| \leq \epsilon$. Since computing μ is very difficult, we try to tune μ in Algorithm (DC) developed in [17] for solving (EiCP) problems. In the first case, we take $\mu = n$ and from the table we observe that Algorithm (DC) can not find the optimal value f^*. In the second case, after extensive simulations we find an appropriate value for μ such that the Algorithm (DC) produces an accurate approximation of the optimal value. From the table we see that our Algorithm (CRCD) is comparable with the Algorithm (DC).

TABLE I

PERFORMANCE OF ALGORITHMS (CRCD), (DC) AND (DC-T). ALGORITHMS (CRCD) AND (DC) ARE DESCRIBED IN PREVIOUS SECTIONS AND ALGORITHM (DC-T) IS A VERSION OF (DC) INVOLVING TUNING OF PARAMETER μ.

n	(DC)			(DC-T)			CRCD	
	μ	CPU / iter	f^*	μ_t	CPU / iter	f^*	CPU / iter	f^*
5000	n	0.01 / 2	82.257	$2n$	0.02 / 8	104.15	0.06 / 105679	104.26
20000	n	0.01 / 2	41.87	$1.45n$	0.16 / 58	52.12	0.07 / 94969	52.09
50000	n	0.02 / 2	58.46	$1.41n$	0.25 / 27	72.83	0.26 / 202705	72.49
75000	n	0.03 / 2	91.77	$1.45n$	0.83 / 59	114.92	0.41 / 300868	114.03
10^5	n	0.05 / 2	100.25	$1.43n$	2.38 / 118	125.27	0.96 / 564346	125.19
$5 \cdot 10^5$	n	0.39 / 2	133.60	$1.43n$	21.06 / 105	167.19	10.73 / 3292800	167.09
$7.5 \cdot 10^5$	n	0.76 / 2	150.35	$1.43n$	39.29 / 105	187.99	18.03 / 4978021	187.89
10^6	n	1.14 / 2	417.83	$1.43n$	65.12 / 107	522.23	27.23 / 7201888	522.09

V. CONCLUSIONS

In this paper we have developed a random coordinate descent algorithm for solving sparse nonconvex optimization problem with composite objective function. Our main theoretical result is the sublinear convergence rate to a stationary point under typical assumptions for composite optimization. Also, we have tested the behavior of our method on solving large-scale sparse eigenvalue complementarity problems. From simulations we observe that our method is comparable with state-of-the art methods developed for this application.

Acknowledgements: This research has received funding from: European Union (FP7/2007–2013) EMBOCON no. 248940; CNCS (project TE, no. 19/11.08.2010); ANCS (project PN II, no. 80EU/2010); POSDRU/89/1.5/S/62557.

REFERENCES

[1] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[2] P. Dorato, *Quantified Multivariate Polynomial Inequalities: The Mathematics of Practical Control Design Problems*, IEEE Control Systems Magazine, 20(5), 48–58, 2000.

[3] J. Judice, M. Raydan, S.S. Rosa and S. A. Santos, *On the solution of the symmetric eigenvalue complementarity problem by the spectral projected gradient algorithm*, Computational Optimization and Applications, 47, 391-407, 2008.

[4] J. Kurose and R. Simha, *Microeconomic approach to optimal resource allocation in distributed computer systems*, IEEE Transactions on Computers, 38, 705-717, 1989.

[5] M. Mongeau and M. Torki, *Computing eigenelements of real symmetric matrices via optimization*, Computational Optimization and Applications, 29, 263–287, 2004.

[6] I. Necoara and A. Patrascu, *A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints*, submitted to Computational Optimization and Applications, 2012.

[7] I. Necoara, Y. Nesterov and F. Glineur, *A random coordinate descent method on large optimization problems with linear constraints*, Technical Report, University Politehnica Bucharest, 2011, `http://acse.pub.ro/person/ion-necoara/`.

[8] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, Core Discussion Paper 2/2010, 2010.

[9] Y. Nesterov, *Gradient methods for minimizing composite objective function*, CORE Discussion Papers 76/2007, 2007.

[10] M. Queiroz, J. Judice and C. Humes, *The symmetric eigenvalue complementarity problem*, Mathematics of Computation, 73(248), 1849–1863, 2004.

[11] B.N. Parlett, *The Symmetric Eigenvalue Problem. Classics in Applied Mathematics*, SIAM, Philadelphia, 1997.

[12] A. Patrascu and I. Necoara, *Random coordinate descent methods for structured nonconvex optimization problems*, Technical Report, University Politehnica Bucharest, `http://acse.pub.ro/person/ion-necoara/`.

[13] P. Richtarik and M. Takac, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, submitted to Mathematical Programming, 2012.

[14] R. T. Rockafeller, *The elementary vectors of a subspace in* $\mathbb{R}^N$, Combinatorial Mathematics and its Applications, Proceedings of Chapel Hill Conference, 104-127, 1969.

[15] R.T. Rockafeller, *Network flows and Monotropic Optimization*, Athena Scientific, 1998.

[16] A. Seeger, *Eigenvalue analysis of equilibrium processes defined by linear complementarity conditions*, Linear Algebra and Its Applications, 292, 1–14, 1999.

[17] H. A. L. Thi, M. Moeini, T.P. Dihn and J. Judice, *A DC programming approach for solving the symmetric Eigenvalue Complementarity Problem*, Computational Optimization and Applications, 51, 1097–1117, 2012.

[18] P. Tseng and S. Yun, *A coordinate gradient descent for nonsmooth separable minimization*, Technical Report, Department of Mathematics, University of Washington, to appear in Mathematical Programming Series B, 2007.

[19] P. Tseng and S. Yun, *A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization*, Journal of Optimization Theory and Applications, 140, 513-535, 2009.