

Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization

Andrei Patrascu · Ion Necoara

Received: 4 May 2013 / Accepted: 17 January 2014 / Published online: 2 February 2014
© Springer Science+Business Media New York 2014

Abstract In this paper we analyze several new methods for solving nonconvex optimization problems with the objective function consisting of a sum of two terms: one is nonconvex and smooth, and another is convex but simple and its structure is known. Further, we consider both cases: unconstrained and linearly constrained nonconvex problems. For optimization problems of the above structure, we propose random coordinate descent algorithms and analyze their convergence properties. For the general case, when the objective function is nonconvex and composite we prove asymptotic convergence for the sequences generated by our algorithms to stationary points and sublinear rate of convergence in expectation for some optimality measure. Additionally, if the objective function satisfies an error bound condition we derive a local linear rate of convergence for the expected values of the objective function. We also present extensive numerical experiments for evaluating the performance of our algorithms in comparison with state-of-the-art methods.

Keywords Large-scale nonconvex optimization · Random coordinate descent algorithms · Convergence analysis · Asymptotic convergence · Convergence rate in expectation

1 Introduction

Coordinate descent methods are among the first algorithms used for solving general minimization problems and are some of the most successful in the large-scale optimization field [3]. Roughly speaking, coordinate descent methods are based on the strategy of updating one

The research leading to these results has received funding from: the European Union (FP7/2007–2013) EMBOCON under grant agreement no 248940; CNCS (project TE-231, 19/11.08.2010); ANCS (project PN II, 80EU/2010); POSDRU/89/1.5/S/62557.

A. Patrascu · I. Necoara (✉)
Automatic Control and Systems Engineering Department, University Politehnica Bucharest,
060042 Bucharest, Romania
e-mail: ion.necoara@acse.pub.ro

A. Patrascu
e-mail: andrei.patrascu@acse.pub.ro

(block) coordinate of the vector of variables per iteration using some index selection procedure (e.g. cyclic, greedy, random). This often reduces drastically the complexity per iteration and memory requirements, making these methods simple and scalable. There exist numerous papers dealing with the convergence analysis of this type of methods [1, 10, 14, 15, 22, 31], which confirm the difficulties encountered in proving the convergence for nonconvex and nonsmooth objective functions. For instance, regarding coordinate minimization of nonconvex functions, Powell [22] provided some examples of differentiable functions whose properties lead the algorithm to a closed loop. Also, proving convergence of coordinate descent methods for minimization of nondifferentiable objective functions is challenging [1]. However, for nonconvex and nonsmooth objective functions with certain structure (e.g. composite objective functions) there are available convergence results for coordinate descent methods based on greedy index selection [2, 10, 31] or random index selection [11]. Recently, Nesterov [18] derived complexity results for random coordinate gradient descent methods for solving smooth and convex optimization problems. In [24] the authors generalized Nesterov's results to convex problems with composite objective functions. An extensive complexity analysis of coordinate gradient descent methods for solving linearly constrained optimization problems with convex (composite) objective function can be found in [2, 13–15].

In this paper we also consider large-scale nonconvex optimization problems with the objective function consisting of a sum of two terms: one is nonconvex, smooth and given by a black-box oracle, and another is convex but simple and its structure is known. Further, we analyze unconstrained but also singly linearly constrained nonconvex problems. We also assume that the dimension of the problem is so large that traditional optimization methods cannot be directly employed since basic operations, such as the updating of the gradient, are too computationally expensive. These types of problems arise in many fields such as data analysis (classification, text mining) [4, 6], systems and control theory (optimal control, pole assignment by static output feedback) [7, 8, 16, 20], machine learning [6, 14, 25, 28, 33] and truss topology design [9, 23]. The goal of this paper is to analyze several new random coordinate gradient descent methods suited for large-scale nonconvex problems with composite objective function. Recently, after our paper came under review, a variant of random coordinate descent method for solving composite nonconvex problems was also proposed in [11]. For our coordinate descent algorithm, which is designed to minimize unconstrained composite nonconvex objective functions, we prove asymptotic convergence of the generated sequence to stationary points and sublinear rate of convergence in expectation for some optimality measure. Additionally, if the objective function satisfies an error bound condition, a local linear rate of convergence for expected values of the objective function is obtained. We also provide convergence analysis for a coordinate descent method designed for solving singly linearly constrained nonconvex problems and obtain similar results as in the unconstrained case. Note that our analysis is very different from the convex case [13–15, 18, 24] and is based on the notion of optimality measure and a supermartingale convergence theorem. Furthermore, unlike to other coordinate descent methods for nonconvex problems, our algorithms offer some important advantages, e.g. due to the randomization our algorithms are simpler and are adequate for modern computational architectures. We also present the results of preliminary computational experiments, which confirm the superiority of our methods compared with other algorithms for large-scale nonconvex optimization.

Contribution: The contribution of the paper can be summarized as follows:

- (a) For unconstrained problems we propose a 1-random coordinate descent method (1-RCD), that involves at each iteration the solution of an optimization subproblem with

- respect to only one (block) variable while keeping all others fixed. We show that this solution can be usually computed in closed form (Sect. 2.2).
- (b) For the linearly constrained case we propose a 2-random coordinate descent method (2-RCD), that involves at each iteration the solution of a subproblem depending on two (block) variables while keeping all other variables fixed. We show that in most cases this solution can be found in linear time (Sect. 3.1).
 - (c) For each of the algorithms we introduce some optimality measure and devise a convergence analysis using this framework. In particular, for both algorithms, (1-RCD) and (2-RCD), we establish asymptotic convergence of the generated sequences to stationary points (Theorems 1 and 4) and sublinear rate of convergence for the expected values of the corresponding optimality measures (Theorems 2 and 5).
 - (d) If the objective function satisfies an error bound condition a local linear rate of convergence for the expected values of the objective function is proved (Theorem 3).

Content: The structure of the paper is as follows. In Sect. 2 we introduce a 1-random coordinate descent algorithm for unconstrained minimization of nonconvex composite functions. Further, we analyze the convergence properties of the algorithm under standard assumptions. Also, under the error bound assumption we obtain linear convergence rate for the expected values of objective function. In Sect. 3 we derive a 2-random coordinate descent method for solving singly linearly constrained nonconvex problems and analyze its convergence. In Sect. 4 we report numerical results on large-scale eigenvalue complementarity problems, which is an important application in control theory.

Notation: We consider the space \mathbb{R}^n composed by column vectors. For $x, y \in \mathbb{R}^n$ we denote the scalar product by $\langle x, y \rangle = x^T y$ and $\|x\| = (x^T x)^{1/2}$. We use the same notation $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ for scalar products and norms in spaces of different dimensions. For some norm $\|\cdot\|_\alpha$ in \mathbb{R}^n , its dual norm is defined by $\|y\|_\alpha^* = \max_{\|x\|_\alpha=1} \langle y, x \rangle$. We consider the following decomposition of the variable dimension: $n = \sum_{i=1}^N n_i$. Also, we denote a block decomposition of $n \times n$ identity matrix by $I_n = [U_1 \dots U_N]$, where $U_i \in \mathbb{R}^{n \times n_i}$. For brevity we use the following notation: for all $x \in \mathbb{R}^n$ and $i, j = 1, \dots, N$, we denote:

$$\begin{aligned}
 x_i &= U_i^T x \in \mathbb{R}^{n_i}, & \nabla_i f(x) &= U_i^T \nabla f(x) \in \mathbb{R}^{n_i} \\
 x_{ij} &= \begin{bmatrix} x_i^T & x_j^T \end{bmatrix}^T \in \mathbb{R}^{n_i+n_j}, & \nabla_{ij} f(x) &= \begin{bmatrix} \nabla_i f(x)^T & \nabla_j f(x)^T \end{bmatrix}^T \in \mathbb{R}^{n_i+n_j}.
 \end{aligned}$$

2 Unconstrained minimization of composite objective functions

In this section we analyze a variant of random block coordinate gradient descent method, which we call the *1-random coordinate descent method* (1-RCD), for solving large-scale unconstrained nonconvex problems with composite objective function. The method involves at each iteration the solution of an optimization subproblem only with respect to one (block) variable while keeping all other variables fixed. After discussing several necessary mathematical preliminaries, we introduce an optimality measure, which will be the basis for the construction and analysis of Algorithm (1-RCD). We establish asymptotic convergence of the sequence generated by Algorithm (1-RCD) to a stationary point and then we show sublinear rate of convergence in expectation for the corresponding optimality measure. For some well-known particular cases of nonconvex objective functions arising frequently in applications, the complexity per iteration of our Algorithm (1-RCD) is of order $\mathcal{O}(\max_i n_i)$.

2.1 Problem formulation

The problem of interest in this section is the unconstrained nonconvex minimization problem with composite objective function:

$$F^* = \min_{x \in \mathbb{R}^n} F(x) \quad (:= f(x) + h(x)), \tag{1}$$

where the function f is smooth and h is a convex, separable, nonsmooth function. Since h is nonsmooth, then for any $x \in \text{dom}(h)$ we denote by $\partial h(x)$ the subdifferential (set of subgradients) of h at x . The smooth and nonsmooth components in the objective function of (1) satisfy the following assumptions:

Assumption 1 (i) The function f has block coordinate Lipschitz continuous gradient, i.e. there are constants $L_i > 0$ such that:

$$\|\nabla_i f(x + U_i s_i) - \nabla_i f(x)\| \leq L_i \|s_i\| \quad \forall s_i \in \mathbb{R}^{n_i}, x \in \mathbb{R}^n, i = 1, \dots, N.$$

(ii) The function h is proper, convex, continuous and block separable:

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

where the functions $h_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}$ are convex for all $i = 1, \dots, N$.

These assumptions are typical for the coordinate descent framework and the reader can find similar variants in [11, 14, 15, 18, 31]. An immediate consequence of Assumption 1 (i) is the following well-known inequality [17]:

$$|f(x + U_i s_i) - f(x) - \langle \nabla_i f(x), s_i \rangle| \leq \frac{L_i}{2} \|s_i\|^2 \quad \forall s_i \in \mathbb{R}^{n_i}, x \in \mathbb{R}^n. \tag{2}$$

Based on this quadratic approximation of function f we get the inequality:

$$F(x + U_i s_i) \leq f(x) + \langle \nabla_i f(x), s_i \rangle + \frac{L_i}{2} \|s_i\|^2 + h(x + U_i s_i) \quad \forall s_i \in \mathbb{R}^{n_i}, x \in \mathbb{R}^n. \tag{3}$$

Given local Lipschitz constants $L_i > 0$ for $i = 1, \dots, N$, we define the vector $L = [L_1 \dots L_N]^T \in \mathbb{R}^N$, the diagonal matrix $D_L = \text{diag}(L_1 I_{n_1}, \dots, L_N I_{n_N}) \in \mathbb{R}^{n \times n}$ and the following pair of dual norms:

$$\|x\|_L = \left(\sum_{i=1}^N L_i \|x_i\|^2 \right)^{1/2} \quad \forall x \in \mathbb{R}^n, \quad \|y\|_L^* = \left(\sum_{i=1}^N L_i^{-1} \|y_i\|^2 \right)^{1/2} \quad \forall y \in \mathbb{R}^n.$$

Under Assumption 1, we can state the first order necessary optimality conditions for the nonconvex optimization problem (1): if $x^* \in \mathbb{R}^n$ is a local minimum for (1), then the following relation holds

$$0 \in \nabla f(x^*) + \partial h(x^*).$$

Any vector x^* satisfying this relation is called a *stationary point* for nonconvex problem (1).

2.2 A 1-random coordinate descent algorithm

We analyze a variant of random coordinate descent method suitable for solving large-scale nonconvex problems of the form (1). Let $i \in \{1, \dots, N\}$ be a random variable and $p_{i_k} = \Pr(i = i_k)$ be its probability distribution. Given a point x , one block is chosen randomly with respect to the probability distribution p_i and the quadratic model (3) derived from the composite objective function is minimized with respect to this block of coordinates (see also [18,24]). Our method has the following iteration: given an initial point x_0 , then for all $k \geq 0$

Algorithm (1-RCD)

1. Choose randomly a block of coordinates i_k with probability p_{i_k}
2. Set $x^{k+1} = x^k + U_{i_k}d_{i_k}$,

where the direction d_{i_k} is computed as follows:

$$d_{i_k} = \arg \min_{s_{i_k} \in \mathbb{R}^{n_{i_k}}} f(x^k) + \langle \nabla_{i_k} f(x^k), s_{i_k} \rangle + \frac{L_{i_k}}{2} \|s_{i_k}\|^2 + h(x^k + U_{i_k}s_{i_k}). \tag{4}$$

Note that the direction d_{i_k} is a minimizer of the quadratic approximation model given in (3). Further, from Assumption 1 (ii) we see that $h(x^k + U_{i_k}s_{i_k}) = h_{i_k}(x_{i_k}^k + s_{i_k}) + \sum_{i \neq i_k} h_i(x_i^k)$ and thus for computing d_{i_k} we only need to know the function $h_{i_k}(\cdot)$. An important property of our algorithm is that for certain particular cases of function h , the complexity per iteration of Algorithm (1-RCD) is very low. In particular, for certain simple functions h , very often met in many applications from signal processing, machine learning and optimal control, the direction d_{i_k} can be computed in closed form, e.g.:

- (I) For some $l, u \in \mathbb{R}^n$, with $l \leq u$, we consider the box indicator function

$$h(x) = \begin{cases} 0 & \text{if } l \leq x \leq u \\ \infty & \text{otherwise.} \end{cases} \tag{5}$$

In this case the direction d_{i_k} has the explicit expression:

$$d_{i_k} = \left[x_{i_k}^k - \frac{1}{L_{i_k}} \nabla_{i_k} f(x^k) \right]_{[l_{i_k}, u_{i_k}]} \quad \forall i_k = 1, \dots, N,$$

where $[x]_{[l, u]}$ is the orthogonal projection of vector x on box set $[l, u]$.

- (II) Given a nonnegative scalar $\beta \in \mathbb{R}_+$, we consider the ℓ_1 -regularization function defined by the 1-norm

$$h(x) = \beta \|x\|_1. \tag{6}$$

In this case, considering $n = N$, the direction d_{i_k} has the explicit expression:

$$d_{i_k} = \text{sgn}(t_{i_k}) \cdot \max \left\{ |t_{i_k}| - \frac{\beta}{L_{i_k}}, 0 \right\} - x_{i_k} \quad \forall i_k = 1, \dots, n,$$

where $t_{i_k} = x_{i_k} - \frac{1}{L_{i_k}} \nabla_{i_k} f(x^k)$.

In these examples the arithmetic complexity of computing the next iterate x^{k+1} , once $\nabla_{i_k} f(x^k)$ is known, is of order $\mathcal{O}(n_{i_k})$. The reader can find other favorable examples of nonsmooth functions h which preserve the low iteration complexity of Algorithm (1-RCD) (see also [14,31] for other examples). Note that most of the (coordinate descent) methods designed for solving nonconvex problems usually have complexity per iteration at least of

order $\mathcal{O}(n)$ (see e.g. [31], where the authors analyze a greedy coordinate descent method). Coordinate descent methods that have similar complexity per iteration as our random method can be found e.g. in [30], where the index selection is made cyclically (Gauss-Seidel rule). But Algorithm (1-RCD) also offers other important advantages, e.g. due to the randomization the algorithm is adequate for modern computational architectures (e.g. distributed and parallel architectures) [16,25].

We assume that the sequence of random variables i_0, \dots, i_k are i.i.d. In the sequel, we use the notation ξ^k for the entire history of random index selection

$$\xi^k = \{i_0, \dots, i_k\}$$

and notation

$$\phi^k = E \left[F \left(x^k \right) \right],$$

for the expectation taken w.r.t. ξ^{k-1} . Given $s, x \in \mathbb{R}^n$, we introduce the following function and the associated map (operator):

$$\begin{aligned} \psi_L(s; x) &= f(x) + \langle \nabla f(x), s \rangle + \frac{1}{2} \|s\|_L^2 + h(x + s), \\ d_L(x) &= \arg \min_{s \in \mathbb{R}^n} f(x) + \langle \nabla f(x), s \rangle + \frac{1}{2} \|s\|_L^2 + h(x + s). \end{aligned} \tag{7}$$

Based on this map, we now introduce an *optimality measure* which will be the basis for the analysis of Algorithm (1-RCD):

$$M_1(x, L) = \|D_L \cdot d_L(x)\|_L^*.$$

The map $M_1(x, L)$ is an optimality measure for optimization problem (1) in the sense that it is positive for all nonstationary points and zero for stationary points (see Lemma 1 below):

Lemma 1 *For any given vector $\tilde{L} \in \mathbb{R}^N$ with positive entries, a vector $x^* \in \mathbb{R}^n$ is a stationary point for problem (1) if and only if the value $M_1(x^*, \tilde{L}) = 0$.*

Proof Based on the optimality conditions of subproblem (7), it can be easily shown that if $M_1(x^*, \tilde{L}) = 0$, then x^* is a stationary point for the original problem (1). We prove the converse implication by contradiction. Assume that x^* is a stationary point for (1) and $M_1(x^*, \tilde{L}) > 0$. It follows that $d_{\tilde{L}}(x^*)$ is a nonzero solution of subproblem (7). Then, there exist the subgradients $g(x^*) \in \partial h(x^*)$ and $g(x^* + d_{\tilde{L}}(x^*)) \in \partial h(x^* + d_{\tilde{L}}(x^*))$ such that the optimality conditions for optimization problems (1) and (7) can be written as:

$$\begin{cases} \nabla f(x^*) + g(x^*) = 0 \\ \nabla f(x^*) + D_{\tilde{L}} d_{\tilde{L}}(x^*) + g(x^* + d_{\tilde{L}}(x^*)) = 0. \end{cases}$$

Taking the difference of the two relations above and considering the inner product with $d_{\tilde{L}}(x^*) \neq 0$ on both sides of the equation, we get:

$$\|d_{\tilde{L}}(x^*)\|_{\tilde{L}}^2 + \langle g(x^* + d_{\tilde{L}}(x^*)) - g(x^*), d_{\tilde{L}}(x^*) \rangle = 0.$$

From convexity of the function h we see that both terms in the above sum are nonnegative and thus $d_{\tilde{L}}(x^*) = 0$, which contradicts our hypothesis. In conclusion $M_1(x^*, \tilde{L}) = 0$. \square

Note that $\psi_L(s; x)$ is an 1-strongly convex function in the variable s w.r.t. norm $\|\cdot\|_L$ and thus $d_L(x)$ is unique and the following inequality holds:

$$\psi_L(s; x) \geq \psi_L(d_L(x); x) + \frac{1}{2} \|d_L(x) - s\|_L^2 \quad \forall x, s \in \mathbb{R}^n. \tag{8}$$

2.3 Convergence of algorithm (1-RCD)

In this section, we analyze the convergence properties of Algorithm (1-RCD). Firstly, we prove the asymptotic convergence of the sequence generated by Algorithm (1-RCD) to stationary points. For proving the asymptotic convergence we use the following supermartingale convergence result of Robbins and Siegmund (see [21, Lemma 11 on page 50]):

Lemma 2 *Let v_k, u_k and α_k be three sequences of nonnegative random variables such that*

$$E[v_{k+1}|\mathcal{F}_k] \leq (1 + \alpha_k)v_k - u_k \quad \forall k \geq 0 \text{ a.s.} \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k < \infty \text{ a.s.},$$

where \mathcal{F}_k denotes the collections $v_0, \dots, v_k, u_0, \dots, u_k, \alpha_0, \dots, \alpha_k$. Then, we have $\lim_{k \rightarrow \infty} v_k = v$ for a random variable $v \geq 0$ a.s. and $\sum_{k=0}^{\infty} u_k < \infty$ a.s.

In the next lemma we prove that Algorithm (1-RCD) is a descent method, i.e. the objective function is nonincreasing along its iterations:

Lemma 3 *Let x^k be the sequence generated by Algorithm (1-RCD) under Assumption 1. Then, the following relation holds:*

$$F(x^{k+1}) \leq F(x^k) - \frac{L_{i_k}}{2} \|d_{i_k}\|^2 \quad \forall k \geq 0. \tag{9}$$

Proof From the optimality conditions of subproblem (4) we have that there exists a subgradient $g(x_{i_k}^k + d_{i_k}) \in \partial h_{i_k}(x_{i_k}^k + d_{i_k})$ such that:

$$\nabla_{i_k} f(x^k) + L_{i_k} d_{i_k} + g(x_{i_k}^k + d_{i_k}) = 0.$$

On the other hand, since the function h_{i_k} is convex, according to Assumption 1 (ii), the following inequality holds:

$$h_{i_k}(x_{i_k}^k + d_{i_k}) - h_{i_k}(x_{i_k}^k) \leq \langle g(x_{i_k}^k + d_{i_k}), d_{i_k} \rangle$$

Applying the previous two relations in (3) and using the separability of the function h , then under Assumption 1 (ii) we have that

$$\begin{aligned} F(x^{k+1}) &\leq F(x^k) + \langle \nabla_{i_k} f(x^k), d_{i_k} \rangle + \frac{L_{i_k}}{2} \|d_{i_k}\|^2 + h_{i_k}(x_{i_k}^k + d_{i_k}) - h_{i_k}(x_{i_k}^k) \\ &\leq F(x^k) + \langle \nabla_{i_k} f(x^k), d_{i_k} \rangle + \frac{L_{i_k}}{2} \|d_{i_k}\|^2 + \langle g(x_{i_k}^k + d_{i_k}), d_{i_k} \rangle \\ &\leq F(x^k) - \frac{L_{i_k}}{2} \|d_{i_k}\|^2. \end{aligned}$$

□

Using Lemma 3, we state the following result regarding the asymptotic convergence of Algorithm (1-RCD).

Theorem 1 *If Assumption 1 holds for the composite objective function F of problem (1) and the sequence x^k is generated by Algorithm (1-RCD) using the uniform distribution, then the following statements are valid:*

- (i) *The sequence of random variables $M_1(x^k, L)$ converges to 0 a.s. and the sequence $F(x^k)$ converges to a random variable \bar{F} a.s.*

(ii) Any accumulation point of the sequence x^k is a stationary point for optimization problem (1).

Proof (i) From Lemma 3 we get:

$$F(x^{k+1}) - F^* \leq F(x^k) - F^* - \frac{L_{i_k}}{2} \|d_{i_k}\|^2 \quad \forall k \geq 0.$$

We now take the expectation conditioned on ξ^{k-1} and note that i_k is independent on the past ξ^{k-1} , while x^k is fully determined by ξ^{k-1} . We thus obtain:

$$\begin{aligned} E[F(x^{k+1}) - F^* | \xi^{k-1}] &\leq F(x^k) - F^* - \frac{1}{2} E[L_{i_k} \cdot \|d_{i_k}\|^2 | \xi^{k-1}] \\ &\leq F(x^k) - F^* - \frac{1}{2N} \|d_L(x^k)\|_L^2. \end{aligned}$$

Using the supermartingale convergence theorem given in Lemma 2 in the previous inequality, we can ensure that

$$\lim_{k \rightarrow \infty} F(x^k) - F^* = \theta \quad \text{a.s.}$$

for a random variable $\theta \geq 0$ and thus $\bar{F} = \theta + F^*$. Further, due to almost sure convergence of sequence $F(x^k)$, it can be easily seen that $\lim_{k \rightarrow \infty} F(x^k) - F(x^{k+1}) = 0$ a.s. From $x^{k+1} - x^k = U_{i_k} d_{i_k}$ and Lemma 3 we have:

$$\frac{L_{i_k}}{2} \|d_{i_k}\|^2 = \frac{L_{i_k}}{2} \|x^{k+1} - x^k\|^2 \leq F(x^k) - F(x^{k+1}) \quad \forall k \geq 0,$$

which implies that

$$\lim_{k \rightarrow \infty} \|x^{k+1} - x^k\| = 0 \quad \text{and} \quad \lim_{k \rightarrow \infty} \|d_{i_k}\| = 0 \quad \text{a.s.}$$

As $\|d_{i_k}\| \rightarrow 0$ a.s., we can conclude that the random variable $E[\|d_{i_k}\| | \xi^{k-1}] \rightarrow 0$ a.s. or equivalently $M_1(x^k, L) \rightarrow 0$ a.s.

(ii) For brevity we assume that the entire sequence x^k generated by Algorithm (1-RCD) is convergent. Let \bar{x} be the limit point of the sequence x^k . In the first part of the theorem we proved that the sequence of random variables $d_L(x^k)$ converges to 0 a.s. Using the definition of $d_L(x^k)$ we have:

$$\begin{aligned} f(x^k) + \langle \nabla f(x^k), d_L(x^k) \rangle + \frac{1}{2} \|d_L(x^k)\|_L^2 + h(x^k + d_L(x^k)) \\ \leq f(x^k) + \langle \nabla f(x^k), s \rangle + \frac{1}{2} \|s\|_L^2 + h(x^k + s) \quad \forall s \in \mathbb{R}^n, \end{aligned}$$

and taking the limit $k \rightarrow \infty$ and using Assumption 1 (ii) we get:

$$F(\bar{x}) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), s \rangle + \frac{1}{2} \|s\|_L^2 + h(\bar{x} + s) \quad \forall s \in \mathbb{R}^n.$$

This shows that $d_L(\bar{x}) = 0$ is the minimum in subproblem (7) for $x = \bar{x}$ and thus $M_1(\bar{x}, L) = 0$. From Lemma 1 we conclude that \bar{x} is a stationary point for optimization problem (1). \square

The next theorem proves the convergence rate of the optimality measure $M_1(x^k, L)$ towards 0 in expectation.

Theorem 2 *Let F satisfy Assumption 1. Then, the Algorithm (1-RCD) based on the uniform distribution generates a sequence x^k satisfying the following convergence rate for the expected values of the optimality measure:*

$$\min_{0 \leq l \leq k} E \left[\left(M_1(x^l, L) \right)^2 \right] \leq \frac{2N (F(x^0) - F^*)}{k + 1} \quad \forall k \geq 0.$$

Proof For simplicity of the exposition we use the following notation: given the current iterate x , denote the next iterate $x^+ = x + U_i d_i$, where direction d_i is given by (4) for some random chosen index i w.r.t. uniform distribution. For brevity, we also adapt the notation of expectation upon the entire history, i.e. (ϕ, ϕ^+, ξ) instead of $(\phi^k, \phi^{k+1}, \xi^{k-1})$. From Assumption 1 and inequality (3) we have:

$$F(x^+) \leq f(x) + \langle \nabla_i f(x), d_i \rangle + \frac{L_i}{2} \|d_i\|^2 + h_i(x_i + d_i) + \sum_{j \neq i} h_j(x_j).$$

We now take the expectation conditioned on ξ :

$$\begin{aligned} E[F(x^+) | \xi] &\leq E \left[f(x) + \langle \nabla_i f(x), d_i \rangle + \frac{L_i}{2} \|d_i\|^2 + h_i(x_i + d_i) + \sum_{j \neq i} h_j(x_j) \mid \xi \right] \\ &\leq f(x) + \frac{1}{N} \left[\langle \nabla f(x), d_L(x) \rangle + \frac{1}{2} \|d_L(x)\|_L^2 + h(x + d_L(x)) + (N - 1)h(x) \right]. \end{aligned}$$

After rearranging the above expression we get:

$$E[F(x^+) | \xi] \leq \left(1 - \frac{1}{N} \right) F(x) + \frac{1}{N} \psi_L(d_L(x); x). \tag{10}$$

Now, by taking the expectation in (10) w.r.t. ξ we obtain:

$$\phi^+ \leq \left(1 - \frac{1}{N} \right) \phi + E \left[\frac{1}{N} \psi_L(d_L(x); x) \right], \tag{11}$$

and then using the 1–strong convexity property of ψ_L we get:

$$\begin{aligned} \phi - \phi^+ &\geq \phi - \left(1 - \frac{1}{N} \right) \phi - \frac{1}{N} E [\psi_L(d_L(x); x)] \\ &= \frac{1}{N} (E [\psi_L(0; x)] - E [\psi_L(d_L(x); x)]) \\ &\geq \frac{1}{2N} E [\|d_L(x)\|_L^2] = \frac{1}{2N} E [(M_1(x, L))^2]. \end{aligned} \tag{12}$$

Now coming back to the notation dependent on k and summing w.r.t. the entire history we have:

$$\frac{1}{2N} \sum_{l=0}^k E [(M_1(x^l, L))^2] \leq \phi^0 - F^*,$$

which leads to the statement of the theorem. □

It is important to note that the convergence rate for the Algorithm (1-RCD) given in Theorem 2 is typical for the class of first order methods designed for solving nonconvex and nonsmooth optimization problems (see e.g. [19] for more details). Recently, after our paper came under review, a variant of 1-random coordinate descent method for solving composite

nonconvex problems was also proposed in [11]. However, the authors in [11] do not provide complexity results for their algorithm, but only asymptotic convergence in expectation. Note also that our convergence results are different from the convex case [18,24], since here we introduce another optimality measure and we use the supermartingale convergence theorem in the analysis.

Furthermore, when the objective function F is smooth and nonconvex, i.e. $h = 0$, the first order necessary conditions of optimality become $\nabla f(x^*) = 0$. Also, note that in this case, the optimality measure $M_1(x, L)$ is given by: $M_1(x, L) = \|\nabla f(x)\|_L^*$. An immediate consequence of Theorem 2 in this case is the following result:

Corollary 1 *Let $h = 0$ and f satisfy Assumption 1 (i). Then, in this case, the Algorithm (1-RCD) based on the uniform distribution generates a sequence x^k satisfying the following convergence rate for the expected values of the norm of the gradients:*

$$\min_{0 \leq l \leq k} E \left[\left(\|\nabla f(x^l)\|_L^* \right)^2 \right] \leq \frac{2N (F(x^0) - F^*)}{k + 1} \quad \forall k \geq 0.$$

2.4 Linear convergence rate of Algorithm (1-RCD) under error bound assumption

In this subsection, an improved rate of convergence is shown for Algorithm (1-RCD) under an additional error bound assumption. In what follows, X^* denotes the set of stationary points of optimization problem (1), $\text{dist}(x, S) = \min_{y \in S} \|y - x\|$ and the vector $\mathbf{1} = [1 \dots 1]^T \in \mathbb{R}^N$.

Assumption 2 A local error bound holds for the objective function of optimization problem (1), i.e. for any $\eta \geq F^* = \min_{x \in \mathbb{R}^n} F(x)$ there exist $\tau > 0$ and $\epsilon > 0$ such that

$$\text{dist}(x, X^*) \leq \tau M_1(x, \mathbf{1}) \quad \forall x \in V,$$

where $V = \{x \in \mathbb{R}^n : F(x) \leq \eta, M_1(x, \mathbf{1}) \leq \epsilon\}$. Moreover, there exists $\rho > 0$ such that $\|x^* - y^*\| \geq \rho$ whenever $x^*, y^* \in X^*$ with $f(x^*) \neq f(y^*)$.

For example, Assumption 2 holds for composite objective functions satisfying the following properties (see [31,32] for more examples):

- (i) f is a quadratic function (even nonconvex) and h is polyhedral.
- (ii) f is strongly convex and has Lipschitz continuous gradient.

Note that the box indicator function (5) and ℓ_1 -regularization function (6) are polyhedral functions. Note also that for strongly convex functions, Assumption 2 is globally satisfied.

In this section, we also assume that function f has global Lipschitz continuous gradient, i.e. there exists a global Lipschitz constant $L_f > 0$ such that:

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

It is well known that this property leads to the following inequality [17]:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L_f}{2} \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n. \tag{13}$$

For a given convex function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ we also define the proximal map $\text{prox}_h(x) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ as $\text{prox}_h(x) = \arg \min_{y \in \mathbb{R}^n} \frac{1}{2} \|y - x\|^2 + h(y)$. For convex functions and \mathbb{R}^n endowed with the Euclidean norm, Nesterov shows in [18] the following relation between L_f and L_i : $L_f \leq \sum_{i=1}^N L_i$. However, in the nonconvex case we cannot derive a similar inequality, but only the obvious relation: $L_i \leq L_f$. In order to analyze the convergence properties of Algorithm

(1-RCD) for minimizing a composite objective function which satisfies Assumption 2, we require the following auxiliary result:

Lemma 4 *Let $h : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. Then, the map $\omega : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ defined by*

$$\omega(\alpha) = \frac{\| \text{prox}_{\alpha h}(x + \alpha d) - x \|}{\alpha},$$

is nonincreasing w.r.t. α for any $x, d \in \mathbb{R}^n$.

Proof Note that this lemma is a generalization of [5, Lemma 2.2] from the projection operator to the “prox” operator case. The proof of this generalization is given in Appendix. \square

Using separability of h according to Assumption 1 (ii), it is easy to see that the map $d_L(x)$ satisfies:

$$x + d_L(x) = \arg \min_{y \in \mathbb{R}^n} \frac{1}{2} \|y - x + D_L^{-1} \nabla f(x)\|^2 + \sum_{i=1}^N \frac{1}{L_i} h_i(y_i),$$

and in a more compact notation we have:

$$(d_L(x))_i = \text{prox}_{\frac{1}{L_i} h_i}(x_i - 1/L_i \nabla_i f(x)) - x_i \quad \forall i = 1, \dots, N.$$

Using this expression in Lemma 4, we conclude that:

$$\|(d_1(x))_i\| \leq \max\{1, L_i\} \cdot \|(d_L(x))_i\| \quad \forall i = 1, \dots, N \tag{14}$$

and moreover,

$$M_1(x, \mathbf{1}) \leq \max_{1 \leq i \leq N} \{1, 1/\sqrt{L_i}\} \cdot M_1(x, L). \tag{15}$$

Further, we denote $\tau_L = \max_{1 \leq i \leq N} \{1, 1/\sqrt{L_i}\}$. The following theorem shows that Algorithm (1-RCD) for minimizing composite functions with error bound (Assumption 2) has linear convergence rate for the expected values of the objective function:

Theorem 3 *Under Assumptions 1 and 2, let x^k be the sequence generated by Algorithm (1-RCD) with uniform probabilities. Then, we have the following linear convergence rate for the expected values of the objective function:*

$$\phi^k - \bar{F} \leq \left(1 - \frac{1}{N[\tau \tau_L(L_f + \bar{L}) + 1]} \right)^k (F(x^0) - \bar{F})$$

for any k sufficiently large, where $\bar{L} = \max_{1 \leq j \leq N} L_j$ and $\bar{F} = F(x^*)$ for some stationary point x^* of (1).

Proof As in the previous section, for a simple exposition we drop k from our derivations: e.g. the current point is denoted x , and $x^+ = x + U_i d_i$, where direction d_i is given by Algorithm (1-RCD) for some random selection of index i . Similarly, we use (ϕ, ϕ^+, ξ) instead of $(\phi^k, \phi^{k+1}, \xi^{k-1})$. From the Lipschitz continuity relation (13) we have:

$$f(x) + \langle \nabla f(x), y - x \rangle \leq f(y) + \frac{L_f}{2} \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n.$$

Adding the term $\frac{1}{2}\|x - y\|_L^2 + h(y) + (N - 1)F(x)$ in both sides of the previous inequality and then minimizing w.r.t. $s = y - x$ we get:

$$\begin{aligned} & \min_{s \in \mathbb{R}^n} f(x) + \langle \nabla f(x), s \rangle + \frac{1}{2}\|s\|_L^2 + h(x + s) + (N - 1)F(x) \\ & \leq \min_{s \in \mathbb{R}^n} F(x + s) + \frac{L_f}{2}\|s\|^2 + \frac{1}{2}\|s\|_L^2 + (N - 1)F(x). \end{aligned}$$

Based on the definition of ψ_L we have:

$$\begin{aligned} \psi_L(d_L(x); x) + (N - 1)F(x) & \leq \min_{s \in \mathbb{R}^n} F(x + s) + \frac{L_f + \bar{L}}{2}\|s\|^2 + (N - 1)F(x) \\ & \leq F(x^*) + \frac{L_f + \bar{L}}{2}\|x - x^*\|^2 + (N - 1)F(x), \end{aligned}$$

for any x^* stationary point, i.e. $x^* \in X^*$. By taking expectation w.r.t. ξ and dividing by N , results in:

$$\frac{1}{N}E[\psi_L(d_L(x); x)] + \left(1 - \frac{1}{N}\right)\phi \leq \frac{1}{N}\left(F(x^*) + \frac{L_f + \bar{L}}{2}E[\|x - x^*\|^2] + (N - 1)\phi\right).$$

Now, we come back to the notation dependent on k . Since the sequence $F(x^k)$ is nonincreasing (according to Lemma 3), then $F(x^k) \leq F(x^0)$ for all k . Further, $M_1(x, \mathbf{1})$ converges to 0 a.s. according to Theorem 1 and inequality (15). Then, from Assumption 2 it follows that there exist $\tau > 0$ and \bar{k} such that

$$\|x^k - \bar{x}^k\| \leq \tau M_1(x, \mathbf{1}) \quad \forall k \geq \bar{k},$$

where $\bar{x}^k \in X^*$ satisfies $\|x^k - \bar{x}^k\| = \text{dist}(x^k, X^*)$. It also follows that $\|x^k - \bar{x}^k\|$ converges to 0 a.s. and then using the second part of Assumption 2 we can conclude that eventually the sequence \bar{x}^k settles down at some isocost surface of F (see also [31]), i.e. there exists some $\hat{k} \geq \bar{k}$ and a scalar \bar{F} such that

$$F(\bar{x}^k) = \bar{F} \quad \forall k \geq \hat{k}.$$

Using (11), assuming $k \geq \hat{k}$ and taking into account that $\bar{x}^k \in X^*$, i.e. \bar{x}^k is a stationary point, we have:

$$\phi^{k+1} \leq \frac{1}{N}\left(\bar{F} + \tau \frac{L_f + \bar{L}}{2}E\left[\|d_1(x^k)\|^2\right] + (N - 1)\phi^k\right).$$

Further, by combining (12) and (15) we get:

$$\phi^{k+1} \leq \frac{1}{N}\left(\bar{F} + N\tau\tau_L(L_f + \bar{L})\left(\phi^k - \phi^{k+1}\right) + (N - 1)\phi^k\right),$$

Multiplying with N we get:

$$\phi^{k+1} - \bar{F} \leq (N\tau\tau_L(L_f + \bar{L}) + N - 1)\left(\phi^k - \bar{F} + \bar{F} - \phi^{k+1}\right).$$

Finally, we get the linear convergence of the sequence ϕ^k :

$$\phi^{k+1} - \bar{F} \leq \left(1 - \frac{1}{N\tau\tau_L(L_f + \bar{L}) + N}\right)\left(\phi^k - \bar{F}\right).$$

□

In [31], Tseng obtained a similar result for a block coordinate descent method with greedy (Gauss-Southwell) index selection. However, due to randomization, our Algorithm (1-RCD) has a much lower complexity per iteration than the complexity per iteration of Tseng’s coordinate descent algorithm.

3 Constrained minimization of composite objective functions

In this section we present a variant of random block coordinate gradient descent method for solving large-scale nonconvex optimization problems with composite objective function and a single linear equality constraint:

$$\begin{aligned}
 F^* &= \min_{x \in \mathbb{R}^n} F(x) \quad (:= f(x) + h(x)) \\
 \text{s.t.} & \quad a^T x = b,
 \end{aligned}
 \tag{16}$$

where $a \in \mathbb{R}^n$ is a nonzero vector and functions f and h satisfy similar conditions as in Assumption 1. In particular, the smooth and nonsmooth part of the objective function in (16) satisfy:

Assumption 3 (i) The function f has 2-block coordinate Lipschitz continuous gradient, i.e. there are constants $L_{ij} > 0$ such that:

$$\|\nabla_{ij} f(x + U_i s_i + U_j s_j) - \nabla_{ij} f(x)\| \leq L_{ij} \|s_{ij}\|$$

for all $s_{ij} = [s_i^T \ s_j^T]^T \in \mathbb{R}^{n_i+n_j}$, $x \in \mathbb{R}^n$ and $i, j = 1, \dots, N$.

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

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

where the functions $h_i : \mathbb{R} \rightarrow \mathbb{R}$ are convex for all $i = 1, \dots, n$.

Note that these assumptions are frequently used in the area of coordinate descent methods for convex minimization, e.g. [2, 13–15, 31]. Based on this assumption the first order necessary optimality conditions become: if x^* is a local minimum of (16), then there exists a scalar λ^* such that:

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

Any vector x^* satisfying this relation is called a *stationary point* for nonconvex problem (16). For a simpler exposition in the following sections we use a context-dependent notation as follows: let $x = \sum_{i=1}^N U_i x_i \in \mathbb{R}^n$ and $x_{ij} = [x_i^T \ x_j^T]^T \in \mathbb{R}^{n_i+n_j}$, then by addition with a vector in the extended space $y \in \mathbb{R}^n$, i.e., $y + x_{ij}$, we understand $y + U_i x_i + U_j x_j$. Also, by the inner product $\langle y, x_{ij} \rangle$ we understand $\langle y, x_{ij} \rangle = \langle y_i, x_i \rangle + \langle y_j, x_j \rangle$. Based on Assumption 3 (i) the following inequality holds [14]:

$$|f(x + s_{ij}) - f(x) + \langle \nabla_{ij} f(x), s_{ij} \rangle| \leq \frac{L_{ij}}{2} \|s_{ij}\|^2 \quad \forall x \in \mathbb{R}^n, \ s_{ij} \in \mathbb{R}^{n_i+n_j} \tag{17}$$

and then we can bound the function F with the following quadratic expression:

$$F(x + s_{ij}) \leq f(x) + \langle \nabla_{ij} f(x), s_{ij} \rangle + \frac{L_{ij}}{2} \|s_{ij}\|^2 + h(x + s_{ij}) \quad \forall s_{ij} \in \mathbb{R}^{n_i+n_j}, \ x \in \mathbb{R}^n. \tag{18}$$

Given local Lipschitz constants $L_{ij} > 0$ for $i \neq j \in \{1, \dots, N\}$, we define the vector $\Gamma \in \mathbb{R}^N$ with the components $\Gamma_i = \frac{1}{N} \sum_{j=1}^N L_{ij}$, the diagonal matrix $D_\Gamma = \text{diag}(\Gamma_1 I_{n_1}, \dots, \Gamma_N I_{n_N}) \in \mathbb{R}^{n \times n}$ and the following pair of dual norms:

$$\|x\|_\Gamma = \left(\sum_{i=1}^N \Gamma_i \|x_i\|^2 \right)^{1/2} \quad \forall x \in \mathbb{R}^n, \quad \|y\|_\Gamma^* = \left(\sum_{i=1}^N \Gamma_i^{-1} \|y_i\|^2 \right)^{1/2} \quad \forall y \in \mathbb{R}^n.$$

3.1 A 2-random coordinate descent algorithm

Let (i, j) be a two dimensional random variable, where $i, j \in \{1, \dots, N\}$ with $i \neq j$ and $p_{i_k j_k} = \Pr((i, j) = (i_k, j_k))$ be its probability distribution. Given a feasible x , two blocks are chosen randomly with respect to a given probability distribution p_{ij} and the quadratic model (18) is minimized with respect to these coordinates. Our method has the following iteration: given a feasible initial point x^0 , that is $a^T x^0 = b$, then for all $k \geq 0$

Algorithm (2-RCD)

1. Choose randomly 2 block coordinates (i_k, j_k) with probability $p_{i_k j_k}$
2. Set $x^{k+1} = x^k + U_{i_k} d_{i_k} + U_{j_k} d_{j_k}$,

where directions $d_{i_k j_k} = [d_{i_k}^T \ d_{j_k}^T]^T$ minimize the quadratic model (18):

$$\begin{aligned} d_{i_k j_k} &= \arg \min_{s_{i_k j_k}} f(x^k) + \langle \nabla_{i_k j_k} f(x^k), s_{i_k j_k} \rangle + \frac{L_{i_k j_k}}{2} \|s_{i_k j_k}\|^2 + h(x^k + s_{i_k j_k}) \\ \text{s.t.} \quad &a_{i_k}^T s_{i_k} + a_{j_k}^T s_{j_k} = 0. \end{aligned} \tag{19}$$

The reader should note that for problems with simple separable functions h (e.g. box indicator function (5), ℓ_1 -regularization function (6)) the arithmetic complexity of computing the direction d_{ij} is $\mathcal{O}(n_i + n_j)$ (see [14, 31] for a detailed discussion). 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. box indicator function or ℓ_1 -regularization function) [14]. Note that other (coordinate descent) methods designed for solving nonconvex problems subject to a single linear equality constraint have complexity per iteration at least of order $\mathcal{O}(n)$ [2, 10, 29, 31]. We can consider more than one equality constraint in the optimization model (16). However, in this case the analysis of Algorithm (2-RCD) is involved and the complexity per iteration is much higher (see [14, 31] for a detailed discussion).

We assume that for every pair (i, j) we have $p_{ij} = p_{ji}$ and $p_{ii} = 0$, resulting in $\frac{N(N-1)}{2}$ different pairs (i, j) . We define the subspace $S = \{s \in \mathbb{R}^n : a^T s = 0\}$ and the local subspace w.r.t. the pair (i, j) as $S_{ij} = \{x \in S : x_l = 0 \ \forall l \neq i, j\}$. Also, we denote $\xi^k = \{(i_0, j_0), \dots, (i_k, j_k)\}$ and $\phi^k = E[F(x^k)]$ for the expectation taken w.r.t. ξ^{k-1} . Given a constant $\alpha > 0$ and a vector with positive entries $L \in \mathbb{R}^N$, the following property is valid for ψ_L :

$$\psi_{\alpha L}(s; x) = f(x) + \langle \nabla f(x), s \rangle + \frac{\alpha}{2} \|s\|_L^2 + h(x + s). \tag{20}$$

Since in this section we deal with linearly constrained problems, we need to adapt the definition for the map $d_L(x)$ introduced in Sect. 2. Thus, for any vector with positive entries $L \in \mathbb{R}^N$ and $x \in \mathbb{R}^n$, we define the following map:

$$d_L(x) = \arg \min_{s \in S} f(x) + \langle \nabla f(x), s \rangle + \frac{1}{2} \|s\|_L^2 + h(x + s). \tag{21}$$

In order to analyze the convergence of Algorithm (2-RCD), we introduce an *optimality measure*:

$$M_2(x, \Gamma) = \|D_\Gamma \cdot d_{N\Gamma}(x)\|_\Gamma^*.$$

Lemma 5 *For any given vector $\tilde{\Gamma}$ with positive entries, a vector $x^* \in \mathbb{R}^n$ is a stationary point for problem (16) if and only if the quantity $M_2(x^*, \tilde{\Gamma}) = 0$.*

Proof Based on the optimality conditions of subproblem (21), it can be easily shown that if $M_2(x^*, \tilde{\Gamma}) = 0$, then x^* is a stationary point for the original problem (16). We prove the converse implication by contradiction. Assume that x^* is a stationary point for (16) and $M_2(x^*, \tilde{\Gamma}) > 0$. It follows that $d_{N\tilde{\Gamma}}(x^*)$ is a nonzero solution of subproblem (21) for $x = x^*$. Then, there exist the subgradients $g(x^*) \in \partial h(x^*)$ and $g(x^* + d_{N\tilde{\Gamma}}(x^*)) \in \partial h(x^* + d_{N\tilde{\Gamma}}(x^*))$ and two scalars $\gamma, \lambda \in \mathbb{R}$ such that the optimality conditions for optimization problems (16) and (21) can be written as:

$$\begin{cases} \nabla f(x^*) + g(x^*) + \lambda a = 0 \\ \nabla f(x^*) + D_{N\tilde{\Gamma}}d_{N\tilde{\Gamma}}(x^*) + g(x^* + d_{N\tilde{\Gamma}}(x^*)) + \gamma a = 0. \end{cases}$$

Taking the difference of the two relations above and considering the inner product with $d_{N\tilde{\Gamma}}(x^*) \neq 0$ on both sides of the equation, we get:

$$\|d_{N\tilde{\Gamma}}(x^*)\|_{\tilde{\Gamma}}^2 + \frac{1}{N} \langle g(x^* + d_{N\tilde{\Gamma}}(x^*)) - g(x^*), d_{N\tilde{\Gamma}}(x^*) \rangle = 0,$$

where we used that $a^T d_{N\tilde{\Gamma}}(x^*) = 0$. From convexity of the function h we see that both terms in the above sum are nonnegative and thus $d_{N\tilde{\Gamma}}(x^*) = 0$, which contradicts our hypothesis. In conclusion, we get $M_2(x^*, \tilde{\Gamma}) = 0$. □

3.2 Convergence of algorithm (2-RCD)

In order to provide the convergence results of Algorithm (2-RCD), we have to introduce some definitions and auxiliary results. We denote by $\text{supp}(x)$ the set of indexes corresponding to the nonzero coordinates in the vector $x \in \mathbb{R}^n$.

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 $j = 1, \dots, n$.

We introduce the notion of elementary vectors for the linear subspace $S = \text{Null}(a^T)$.

Definition 2 An elementary vector d of S is a vector $d \in S$ for which there is no nonzero $d' \in S$ conformal to d and $\text{supp}(d') \neq \text{supp}(d)$.

We now present some results for elementary vectors and conformal realization, whose proofs can be found in [26, 27, 31]. A particular case of Exercise 10.6 in [27] and an interesting result in [26] provide us the following lemma:

Lemma 6 [26, 27] *Given $d \in S$, if d is an elementary vector, then $|\text{supp}(d)| \leq 2$. Otherwise, d has a conformal realization $d = d^1 + \dots + d^s$, where $s \geq 2$ and $d^t \in S$ are elementary vectors conformal to d for all $t = 1, \dots, s$.*

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

Lemma 7 [31] *Let h be componentwise separable and convex. For any $x, x + d \in \text{dom}h$, let d be expressed as $d = d^1 + \dots + d^s$ for some $s \geq 2$ and some nonzero $d^t \in \mathbb{R}^n$ conformal to d for all $t = 1, \dots, s$. Then,*

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

where $d^t \in S$ are elementary vectors conformal to d for all $t = 1, \dots, s$.

Lemma 8 *If Assumption 3 holds and sequence x^k is generated by Algorithm (2-RCD) using the uniform distribution, then the following inequality is valid:*

$$\begin{aligned} & E \left[\psi_{L_{i_k j_k} \mathbf{1}} \left(d_{i_k j_k}; x^k \right) \mid \xi^{k-1} \right] \\ & \leq \left(1 - \frac{2}{N(N-1)} \right) F \left(x^k \right) + \frac{2}{N(N-1)} \psi_{N\Gamma} \left(d_{N\Gamma} \left(x^k \right); x^k \right) \quad \forall k \geq 0. \end{aligned}$$

Proof As in the previous sections, for a simple exposition we drop k from our derivations: e.g. the current point is denoted x , next iterate $x^+ = x + U_i d_i + U_j d_j$, where direction d_{ij} is given by Algorithm (2-RCD) for some random selection of pair (i, j) and ξ instead of ξ^{k-1} . From the relation (20) and the property of minimizer d_{ij} we have:

$$\psi_{L_{ij} \mathbf{1}} \left(d_{ij}; x \right) \leq \psi_{L_{ij} \mathbf{1}} \left(s_{ij}; x \right) \quad \forall s_{ij} \in S_{ij}.$$

Taking expectation in both sides w.r.t. random variable (i, j) conditioned on ξ and recalling that $p_{ij} = \frac{2}{N(N-1)}$, we get:

$$\begin{aligned} & E[\psi_{L_{ij} \mathbf{1}}(d_{ij}; x) \mid \xi] \\ & \leq f(x) + \frac{2}{N(N-1)} \left[\sum_{i,j} \langle \nabla_{ij} f(x), s_{ij} \rangle + \sum_{i,j} \frac{L_{ij}}{2} \|s_{ij}\|^2 + \sum_{i,j} h(x + s_{ij}) \right] \\ & = f(x) + \frac{2}{N(N-1)} \left[\sum_{i,j} \langle \nabla_{ij} f(x), s_{ij} \rangle + \sum_{i,j} \frac{1}{2} \|\sqrt{L_{ij}} s_{ij}\|^2 + \sum_{i,j} h(x + s_{ij}) \right], \end{aligned}$$

for all $s_{ij} \in S_{ij}$. We can apply Lemma 7 for coordinatewise separable functions $\|\cdot\|^2$ and $h(\cdot)$ and we obtain:

$$\begin{aligned} E[\psi_{L_{ij} \mathbf{1}}(d_{ij}; x) \mid \xi] & \leq f(x) + \frac{2}{N(N-1)} \left[\left\langle \nabla f(x), \sum_{i,j} s_{ij} \right\rangle + \frac{1}{2} \left\| \sum_{i,j} \sqrt{L_{ij}} s_{ij} \right\|^2 \right. \\ & \quad \left. + h \left(x + \sum_{i,j} s_{ij} \right) + \left(\frac{N(N-1)}{2} - 1 \right) h(x), \right] \end{aligned}$$

for all $s_{ij} \in S_{ij}$. From Lemma 6 it follows that any $s \in S$ has a conformal realization defined by $s = \sum_t s^t$, where the vectors $s^t \in S$ are elementary vectors conformal to s . Therefore, observing that every elementary vector s^t has at most two nonzero blocks, then any vector $s \in S$ can be generated by $s = \sum_{i,j} s_{ij}$, where $s_{ij} \in S$ are conformal to s and have at most two nonzero blocks, i.e. $s_{ij} \in S_{ij}$ for some pair (i, j) . Due to conformal property of the vectors s_{ij} , the expression $\|\sum_{i,j} \sqrt{L_{ij}} s_{ij}\|^2$ is nondecreasing in the weights L_{ij} and taking in account that $L_{ij} \leq \min\{N\Gamma_i, N\Gamma_j\}$, the previous inequality leads to:

$$\begin{aligned}
 & E[\psi_{L_{ij}\mathbf{1}}(d_{ij}; x) | \xi] \\
 & \leq f(x) + \frac{2}{N(N-1)} \left[\left\langle \nabla f(x), \sum_{i,j} s_{ij} \right\rangle + \frac{1}{2} \left\| \sum_{i,j} D_{N\Gamma}^{1/2} s_{ij} \right\|^2 + h \left(x + \sum_{i,j} s_{ij} \right) \right. \\
 & \quad \left. + \left(\frac{N(N-1)}{2} - 1 \right) h(x) \right] \\
 & = f(x) + \frac{2}{N(N-1)} \left[\langle \nabla f(x), s \rangle + \frac{1}{2} \|\sqrt{N} D_{\Gamma}^{1/2} s\|^2 + h(x+s) + \left(\frac{N(N-1)}{2} - 1 \right) h(x), \right]
 \end{aligned}$$

for all $s \in S$. As the last inequality holds for any vector $s \in S$, it also holds for the particular vector $d_{N\Gamma}(x) \in S$:

$$\begin{aligned}
 E[\psi_{L_{ij}\mathbf{1}}(d_{ij}; x) | \xi] & \leq \left(1 - \frac{2}{N(N-1)} \right) F(x) + \frac{2}{N(N-1)} \\
 & \quad \times \left[f(x) + \langle \nabla f(x), d_{N\Gamma}(x) \rangle + \frac{N}{2} \|d_{N\Gamma}(x)\|_{\Gamma}^2 + h(x+d_{N\Gamma}(x)) \right] \\
 & = \left(1 - \frac{2}{N(N-1)} \right) F(x) + \frac{2}{N(N-1)} \psi_{N\Gamma}(d_{N\Gamma}(x); x).
 \end{aligned}$$

□

The main convergence properties of Algorithm (2-RCD) are given in the following theorem:

Theorem 4 *If Assumption 3 holds for the composite objective function F of problem (16) and the sequence x^k is generated by Algorithm (2-RCD) using the uniform distribution, then the following statements are valid:*

- (i) *The sequence of random variables $M_2(x^k, \Gamma)$ converges to 0 a.s. and the sequence $F(x^k)$ converges to a random variable \bar{F} a.s.*
- (ii) *Any accumulation point of the sequence x^k is a stationary point for optimization problem (16).*

Proof (i) Using a similar reasoning as in Lemma 3 but for the inequality (18) we can show the following decrease in the objective function for Algorithm (2-RCD) (i.e. Algorithm (2-RCD) is also a descent method):

$$F(x^{k+1}) \leq F(x^k) - \frac{L_{ik,jk}}{2} \|d_{ik,jk}\|^2 \quad \forall k \geq 0. \tag{22}$$

Further, subtracting F^* from both sides, applying expectation conditioned on ξ^{k-1} and then using supermartingale convergence theorem given in Lemma 2 we obtain that $F(x^k)$ converges to a random variable \bar{F} a.s. for $k \rightarrow \infty$. Due to almost sure convergence of sequence $F(x^k)$, it can be easily seen that $\lim_{k \rightarrow \infty} F(x^k) - F(x^{k+1}) = 0$ a.s. Moreover, from (22) we have:

$$\frac{L_{ik,jk}}{2} \|d_{ik,jk}\|^2 = \frac{L_{ik,jk}}{2} \|x^{k+1} - x^k\|^2 \leq F(x^k) - F(x^{k+1}) \quad \forall k \geq 0,$$

which implies that

$$\lim_{k \rightarrow \infty} d_{ik,jk} = 0 \quad \text{and} \quad \lim_{k \rightarrow \infty} \|x^{k+1} - x^k\| = 0 \quad \text{a.s.}$$

As in the previous section, for a simple exposition we drop k from our derivations: e.g. the current point is denoted x , next iterate $x^+ = x + U_i d_i + U_j d_j$, where direction d_{ij} is given by Algorithm (2-RCD) for some random selection of pair (i, j) and ξ stands for ξ^{k-1} . From Lemma 8, we obtain a sequence which bounds from below $\psi_{N\Gamma}(d_{N\Gamma}(x); x)$ as follows:

$$\frac{N(N-1)}{2} E[\psi_{L_{ij}1}(d_{ij}; x) | \xi] + \left(1 - \frac{N(N-1)}{2}\right) F(x) \leq \psi_{N\Gamma}(d_{N\Gamma}(x); x).$$

On the other hand, from Lemma 6 it follows that any $s \in S$ has a conformal realization defined by $s = \sum_{i,j} s_{ij}$, where $s_{ij} \in S$ are conformal to s and have at most two nonzero blocks, i.e. $s_{ij} \in S_{ij}$ for some pair (i, j) . Using now Jensen inequality we derive another sequence which bounds $\psi_{N\Gamma}(d_{N\Gamma}(x); x)$ from above:

$$\begin{aligned} \psi_{N\Gamma}(d_{N\Gamma}(x); x) &= \min_{s \in S} f(x) + \langle \nabla f(x), s \rangle + \frac{1}{2} \|s\|_{N\Gamma}^2 + h(x + s) \\ &= \min_{s_{ij} \in S_{ij}} \left[f(x) + \left\langle \nabla f(x), \sum_{i,j} s_{ij} \right\rangle + \frac{1}{2} \left\| \sum_{i,j} s_{ij} \right\|_{N\Gamma}^2 + h\left(x + \sum_{i,j} s_{ij}\right) \right] \\ &= \min_{\tilde{s}_{ij} \in S_{ij}} f(x) + \frac{1}{N(N-1)} \left\langle \nabla f(x), \sum_{i,j} \tilde{s}_{ij} \right\rangle + \frac{1}{2} \left\| \frac{1}{N(N-1)} \sum_{i,j} \tilde{s}_{ij} \right\|_{N\Gamma}^2 \\ &\quad + h\left(x + \frac{1}{N(N-1)} \sum_{i,j} \tilde{s}_{ij}\right) \\ &\leq \min_{\tilde{s}_{ij} \in S_{ij}} f(x) + \frac{1}{N(N-1)} \sum_{i,j} \langle \nabla f(x), \tilde{s}_{ij} \rangle + \frac{1}{2N(N-1)} \sum_{i,j} \|\tilde{s}_{ij}\|_{N\Gamma}^2 \\ &\quad + \frac{1}{N(N-1)} \sum_{i,j} h(x + \tilde{s}_{ij}) = E[\psi_{N\Gamma}(d_{ij}; x) | \xi], \end{aligned}$$

where we used the notation $\tilde{s}_{ij} = N(N-1)s_{ij}$. If we come back to the notation dependent on k , then using Assumption 3 (ii) and the fact that $d_{i_k j_k} \rightarrow 0$ a.s. we obtain that $E[\psi_{N\Gamma}(d_{i_k j_k}; x^k) | \xi^{k-1}]$ converges to \bar{F} a.s. for $k \rightarrow \infty$. We conclude that both sequences, lower and upper bounds of $\psi_{N\Gamma}(d_{N\Gamma}(x^k); x^k)$ from above, converge to \bar{F} a.s., hence $\psi_{N\Gamma}(d_{N\Gamma}(x^k); x^k)$ converges to \bar{F} a.s. for $k \rightarrow \infty$. A trivial case of strong convexity relation (8) leads to:

$$\psi_{N\Gamma}(0; x^k) \geq \psi_{N\Gamma}(d_{N\Gamma}(x^k); x^k) + \frac{N}{2} \|d_{N\Gamma}(x^k)\|_{\Gamma}^2.$$

Note that $\psi_{N\Gamma}(0; x^k) = F(x^k)$ and since both sequences $\psi_{N\Gamma}(0; x^k)$ and $\psi_{N\Gamma}(d_{N\Gamma}(x^k); x^k)$ converge to \bar{F} a.s. for $k \rightarrow \infty$, from the above strong convexity relation it follows that the sequence $M_2(x^k; \Gamma) = \|d_{N\Gamma}(x^k)\|_{\Gamma}$ converges to 0 a.s. for $k \rightarrow \infty$.

(ii) The proof follows the same ideas as in the proof of Theorem 1 (ii). □

We now present the convergence rate for Algorithm (2-RCD).

Theorem 5 *Let F satisfy Assumption 3. Then, the Algorithm (2-RCD) based on the uniform distribution generates a sequence x^k satisfying the following convergence rate for the expected values of the optimality measure:*

$$\min_{0 \leq l \leq k} E \left[\left(M_2(x^l, \Gamma) \right)^2 \right] \leq \frac{N(F(x^0) - F^*)}{k + 1} \quad \forall k \geq 0.$$

Proof Given the current feasible point x , denote $x^+ = x + U_i d_i + U_j d_j$ as the next iterate, where direction (d_i, d_j) is given by Algorithm (2-RCD) for some random chosen pair (i, j) and we use the notation (ϕ, ϕ^+, ξ) instead of $(\phi^k, \phi^{k+1}, \xi^{k-1})$. Based on Lipschitz inequality (18) we derive:

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

Taking expectation conditioned on ξ in both sides and using Lemma 8 we get:

$$E [F(x^+) | \xi] \leq \left(1 - \frac{2}{N(N-1)}\right) F(x) + \frac{2}{N(N-1)} \psi_{N\Gamma}(d_{N\Gamma}(x); x).$$

Taking now expectation w.r.t. ξ , we can derive:

$$\begin{aligned} \phi - \phi^+ &\geq E[\psi_{N\Gamma}(0; x)] - \left(1 - \frac{2}{N(N-1)}\right) E[\psi_{N\Gamma}(0; x)] \\ &\quad - \frac{2}{N(N-1)} E[\psi_{N\Gamma}(d_{N\Gamma}(x); x)] \\ &= \frac{2}{N(N-1)} (E[\psi_{N\Gamma}(0; x)] - E[\psi_{N\Gamma}(d_{N\Gamma}(x); x)]) \\ &\geq \frac{1}{N-1} E[\|d_{N\Gamma}(x)\|_\Gamma^2] \geq \frac{1}{N} E[(M_2(x, \Gamma))^2], \end{aligned}$$

where we used the strong convexity property of function $\psi_{N\Gamma}(s; x)$. Now, considering iteration k and summing up with respect to entire history we get:

$$\frac{1}{N} \sum_{l=0}^k E \left[(M_2(x^l, \Gamma))^2 \right] \leq F(x^0) - F^*.$$

This inequality leads us to the above result. □

3.3 Constrained minimization of smooth objective functions

We now study the convergence of Algorithm (2-RCD) on the particular case of optimization model (16) with $h = 0$. For this particular case a feasible point x^* is a *stationary point* for (16) if there exists $\lambda^* \in \mathbb{R}$ such that:

$$\nabla f(x^*) + \lambda^* a = 0 \quad \text{and} \quad a^T x^* = b. \tag{23}$$

For any feasible point x , note that exists $\lambda \in \mathbb{R}$ such that:

$$\nabla f(x) = \nabla f(x)_\perp - \lambda a,$$

where $\nabla f(x)_\perp$ is the projection of the gradient vector $\nabla f(x)$ onto the subspace S orthogonal to the vector a . Since $\nabla f(x)_\perp = \nabla f(x) + \lambda a$, we defined a particular optimality measure:

$$M_3(x, \mathbf{1}) = \|\nabla f(x)_\perp\|.$$

In this case the iteration of Algorithm (2-RCD) is a projection onto a hyperplane so that the direction $d_{i_k j_k}$ can be computed in closed form. We denote by $Q_{ij} \in \mathbb{R}^{n \times n}$ the symmetric matrix with all blocks zeros except:

$$Q_{ij}^{ii} = I_{n_i} - \frac{a_i a_i^T}{a_i^T a_i}, \quad Q_{ij}^{jj} = -\frac{a_i a_j^T}{a_{ij}^T a_{ij}}, \quad Q_{ij}^{jj} = I_{n_j} - \frac{a_j a_j^T}{a_{ij}^T a_{ij}}.$$

It is straightforward to see that Q_{ij} is positive semidefinite (notation $Q_{ij} \geq 0$) and $Q_{ij}a = 0$ for all pairs (i, j) with $i \neq j$. Given a probability distribution p_{ij} , let us define the matrix:

$$Q = \sum_{i,j} \frac{p_{ij}}{L_{ij}} Q_{ij},$$

that is also symmetric and positive semidefinite, since $L_{ij}, p_{ij} > 0$ for all (i, j) . Furthermore, since we consider all possible pairs (i, j) , with $i \neq j \in \{1, \dots, N\}$, it can be shown that the matrix Q has an eigenvalue $v_1(Q) = 0$ (which is a simple eigenvalue) with the associated eigenvector a . It follows that $v_2(Q)$ (the second smallest eigenvalue of Q) is positive. Since $h = 0$, we have $F = f$. Using the same reasoning as in the previous sections we can easily show that the sequence $f(x^k)$ satisfies the following decrease:

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L_{ij}} \nabla f(x^k)^T Q_{ij} \nabla f(x^k) \quad \forall k \geq 0. \tag{24}$$

We now give the convergence rate of Algorithm (2-RCD) for this particular case:

Theorem 6 *Let $h = 0$ and f satisfy Assumption 3 (i). Then, Algorithm (2-RCD) based on a general probability distribution p_{ij} generates a sequence x^k satisfying the following convergence rate for the expected values of the norm of the projected gradients onto subspace S :*

$$\min_{0 \leq l \leq k} E \left[\left(M_3(x^l, \mathbf{1}) \right)^2 \right] \leq \frac{2(F(x^0) - F^*)}{v_2(Q)(k + 1)}.$$

Proof As in the previous section, for a simple exposition we drop k from our derivations: e.g. the current point is denoted x , and $x^+ = x + U_i d_i + U_j d_j$, where direction d_{ij} is given by Algorithm (2-RCD) for some random selection of pair (i, j) . Since $h = 0$, we have $F = f$. From (24) we have the following decrease: $f(x^+) \leq f(x) - \frac{1}{2L_{ij}} \nabla f(x)^T Q_{ij} \nabla f(x)$. Taking now expectation conditioned in ξ in this inequality we have:

$$E[f(x^+) | \xi] \leq f(x) - \frac{1}{2} \nabla f(x)^T Q \nabla f(x).$$

From the above decomposition of the gradient $\nabla f(x) = \nabla f(x)_\perp - \lambda a$ and the observation that $Qa = 0$, we conclude that the previous inequality does not change if we replace $\nabla f(x)$ with $\nabla f(x)_\perp$:

$$E[f(x^+) | \xi] \leq f(x) - \frac{1}{2} \nabla f(x)_\perp^T Q \nabla f(x)_\perp.$$

Note that $\nabla f(x)_\perp$ is included in the orthogonal complement of the span of vector a , so that the above inequality can be relaxed to:

$$E[f(x^+) | \xi] \leq f(x) - \frac{1}{2} v_2(Q) \|\nabla f(x)_\perp\|^2 = f(x) - \frac{v_2(Q)}{2} (M_3(x, \mathbf{1}))^2. \tag{25}$$

Coming back to the notation dependent on k and taking expectation in both sides of inequality (25) w.r.t. ξ^{k-1} , we have:

$$\phi^k - \phi^{k+1} \geq \frac{v_2(Q)}{2} E \left[\left(M_3(x^k, \mathbf{1}) \right)^2 \right].$$

Summing w.r.t. the entire history, we obtain the above result. □

Note that our convergence proofs given in this section (Theorems 4, 5 and 6) are different from the convex case [13–15], since here we introduce another optimality measure and we use supermartingale convergence theorem in the analysis. It is important to see that the convergence rates for the Algorithm (2-RCD) given in Theorems 5 and 6 are typical for the class of first order methods designed for solving nonconvex and nonsmooth optimization problems, e.g. in [2, 19] similar results are obtained for other gradient based methods designed to solve nonconvex problems.

4 Numerical experiments

In this section we analyze the practical performance of the random coordinate descent methods derived in this paper and compare our algorithms with some recently developed state-of-the-art algorithms from the literature. Coordinate descent methods are one of the most efficient classes of algorithms for large-scale optimization problems. Therefore, we present extensive numerical simulation for large-scale nonconvex problems with dimension ranging from $n = 10^3$ to $n = 10^7$. For numerical experiments, we implemented all the algorithms in C code and we performed our tests on a PC with Intel Xeon E5410 CPU and 8 Gb RAM memory.

For tests we choose as application the *eigenvalue complementarity problem*. It is well-known that many applications from mathematics, physics and engineering require the efficient computation of eigenstructure of some symmetric matrix. A brief list of these applications includes optimal control, stability analysis of dynamic systems, structural dynamics, electrical networks, quantum chemistry, chemical reactions and economics (see [7, 12, 20, 29] and the reference therein for more details). The eigenvalues of a symmetric matrix A have an elementary definition as the roots of the characteristic polynomial $\det(A - \lambda I)$. In realistic applications the eigenvalues can have an important role, for example to describe expected long-time behavior of a dynamical system, or to be only intermediate values of a computational method. For many applications the optimization approach for eigenvalues computation is better than the algebraic one. Although, the eigenvalues computation can be formulated as a convex problem, the corresponding feasible set is complex so that the projection on this set is numerically very expensive, at least of order $\mathcal{O}(n^2)$. Therefore, classical methods for convex optimization are not adequate for large-scale eigenvalue problems. To obtain a lower iteration complexity as $\mathcal{O}(n)$ or even $\mathcal{O}(p)$, where $p \ll n$, an appropriate way to approach these problems is through nonconvex formulation and using coordinate descent methods. A classical optimization problem formulation involves the Rayleigh quotient as the objective function of some nonconvex optimization problem [12]. The eigenvalue complementarity problem (EiCP) is an extension of the classical eigenvalue problem, which can be stated as: *given matrices A and B , find $v \in \mathbb{R}$ and $x \neq 0$ such that*

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

If matrices A and B are symmetric, then we have *symmetric (EiCP)*. It has been shown in [29] that symmetric (EiCP) is equivalent with finding a stationary point of a generalized Rayleigh quotient on the simplex:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} & \frac{x^T A x}{x^T B x} \\ \text{s.t.} & \mathbf{1}^T x = 1, x \geq 0, \end{aligned}$$

where we recall that $\mathbf{1} = [1 \dots 1]^T \in \mathbb{R}^n$. A widely used alternative formulation of (EiCP) problem is the nonconvex logarithmic formulation (see [8,29]):

$$\begin{aligned} \max_{x \in \mathbb{R}^n} f(x) & \left(= \ln \frac{x^T Ax}{x^T Bx} \right) \\ \text{s.t.} & \mathbf{1}^T x = 1, \quad x \geq 0. \end{aligned} \tag{26}$$

Note that optimization problem (26) is a particular case of (16), where h is the indicator function of the nonnegative orthant. In order to have a well-defined objective function for the logarithmic case, in the most of the aforementioned papers the authors assumed positive definiteness of matrices $A = [a_{ij}]$ and $B = [b_{ij}]$. In this paper, in order to have a more practical application with a highly nonconvex objective function [7], we consider the class of nonnegative matrices, i.e. $A, B \geq 0$, with positive diagonal elements, i.e. $a_{ii} > 0$ and $b_{ii} > 0$ for all $i = 1, \dots, n$. For this class of matrices the problem (26) is also well-defined on the simplex. Based on Perron-Frobenius theorem, we have that for matrices A that are also irreducible and $B = I_n$ the corresponding stationary point of the (EiCP) problem (26) is the global minimum of this problem or equivalently is the Perron vector, so that any accumulation point of the sequence generated by our Algorithm (2-RCD) is also a global minimizer. In order to apply our Algorithm (2-RCD) on the logarithmic formulation of the (EiCP) problem (26), we have to compute an approximation of the Lipschitz constants L_{ij} . For brevity, we introduce the notation $\Delta_n = \{x \in \mathbb{R}^n : \mathbf{1}^T x = 1, x \geq 0\}$ for the standard simplex and the function $g_A(x) = \ln x^T Ax$. For a given matrix A , we denote by $A_{ij} \in \mathbb{R}^{(n_i+n_j) \times (n_i+n_j)}$ the 2×2 block matrix of A by taking the pair (i, j) of block rows of matrix A and then the pair (i, j) of block columns of A .

Lemma 9 *Given a nonnegative matrix $A \in \mathbb{R}^{n \times n}$ such that $a_{ii} \neq 0$ for all $i = 1, \dots, n$, then the function $g_A(x) = \ln x^T Ax$ has 2 block coordinate Lipschitz gradient on the standard simplex, i.e.:*

$$\|\nabla_{ij} g_A(x + s_{ij}) - \nabla_{ij} g_A(x)\| \leq L_{ij}^A \|s_{ij}\|, \quad \forall x, x + s_{ij} \in \Delta_n,$$

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

$$L_{ij}^A \leq \frac{2N}{\min_{1 \leq i \leq N} a_{ii}} \|A_{ij}\|.$$

Proof The Hessian of the function $g_A(x)$ is given by

$$\nabla^2 g_A(x) = \frac{2A}{x^T Ax} - \frac{4(Ax)(Ax)^T}{(x^T Ax)^2}.$$

Note that $\nabla_{ij}^2 g_A(x) = \frac{2A_{ij}}{x^T Ax} - \frac{4(Ax)_{ij}(Ax)_{ij}^T}{(x^T Ax)^2}$. With the same arguments as in [29] we have that: $\|\nabla_{ij}^2 g_A(x)\| \leq \|\frac{2A_{ij}}{x^T Ax}\|$. From the mean value theorem we obtain:

$$\nabla_{ij} g_A(x + s_{ij}) = \nabla_{ij} g_A(x) + \int_0^1 \nabla_{ij}^2 g_A(x + \tau s_{ij}) s_{ij} \, d\tau,$$

for any $x, x + s_{ij} \in \Delta_n$. Taking norm in both sides of the equality results in:

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

Note that $\min_{x \in \Delta_n} x^T A x > 0$ since we have:

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

and the above result can be easily derived. □

Based on the previous notation, the objective function of the logarithmic formulation (26) is given by:

$$\max_{x \in \Delta_n} f(x) \quad (= g_A(x) - g_B(x)) \quad \text{or} \quad \min_{x \in \Delta_n} \bar{f}(x) \quad (= g_B(x) - g_A(x)).$$

Therefore, the local Lipschitz constants L_{ij} of function f are estimated very easily and numerically cheap as:

$$L_{ij} \leq L_{ij}^A + L_{ij}^B = \frac{2N}{\min_{1 \leq i \leq n} a_{ii}} \|A_{ij}\| + \frac{2N}{\min_{1 \leq i \leq n} b_{ii}} \|B_{ij}\| \quad \forall i \neq j.$$

In [29] the authors show that a variant of difference of convex functions (DC) algorithm is very efficient for solving the logarithmic formulation (26). We present extensive numerical experiments for evaluating the performance of our Algorithm (2-RCD) in comparison with the Algorithm (DC). For completeness, we also present the Algorithm (DC) for logarithmic formulation of (EiCP) in the minimization form from [29]: given $x_0 \in \mathbb{R}^n$, for $k \geq 0$ do

Algorithm (DC) [31]

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 QP : $x^{k+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ \frac{\mu}{2} \|x\|^2 - \langle x, y^k \rangle : \mathbf{1}^T x = 1, x \geq 0 \right\}$,

where μ is a parameter chosen in a preliminary stage of the algorithm such that the function $x \mapsto \frac{1}{2} \mu \|x\|^2 + \ln(x^T A x)$ is convex. In both algorithms we use the following stopping criterion: $|f(x^k) - f(x^{k+1})| \leq \epsilon$, where ϵ is some chosen accuracy. Note that Algorithm (DC) is based on full gradient information and in the application (EiCP) the most computations consists of matrix vector multiplication and a projection onto simplex. 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 reduced to $\mathcal{O}(pn)$ operations, where p is the average number of nonzeros in each row of the matrix A and B . Further, there are efficient algorithms for computing the projection onto simplex, e.g. block pivotal principal pivoting algorithm described in [8], whose arithmetic complexity is of order $\mathcal{O}(n)$. As it appears in practice, the value of parameter μ is crucial in the rate of convergence

of Algorithm (DC). The authors in [29] provide an approximation of μ that can be computed easily when the matrix A from (26) is positive definite. However, for general copositive matrices (as the case of nonnegative irreducible matrices considered in this paper) one requires the solution of certain NP-hard problem to obtain a good approximation of parameter μ . On the other hand, for our Algorithm (2-RCD) the computation of the Lipschitz constants L_{ij} is very simple and numerically cheap (see previous lemma). Further, for the scalar case (i.e. $n = N$) the complexity per iteration of our method applied to (EiCP) problem is $\mathcal{O}(p)$ in the sparse case.

In Table 1 we compare the two algorithms: (2-CRD) and (DC). We generated random sparse symmetric nonnegative and irreducible matrices of dimension ranging from $n = 10^3$ to $n = 10^7$ using the uniform distribution. Each row of the matrices has only $p = 10$ nonzero entries. In both algorithms we start from random initial points. In the table we present for each algorithm the final objective function value (F^*), the number of iterations (iter) and the necessary CPU time (in seconds) for our computer to execute all the iterations. As Algorithm (DC) uses the whole gradient information to obtain the next iterate, we also report for Algorithm (2-RCD) the equivalent number of full-iterations which means the total number of iterations divided by $n/2$ (i.e. the number of iterations groups $x^0, x^{n/2}, \dots, x^{kn/2}$). Since computing μ is very difficult for this type of matrices, we try to tune μ in Algorithm (DC). We have tried four values for μ ranging from $0.01n$ to $50n$. We have noticed that if μ is not carefully tuned Algorithm (DC) cannot find the optimal value f^* in a reasonable time. Then, after extensive simulations we find an appropriate value for μ such that Algorithm (DC) produces an accurate approximation of the optimal value. From the table we see that our Algorithm (2-RCD) provides better performance in terms of objective function values and CPU time (in seconds) than Algorithm (DC). We also observe that our algorithm is not sensitive w.r.t. the Lipschitz constants L_{ij} and also w.r.t. the initial point, while Algorithm (DC) is very sensitive to the choice of μ and the initial point.

Further, in Fig. 1 we plot the evolution of the objective function w.r.t. time for Algorithms (2-RCD) and (DC), in logarithmic scale, on a random (EiCP) problem with dimension $n = 5 \cdot 10^5$ (Algorithm (DC) with parameter left: $\mu = 1.42 \cdot n$; right: $\mu = 50 \cdot n$). For a good choice of μ we see that in the initial phase of Algorithm (DC) the reduction in the objective function is very fast, but while approaching the optimum it slows down. On the other hand, due to the sparsity and randomization our proposed algorithm is faster in numerical implementation than the (DC) scheme.

In Fig. 2 we plot the evolution of CPU time, in logarithmic scale, required for solving the problem w.r.t. the average number of nonzeros entries p in each row of the matrix A . We see that for very sparse matrices (i.e. for matrices with relatively small number of nonzeros per row $p \ll n$), our Algorithm (2-RCD) performs faster in terms of CPU time than (DC) method. The main reason is that our method has a simple implementation, does not require the use of other algorithms at each iteration and the arithmetic complexity of an iteration is of order $\mathcal{O}(p)$. On the other hand, Algorithm (DC) is using the block pivotal principal pivoting algorithm described in [8] at each iteration for projection on simplex and the arithmetic complexity of an iteration is of order $\mathcal{O}(pn)$.

We conclude from the theoretical rate of convergence and the previous numerical results that Algorithms (1-RCD) and (2-RCD) are easier to be implemented and analyzed due to the randomization and the typically very simple iteration. Furthermore, on certain classes of problems with sparsity structure, that appear frequently in many large-scale real applications, the practical complexity of our methods is better than that of some well-known methods from the literature. All these arguments make our algorithms to be competitive

Table 1 Performance of algorithms (2-RCD) and (DC) on randomly generated (EiCP) sparse problems with $p = 10$ and random starting point x^0 for different problem dimensions n

n	(DC)				(2-RCD)		
	μ	CPU (s)	Iter	F^*	CPU (sec)	Full-iter	F^*
5×10^3	$0.01n$	0.0001	1	1.32	0.09	56	105.20
	n	0.001	2	82.28			
	$2n$	0.02	18	105.21			
	$50n$	0.25	492	105.21			
2×10^4	$0.01n$	0.01	1	1.56	0.39	50	73.74
	n	0.01	2	59.99			
	$1.43n$	0.59	230	73.75			
	$50n$	0.85	324	73.75			
5×10^4	$0.01n$	0.01	1	1.41	1.75	53	83.54
	n	0.02	2	67.03			
	$1.43n$	1.53	163	83.55			
	$50n$	2.88	324	83.57			
7.5×10^4	$0.01n$	0.01	1	2.40	3.60	61	126.04
	n	0.03	2	101.76			
	$1.45n$	6.99	480	126.05			
	$50n$	4.72	324	126.05			
10^5	$0.01n$	0.02	1	0.83	4.79	53	52.21
	n	0.05	2	41.87			
	$1.43n$	6.48	319	52.22			
	$50n$	6.57	323	52.22			
5×10^5	$0.01n$	0.21	1	2.51	49.84	59	136.37
	n	0.42	2	109.92			
	$1.43n$	94.34	475	136.38			
	$50n$	66.61	324	136.38			
7.5×10^5	$0.01n$	0.44	1	3.11	37.59	38	177.52
	n	0.81	2	143.31			
	$1.43n$	72.80	181	177.52			
	$50n$	135.35	323	177.54			
10^6	$0.01n$	0.67	1	3.60	49.67	42	230.09
	n	1.30	2	184.40			
	$1.43n$	196.38	293	230.09			
	$50n$	208.39	323	230.11			
10^7	$0.01n$	4.69	1	10.83	758.1	41	272.37
	n	22.31	2	218.88			
	$1.45n$	2947.93	325	272.37			
	$50n$	2929.74	323	272.38			

in the large-scale nonconvex optimization framework. Moreover, our methods are suited for recently developed computational architectures (e.g., distributed or parallel architectures [16,25]).

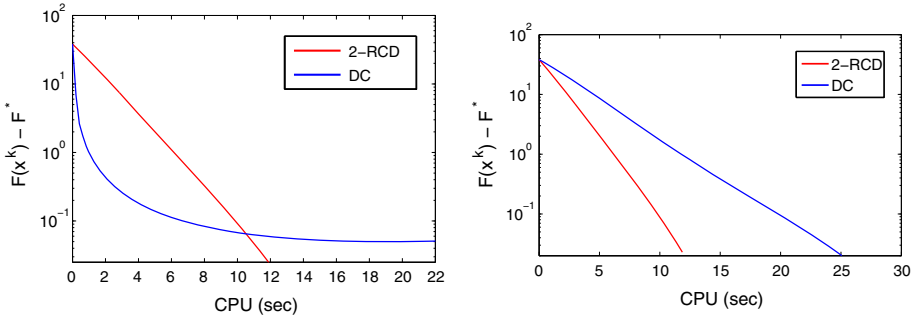
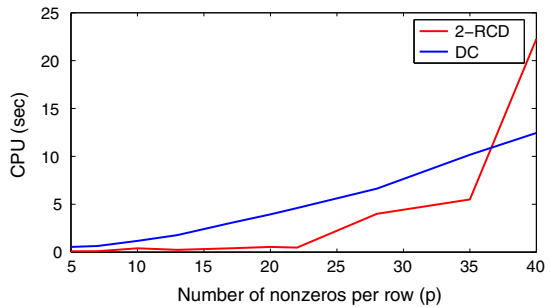


Fig. 1 Performance in terms of function values of Algorithms (2-RCD) and (DC) on a randomly generated (EiCP) problem with $n = 5 \cdot 10^5$: left $\mu = 1.42 \cdot n$ and right $\mu = 50 \cdot n$

Fig. 2 CPU time performance of Algorithms (2-RCD) and (DC) for different values of the sparsity p of the matrix on a randomly generated (EiCP) problem of dimension $n = 2 \times 10^4$



Appendix

Proof of Lemma 4 We derive our proof based on the following remark (see also [5]), for given $u, v \in \mathbb{R}^n$ if $\langle v, u - v \rangle > 0$, then

$$\frac{\|u\|}{\|v\|} \leq \frac{\langle u, u - v \rangle}{\langle v, u - v \rangle}. \tag{27}$$

Let $\alpha > \beta > 0$. Taking $u = \text{prox}_{\alpha h}(x + \alpha d) - x$ and $v = \text{prox}_{\beta h}(x + \beta d) - x$, we show first that inequality $\langle v, u - v \rangle > 0$ holds. Given a real constant $c > 0$, from the optimality conditions corresponding to proximal operator we have:

$$x - \text{prox}_{ch}(x) \in \partial ch(\text{prox}_{ch}(x)).$$

Therefore, from the convexity of h we can derive that:

$$ch(z) \geq ch(\text{prox}_{ch}(y)) + \langle y - \text{prox}_{ch}(y), z - \text{prox}_{ch}(y) \rangle \quad \forall y, z \in \mathbb{R}^n.$$

Taking $c = \alpha$, $z = \text{prox}_{\beta h}(x + \beta d)$ and $y = x + \alpha d$ we have:

$$\langle u, u - v \rangle \leq \alpha (\langle d, u - v \rangle + h(\text{prox}_{\beta h}(x + \beta d)) - h(\text{prox}_{\alpha h}(x + \alpha d))). \tag{28}$$

Also, if $c = \beta$, $z = \text{prox}_{\alpha h}(x + \alpha d)$ and $y = x + \beta d$, then we have:

$$\langle v, u - v \rangle \geq \beta (\langle d, u - v \rangle + h(\text{prox}_{\alpha h}(x + \alpha d)) - h(\text{prox}_{\beta h}(x + \beta d))). \tag{29}$$

Summing these two inequalities and taking in account that $\alpha > \beta$ we get:

$$\langle d, u - v \rangle + h(\text{prox}_h(x + \beta d)) - h(\text{prox}_h(x + \alpha d)) > 0.$$

Therefore, replacing this expression into inequality (28) leads to $\langle v, u - v \rangle > 0$. Finally, from (27), (28) and (29) we get the inequality:

$$\frac{\|u\|}{\|v\|} \leq \frac{\alpha \langle d, u - v \rangle}{\beta \langle d, u - v \rangle},$$

and then the statement of Lemma 4 can be easily derived. \square

References

1. Auslender, A.: *Optimisation Methodes Numeriques*. Masson, Paris (1976)
2. Beck, A.: The 2-Coordinate Descent Method for Solving Double-Sided Simplex Constrained Minimization Problems. Technical Report (2012)
3. Bertsekas, D.: *Nonlinear Programming*. Athena Scientific (1999)
4. Bonettini, S.: Inexact block coordinate descent methods with application to nonnegative matrix factorization. *J. Numer. Anal.* **22**, 1431–1452 (2011)
5. Calamai, P.H., More, J.J.: Projected gradient methods for linearly constrained problems. *Math. Program.* **39**, 93–116 (1987)
6. Chapelle, O., Sindhvani, V., Keerthi, S.: Optimization techniques for semi-supervised support vector machines. *J. Mach. Learn. Res.* **2**, 203–233 (2008)
7. Fainshil, L., Margalot, M.: A maximum principle for positive bilinear control systems with applications to positive linear switched systems. *SIAM J. Control Optim.* **50**, 2193–2215 (2012)
8. Judice, J., Raydan, M., Rosa, S.S., Santos, S.A.: On the solution of the symmetric eigenvalue complementarity problem by the spectral projected gradient algorithm. *Comput. Optim. Appl.* **47**, 391–407 (2008)
9. Kocvara, M., Outrata, J.: Effective reformulations of the truss topology design problem. *Optim. Eng.* (2006)
10. Lin, C.J., Lucidi, S., Palagi, L., Risi, A., Sciarbone, M.: Decomposition algorithm model for singly linearly-constrained problems subject to lower and upper bounds. *J. Optim. Theory Appl.* **141**(1), 107–126 (2009)
11. Lu, Z., Xiao, L.: Randomized Block Coordinate Non-monotone Gradient Method for a Class of Nonlinear Programming. Technical Report (2013)
12. Mongeau, M., Torchi, M.: Computing eigenelements of real symmetric matrices via optimization. *Comput. Optim. Appl.* **29**, 263–287 (2004)
13. Necoara, I., Nesterov, Y., Glineur, F.: A Random Coordinate Descent Method for Large Optimization Problems with Linear Constraints. Technical Report (2011). <http://acse.pub.ro/person/ion-necoara/>
14. Necoara, I., Patrascu, A.: A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. *Comput. Optim. Appl.* (2013)
15. Necoara, I.: Random coordinate descent algorithms for multi-agent convex optimization over networks. *IEEE Trans. Autom. Control* **58**(8), 1–12 (2013)
16. Necoara, I., Clipici, D.: Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: application to distributed MPC. *J. Process Control* **23**(3), 243–253 (2013)
17. Nesterov, Y.: *Introductory Lectures on Convex Optimization*. Kluwer, Dordrecht (2004)
18. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.* **22**(2), 341–362 (2012)
19. Nesterov, Y.: Gradient methods for minimizing composite objective function. *Math. Program.* **140**(1), 125–161 (2013)
20. Parlett, B.N.: *The Symmetric Eigenvalue Problem*. SIAM (1997)
21. Poliak, B.T.: *Introduction to Optimization*. Optimization Software (1987)
22. Powell, M.J.D.: On search directions for minimization algorithms. *Mathematical Programming* (1973)
23. Richtarik, P., Takac, M.: *Efficient Serial and Parallel Coordinate Descent Methods for Huge-Scale Truss Topology Design*. Operations Research Proceedings, Springer, pp. 27–32 (2012)
24. Richtarik, P., Takac, M.: Iteration complexity of randomized block coordinate descent methods for minimizing a composite function. *Mathematical Programming* (2012)

25. Richtarik, P., Takac, M.: Parallel Coordinate Descent Methods for Big Data Optimization, Technical Report, (2012). <http://www.maths.ed.ac.uk/~richtarik/>
26. Rockafeller, R.T.: The elementary vectors of a subspace in \mathbb{R}^N . In: Bose, R.C., Downing, T.A. (eds.) Combinatorial Mathematics and its Applications, Proceedings of the Chapel Hill Conference, pp. 104–127 (1969)
27. Rockafeller, R.T.: Network Flows and Monotropic Optimization. Wiley-Interscience, New York (1984)
28. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.* **14**, 567–599 (2013)
29. Thi, H.A.L., Moeini, M., Dihn, T.P., Judice, J.: A DC programming approach for solving the symmetric eigenvalue complementarity problem. *Comput. Optim. Appl.* **51**, 1097–1117 (2012)
30. Tseng, P., Yun, S.: A coordinate gradient descent for nonsmooth separable minimization. *Math. Program.* **117**, 387–423 (2009)
31. Tseng, P., Yun, S.: A block coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *J. Optim. Theory Appl.* **140**, 513–535 (2009)
32. Tseng, P.: Approximation accuracy, gradient methods and error bound for structured convex optimization. *Math. Program.* **125**(2), 263–295 (2010)
33. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)