WHEN CYCLIC COORDINATE DESCENT OUTPERFORMS RANDOMIZED COORDINATE DESCENT

M. GÜRBÜZBALABAN*, A. OZDAGLAR†, P. PARRILO†, AND N. D. VANLI†

Abstract. Coordinate descent (CD) method is a classical optimization algorithm that has seen a revival of interest because of its competitive performance in machine learning applications. A number of recent papers provided convergence rate estimates for their deterministic (cyclic) and randomized variants that differ in the selection of update coordinates. These estimates suggest randomized coordinate descent (RCD) performs better than cyclic coordinate descent (CCD), although numerical experiments do not provide clear justification for this comparison. In this paper, we provide examples and more generally problem classes for which CCD (or CD with any deterministic order) is faster than RCD in terms of asymptotic worst-case convergence. Furthermore, we provide lower and upper bounds on the amount of improvement on the rate of deterministic CD relative to RCD. The amount of improvement depend on the deterministic order used. We also provide a characterization of the best deterministic order (that leads to the maximum improvement in convergence rate) in terms of the combinatorial properties of the Hessian matrix of the objective function.

1. Introduction. We consider solving smooth convex optimization problems using coordinate descent (CD) methods. The CD method is an iterative algorithm that performs (approximate) global minimizations with respect to a single coordinate (or several coordinates in the case of block CD) in a sequential manner. More specifically, at each iteration k, an index $i_k \in \{1, 2, \ldots, n\}$ is selected and the decision vector is updated to approximately minimize the objective function in the i_k-th coordinate [3, 4]. CD methods can be deterministic or randomized depending on the choice of the update coordinates. If the coordinate indices i_k are chosen in a cyclic manner over the set $\{1, 2, \ldots, n\}$, then the method is called the *cyclic coordinate descent* (CCD) method. When i_k is sampled uniformly from the set $\{1, 2, \ldots, n\}$, the resulting method is called the *randomized coordinate descent* (RCD) method.

CD methods have a long history in optimization and have been used in many applications. The convergence of these methods has been studied extensively in 80s and 90s (cf. [5, 12, 13, 17]). The CD methods have seen a resurgence of recent interest because of their applicability and superior empirical performance in machine learning and large-scale data analysis [8, 9]. Several recent influential papers established non-asymptotic convergence rate estimates under various assumptions. Among these are Nesterov [15], which provided the first global non-asymptotic convergence rates of RCD for convex and smooth problems (see also [20–22] for problems with non-smooth terms), and Beck and Tetruashvili [1], which provided rate estimates for block coordinate gradient descent method, which yields rate results

*Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ 08854, USA. email: mertg@mit.edu

†Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. email: {asuman, parrilo, denizcan}@mit.edu.

for CCD with exact minimization for quadratic problems. Tighter rate estimates with respect to [1] are then presented in [23,24] for CCD. These rate estimates suggest that CCD can be slower than RCD (precisely $\mathcal{O}(n^2)$ times slower, where n is the dimension of the problem), which is puzzling in view of the faster empirical performance of CCD over RCD for various problems. This gap was investigated by Sun and Ye [24], which provided a quadratic problem that attains this performance gap. In this paper, we investigate performance comparison of deterministic and randomized coordinate descent and provide examples and more generally problem classes for which *CCD (or CD with any deterministic order) is faster than RCD* in terms of asymptotic worst-case convergence. Furthermore, we provide lower and upper bounds on the amount of improvement on the rate of deterministic CD relative to RCD. The amount of improvement depends on the deterministic order used. We also provide a characterization of the best deterministic order (that leads to the maximum improvement in convergence rate) in terms of the combinatorial properties of the Hessian matrix of the objective function.

In order to clarify the rate comparison between CCD and RCD, we focus on quadratic optimization problems. In particular, we consider the problem[1]

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^T Ax, \tag{1.1}$$

where A is a positive definite matrix. We consider two problem classes: *i)* A is an M-matrix, i.e., the off-diagonal entries of A are nonpositive. These matrices arise in a large number of applications. A notable example is problems that consider minimization of quadratic forms of graph Laplacians (where $A = D - W$ and W denoes the weighted adjacency graph and D is a diagonal matrix given by $D_{i,i} = \sum_j W_{i,j}$), e.g., for spectral partitioning [6] and semisupervised learning [2]. *ii)* A is a 2-cyclic matrix, whose formal definition is given in Definition 4.2, but an equivalent and insightful definition is the bipartiteness of the underlying induced graph. We build on the seminal work of Young [28] and Varga [26] on the analysis of Gauss-Seidel method for solving linear systems of equations (with matrices satisfying certain properties) and provide a novel analysis that allows us to compare the asymptotic worst-case convergence rate of CCD and RCD for the aforementioned class of problems and establish the faster performance of CCD with any deterministic order.

Outline: In the next section, we formally introduce the CCD and RCD methods. In Section 3, we present the notion of asymptotic convergence rate to compare the CCD and RCD methods and provide a motivating example on which CCD converges faster than RCD. In Section 4, we present classes of problems for which the asymptotic convergence rate of CCD is faster than the one of RCD. We conclude in Section 5 by providing numerical experiments that validates our theoretical results on the performance of the CCD and RCD methods.

[1] For ease of presentation, we consider minimization of $\frac{1}{2}x^T Ax$, yet our results directly extend for problems of the type $\frac{1}{2}x^T Ax - b^T x$ for any $b \neq 0$.

Notation: For a matrix H, we let H_i denote its ith row and $H_{i,j}$ denote its entry at the ith row and jth column. For a vector x, we let x_i denote its ith entry. Throughout the paper, we reserve superscripts for iteration counters of an iterative algorithm and use x^* to denote the optimal solution of problem (1.1). For a vector x, $\|x\|$ denotes its Euclidean norm and for a matrix H, $||H||$ denotes its operator norm. For a matrices, $\geq$ and $\leq$ are entry-wise operators. The matrices I and 0 denote the identity matrix and the zero matrix respectively and their dimensions can be understood from the context.

2. Coordinate Descent Methods. Starting from an initial point $x^0 \in \mathbb{R}^n$, the coordinate descent (CD) method, at each iteration k, picks a coordinate of x, say i_k, and updates the decision vector by performing exact minimization along the i_kth coordinate, which for problem (1.1) yields

$$x^{k+1} = x^k - \frac{1}{A_{i_k,i_k}} A_{i_k} x^k e_{i_k}, \quad k = 0, 1, 2, \ldots, \tag{2.1}$$

where e_{i_k} is the unit vector, whose i_kth entry is 1 and the rest of its entries are 0. Note that this is a special case of the coordinate gradient projection method (see [1]), which at each iteration updates a single coordinate, say coordinate i_k, along the gradient component direction (with the particular step size of $\frac{1}{A_{i_k,i_k}}$). The coordinate index i_k can be selected according to a deterministic or randomized rule:

- When i_k is chosen using the *cyclic rule* with order $1, \ldots, n$ (i.e., $i_k = k \pmod n + 1$), the resulting algorithm is called the cyclic coordinate descent (CCD) method. In order to write the CCD iterations in a matrix form, we introduce the following decomposition

  $$A = D - L - L^T,$$

  where D is the diagonal part of A and $-L$ is the strictly lower triangular part of A. Then, over each epoch $\ell \geq 0$ (where an epoch is defined to be consecutive n iterations), the CCD iterations given in (2.1) can be written as

  $$x_{\text{CCD}}^{(\ell+1)n} = C\, x_{\text{CCD}}^{\ell n}, \quad \text{where} \quad C = (D - L)^{-1} L^T. \tag{2.2}$$

  Note that the epoch in (2.2) is equivalent to one iteration of the Gauss-Seidel (GS) method applied to the first-order optimality condition of (1.1), i.e., applied to the linear system $Ax = 0$ [27].
- When i_k is chosen at random among $\{1, \ldots, n\}$ with probabilities $\{p_1, \ldots, p_n\}$ independently at each iteration k, the resulting algorithm is called the randomized coordinate descent (RCD) method. Given the kth iterate generated by the RCD algorithm, i.e., x_{RCD}^k, we have

  $$\mathbb{E}_k \left[x_{\text{RCD}}^{k+1} \mid x_{\text{RCD}}^k \right] = \left(I - PD^{-1}A \right) x_{\text{RCD}}^k, \tag{2.3}$$

where $P = \mathrm{diag}(p_1, \ldots, p_n)$ and the conditional expectation $\mathbb{E}_k$ is taken over the random variable i_k given x^k_{RCD}. Using the nested property of the expectations, the RCD iterations in expectation over each epoch $\ell \geq 0$ satisfy

$$\mathbb{E}x^{(\ell+1)n}_{\mathrm{RCD}} = R\, \mathbb{E}x^{\ell n}_{\mathrm{RCD}} \quad \text{with} \quad R := \left(I - PD^{-1}A\right)^n. \tag{2.4}$$

3. Comparison of the Convergence Rates of CCD and RCD Methods. In the following subsection, we define our basis of comparison for rates of CCD and RCD methods. To measure the performance of these methods, we use the notion of the average worst-case asymptotic rate that has been studied extensively in the literature for characterizing the rate of iterative algorithms [26]. In Section 3.2, we construct an example, for which the rate of CCD is more than twice the rate of RCD. This raises the question whether the best known convergence rates of CCD in the literature are tight or whether there exist a class of problems for which CCD provably attains better convergence rates than the best known rates for RCD, a question which we will answer in Section 4 .

3.1. Asymptotic Converge Rate for Iterative Algorithms. Consider an iterative algorithm with update rule $x^{(\ell+1)n} = Cx^{\ell n}$ (e.g., the CCD algorithm). The reduction in the distance to the optimal solution of the iterates generated by this algorithm after ℓ epochs is given by

$$\frac{\left|\left|x^{\ell n} - x^*\right|\right|}{||x^0 - x^*||} = \frac{\left|\left|C^\ell(x^0 - x^*)\right|\right|}{||x^0 - x^*||}. \tag{3.1}$$

Note that the right hand side of (3.1) can be as large as $\left|\left|C^\ell\right|\right|$, hence in the worst-case, the average decay of distance at each epoch of this algorithm is $\left|\left|C^\ell\right|\right|^{1/\ell}$. Over any finite epochs $\ell \geq 1$, we have $\left|\left|C^\ell\right|\right|^{1/\ell} \geq \rho(C)$ and $\left|\left|C^\ell\right|\right|^{1/\ell} \to \rho(C)$ as $\ell \to \infty$. Thus, we define the *asymptotic worst-case convergence rate* of an iterative algorithm (with iteration matrix C) as follows

$$\mathcal{R}(C) := \lim_{\ell\to\infty} \sup_{x^0\in\mathbb{R}^n} -\frac{1}{\ell} \log\left(\frac{\left|\left|x^{\ell n} - x^*\right|\right|}{||x^0 - x^*||}\right) = -\log\left(\rho(C)\right). \tag{3.2}$$

We emphasize that this notion has been used extensively for studying the performance of iterative methods such as GS and Jacobi methods [5, 17, 26, 28]. Note that according to our definition in (3.2), larger rate means faster algorithm and we will use these terms interchangably in throughout the paper.

Analogously, for a randomized algorithm with expected update rule $\mathbb{E}x^{(\ell+1)n} = R\,\mathbb{E}x^{\ell n}$ (e.g., the RCD algorithm), we consider the asymptotic convergence of the expected iterate error $\left|\left|\mathbb{E}(x^{\ell n}) - x^*\right|\right|$ and define the asymptotic worst-case convergence rate as

$$\overline{\mathcal{R}}(R) := \lim_{\ell\to\infty} \sup_{x^0\in\mathbb{R}^n} -\frac{1}{\ell} \log\left(\frac{\left|\left|\mathbb{E}(x^{\ell n}) - x^*\right|\right|}{||x^0 - x^*||}\right) = -\log\left(\rho(R)\right), \tag{3.3}$$

Note that in (3.3), we use the distance of the expected iterates $\left|\left|\mathbb{E}x^{\ell n} - x^*\right|\right|$ as our convergence criterion. One can also use the expected distance (or the squared distance) of the iterates $\mathbb{E}\left|\left|x^{\ell n} - x^*\right|\right|$

as the convergence criterion, which is a stronger convergence criterion than the one in (3.3). This follows since $\mathbb{E}\left|\left|x^{\ell n}-x^*\right|\right| \geq \left|\left|\mathbb{E}x^{\ell n}-x^*\right|\right|$ by Jensen's inequality and any convergence rate on $\mathbb{E}\left|\left|x^{\ell n}-x^*\right|\right|$ immediately implies at least the same convergence rate on $\left|\left|\mathbb{E}x^{\ell n}-x^*\right|\right|$ as well. Since we consider the reciprocal case, i.e., obtain a convergence rate on $\left|\left|\mathbb{E}x^{\ell n}-x^*\right|\right|$ and show that it is slower than that of CCD, our results naturally imply that the convergence rate on $\mathbb{E}\left|\left|x^{\ell n}-x^*\right|\right|$ is also slower than that of CCD.

3.2. A Motivating Example. In this section, we provide an example for which the (asymptotic worst-case convergence) rate of CCD is better than the one of RCD and building on this example, in Section 4, we construct a class of problems for which CCD attains a better rate than RCD. For some positive integer $n \geq 1$, consider the $2n \times 2n$ symmetric matrix

$$A = I - L - L^T, \quad \text{where} \quad L = \frac{1}{n^2}\begin{bmatrix} 0_{n\times n} & 0_{n\times n} \\ \mathbb{1}_{n\times n} & 0_{n\times n} \end{bmatrix}, \tag{3.4}$$

and $\mathbb{1}_{n\times n}$ is the $n \times n$ matrix with all entries equal to 1 and $0_{n\times n}$ is the $n \times n$ zero matrix. Noting that A has a special structure (A is equal to the sum of the identity matrix and the rank-two matrix $-L - L^T$), it is easy to check that $1 - 1/n$ and $1 + 1/n$ are eigenvalues of A with the corresponding eigenvectors $\begin{bmatrix}\mathbb{1}_{1\times n} & \mathbb{1}_{1\times n}\end{bmatrix}^T$ and $\begin{bmatrix}\mathbb{1}_{1\times n} & -\mathbb{1}_{1\times n}\end{bmatrix}^T$. The remaining $2n-2$ eigenvalues of A are equal to 1.

The iteration matrix of the CCD algorithm when applied to the problem in (1.1) with the matrix (3.4) can be found as

$$C = \begin{bmatrix} 0_{n\times n} & \frac{1}{n^2}\mathbb{1}_{n\times n} \\ 0_{n\times n} & \frac{1}{n^3}\mathbb{1}_{n\times n} \end{bmatrix}.$$

The eigenvalues of C are all zero except the eigenvalue of $1/n^2$ with the corresponding eigenvector $[n\mathbb{1}_{1\times n}, \mathbb{1}_{1\times n}]^T$. Therefore, $\rho(C) = 1/n^2$ and $\mathcal{R}(C) = -\log(\rho(C)) = 2\log n$. On the other hand, the spectral radius of the expected iteration matrix of RCD can be found as

$$\rho(R) = \left(1 - \frac{\lambda_{\min}(A)}{n}\right)^n \geq 1 - \lambda_{\min}(A) = \frac{1}{n},$$

which yields $\overline{\mathcal{R}}(R) = -\log(\rho(R)) \leq \log n$. Thus, we conclude

$$\frac{\mathcal{R}(C)}{\overline{\mathcal{R}}(R)} \geq 2, \quad \text{for all} \quad n \geq 1.$$

That is, CCD is at least twice as fast as RCD in terms of the the asymptotic rate. This motivates us to investigate if there exists a more general class of problems for which the asymptotic worst-case rate of CCD is larger than that of RCD. The answer to this question turns out to be positive as we describe in the following section.

4. When Deterministic Orders Outperform Randomized Sampling. In this section, we present special classes of problems (of the form (1.1)) for which the asymptotic worst-case rate of CCD is larger than that of RCD. We begin our discussion by highlighting the main assumption we will use in this section.

ASSUMPTION 1. *Hessian matrix A has the following properties:*

(i) A is a symmetric positive definite matrix with smallest eigenvalue $\mu > 0$.

(ii) The diagonal entries of A are 1.

Given any positive definite matrix A with diagonals $D \neq I$, the diagonal entries of the preconditioned matrix $D^{-1/2}AD^{-1/2}$ are 1. Therefore, part *(ii)* of Assumption 1 is mild. The relationship between the smallest eigenvalue of the original matrix and the preconditioned matrix are as follows. Let $\sigma > 0$ and $L_{\max}$ denote the smallest eigenvalue and the largest diagonal entry of the original matrix, respectively. Then, the smallest eigenvalue of the preconditioned matrix satisfies $\mu \geq \sigma/L_{\max}$.

REMARK 1. *For the RCD algorithm, the coordinate index $i_k \in \{1, \ldots, n\}$ (at iteration k) can be chosen using different probability distributions $\{p_1, \ldots, p_n\}$. The most widely used distributions (due to their simplicity) have the form $p_i = \frac{A_{i,i}^{\alpha}}{\sum_{J=1}^{N} A_{j,j}^{\alpha}}$ for a choice of $\alpha \geq 0$ as discussed in [15]. Since by Assumption 1, the diagonal entries of A are* 1*, we have $p_i = \frac{1}{n}$ for all $i \in \{1, \ldots, n\}$ and $\alpha \geq 0$. Therefore, in the rest of the paper, we consider the RCD algorithm with uniform and independent coordinate selection at each iteration.*

In the following lemma, we characterize the spectral radius of the RCD method.

LEMMA 4.1. *Suppose Assumption 1 holds. Then, the spectral radius of the expected iteration matrix R of the RCD algorithm (defined in* (2.4)*) is given by*

$$\rho(R) = \left(1 - \frac{\mu}{n}\right)^n. \tag{4.1}$$

Proof. By Assumption 1, $\mu > 0$ and $\operatorname{tr}(A) = n$, which implies all eigenvalues of the matrix A/n are in the interval $(0, 1)$. Therefore, we have

$$\rho(R) = \lambda_{\max}\left(\left(I - \frac{1}{n}A\right)^n\right) = \left(1 - \frac{1}{n}\lambda_{\min}(A)\right)^n = \left(1 - \frac{\mu}{n}\right)^n.$$

□

In the following sections, we present classes of problems for which CCD attains better convergence rates than RCD.

4.1. Convergence Rate of CCD for 2-Cyclic Matrices. In this section, we introduce the class of 2-cyclic matrices and show that the asymptotic worst-case convergence rate of CCD is more than two times faster than that of RCD.

4.1.1. Definition & Properties.

DEFINITION 4.2 (2-Cyclic Matrix). *A matrix H is 2-cyclic if there exists a permutation matrix P such that*

$$PHP^T = D + \begin{bmatrix} 0 & B_1 \\ B_2 & 0 \end{bmatrix}, \tag{4.2}$$

where the diagonal null submatrices are square and D is a diagonal matrix.

This definition can be interpreted as follows. Let H be a 2-cyclic matrix, i.e., H satisfies (4.2). Then, the graph induced by the matrix $H - D$ is bipartite. The definition in (4.2) is first introduced in [28], where it had an alternative name, called *Property A*. A generalization of this property is later introduced by Varga to the class of p-cyclic matrices [26] where $p \geq 2$ can be arbitrary.

We next introduce the following definition that will be useful in Theorem 4.10 and explicitly identify the class of matrices that satisfy this definition.

DEFINITION 4.3 (Consistently Ordered Matrix). *For a matrix H, let $H = H_D - H_L - H_U$ be its decomposition such that H_D is a diagonal matrix, H_L (and H_U) is a strictly lower (and upper) triangular matrix. If the eigenvalues of the matrix $\alpha H_L + \alpha H_U - \gamma H_D$ are independent of α for any $\gamma \in \mathbb{R}$ and $\alpha \neq 0$, then H is said to be consistently ordered.*

In the next lemma, we highlight the connection between Definitions 4.2 and 4.3.

LEMMA 4.4. *[28, Theorem 4.5] A matrix H is 2-cyclic if and only if there exists a permutation matrix P such that PHP^T is consistently ordered.*

This lemma shows that in order for the lower bounds in Theorem 4.10 to hold with equality, it is necessary and sufficient that the lower triangular part of A can be written as $L = \begin{bmatrix} 0 & 0 \\ B & 0 \end{bmatrix}$, for a real matrix B where the diagonal null submatrices are square matrices of appropriate dimension. However, in Theorem 4.10, we assume that A is an M-matrix, i.e., $L \geq 0$. In the following theorem, we prove that a similar spectral radius equality to Theorem 4.10 holds for consistently ordered 2-cyclic matrices under less restrictive assumptions (by removing the assumption that the off-diagonal entries are non-positive).

4.1.2. Convergence Rates. In the next theorem, we characterize the convergence rate of CCD algorithm applied to a 2-cyclic matrix. Since $\rho(R) \geq 1 - \mu$ by Lemma 4.1, the following theorem indicates that the spectral radius of the CCD iteration matrix is smaller than $\rho^2(R)$.

THEOREM 4.5. *Suppose Assumption 1 holds and A is a consistently ordered 2-cyclic matrix. Then, the spectral radius of the CCD algorithm is given by*

$$\rho(C) = (1 - \mu)^2 .$$

Proof. The eigenvalues of C are the roots of the polynomial

$$\phi_C(\lambda) = \det(\lambda I - C) = 0.$$

As $I - L$ is nonsingular and $\det(I - L) = 1$, we have

$$\begin{aligned}\phi_C(\lambda) &= \det(I - L)\det(\lambda I - C)\\ &= \det(\lambda I - \lambda L - L^T)\\ &= \sqrt{\lambda}\det\left(\sqrt{\lambda} I - \left(\sqrt{\lambda} L + \frac{1}{\sqrt{\lambda}} L^T\right)\right).\end{aligned}$$

Therefore, if $\sqrt{\lambda}$ is an eigenvalue of the matrix $\sqrt{\lambda}L + \frac{1}{\sqrt{\lambda}}L^T$, then λ is an eigenvalue of C. Furthermore, since the eigenvalues of the matrix $\sqrt{\lambda}L + \frac{1}{\sqrt{\lambda}}L^T$ are independent of λ as A is a consistently ordered matrix by definition, then $\sqrt{\lambda}$ is an eigenvalue of $L + L^T$ as well. Consequently, we have $\rho(C) = \rho^2(L + L^T) = \rho^2(I - A) = (1 - \mu)^2$. ∎

REMARK 2. *Note that our motivating example given by* (3.4) *in Section 3.2 is an example of a consistently ordered 2-cyclic matrix where Theorem 4.5 is directly applicable. In fact, for* (3.4)*, we can apply Theorem 4.5 with* $\mu = 1 - 1/n$ *leading to* $\rho(C) = 1/n^2$*, which coincides exactly with our previous computations of* $\rho(C)$ *in Section 3.2. We also give an example in the Appendix B where CCD is twice faster from any arbitrary initialization with probability one.*

The following corollary states that the asymptotic worst-case convergence rate of CCD is more than twice larger than that of RCD for quadratic problems whose Hessian is a 2-cyclic matrix. This corollary directly follows by Theorem 4.5 and definitions (3.2)-(3.3).

COROLLARY 4.6. *Suppose Assumption 1 holds and A is a consistently ordered 2-cyclic matrix. Then, for the constant* $\nu_n > 1$ *as defined in* (4.10)*, the asymptotic worst-case rate of CCD and RCD satisfies*

$$\frac{\mathcal{R}(C)}{\overline{\mathcal{R}}(R)} = 2\nu_n, \quad \text{where} \quad \nu_n := \frac{\log(1-\mu)}{n\log\left(1 - \frac{\mu}{n}\right)}. \tag{4.3}$$

In the following remark, we highlight several properties of the constant ν_n.

REMARK 3. ν_n *is a monotonically increasing function of* n *over the interval* $[1, \infty)$*, where* $\nu_1 = 1$ *and* $\lim_{n\to\infty} \nu_n = \frac{-\log(1-\mu)}{\mu} > 1$*. Furthermore,* $\lim_{\mu\to 0^+} \nu_n = 1$.

4.2. Convergence Rate of CCD for Irreducible M-Matrices. In this section, we first define the class of M-matrices and then present the convergence rate of the CCD algorithm applied to quadratic problems whose Hessian is an M-matrix.

4.2.1. Definition & Properties.

DEFINITION 4.7 (M-matrix). *A real matrix A with $A_{i,j} \le 0$ for all $i \neq j$ is an M-matrix if A has the decomposition $A = sI - B$ such that $B \ge 0$ and $s \ge \rho(B)$.*

We emphasize that M-matrices arise in a variety of applications such as belief propagation over Gaussian graphical models [14] and distributed control of positive systems [19], and has been used to analyze performance of various algorithms in the literature [5, 22, 25]. Furthermore, graph Laplacians are M-matrices, therefore solving linear systems with M-matrices (or equivalently solving (1.1) for an M-matrix A) arise in a variety of applications for analyzing random walks over graphs and distributed optimization and consensus problems over graphs (cf. [11] for a survey). For quadratic problems, the Hessian is an M-matrix if and only if the gradient descent mapping is an isotone operator [5, 22] and in Gaussian graphical models, M-matrices are often referred as attractive models [14].

In the following lemma, we highlight a property of nonsingular M-matrices, which we will use in the following section to characterize the convergence rate of the CCD algorithm applied to quadratic problems whose Hessian is an M-matrix.

LEMMA 4.8. *[18, Theorem 2] A is a nonsingular M-matrix if and only if A^{-1} exists and $A^{-1} \geq 0$.*

Before concluding this section, we introduce the following lemma, which is presented in variuos papers (e.g., [26, Lemma 4.12], [16, Corollary 1.2], [10, Theorem 1]) to analyze the spectral radii of nonnegative matrices. Particularly, this lemma states that if the matrix $e^{\alpha}L+e^{-\alpha}L^T$ is not consistently ordered (where $L \geq 0$ is a strictly lower triangular matrix), then its spectral radius is strictly log-convex in α. The proof of this lemma is presented in Appendix A for completeness.

LEMMA 4.9. *Let $B_\alpha = e^{\alpha}L+e^{-\alpha}L^T$, where $L \geq 0$ is a strictly lower triangular matrix and $\alpha \in \mathbb{R}$. Then, either $\rho(B_\alpha)$ is strictly log-convex in α with $\rho(B_\alpha) > \rho(B_0)$ for all $\alpha \neq 0$ or $\rho(B_\alpha)$ is constant for all $\alpha \in \mathbb{R}$ (i.e., B_α is a consistently ordered matrix).*

4.2.2. Convergence Rates. In the following theorem, we provide lower and upper bounds on the spectral radius of the iteration matrix of CCD for quadratic problems whose Hessian matrix is an irreducible M-matrix. In particular, we show that the spectral radius of the iteration matrix of CCD is strictly smaller than the one of RCD for irreducible M-matrices. Note that the Hessian matrix in our motivating example (in Section 3.2) is an irreducible M-matrix.

THEOREM 4.10. *Suppose Assumption 1 holds, A is an irreducible M-matrix and $n \geq 2$. Then, the iteration matrix of the CCD algorithm $C = (I - L)^{-1}L^T$ satisfies the following inequality*

$$(1 - \mu)^2 \leq \rho(C) \leq \frac{1 - \mu}{1 + \mu}, \tag{4.4}$$

where the inequality on the left holds with equality if and only if A is a consistently ordered matrix.

Proof. Since A is an M-matrix, $I - L$ is an M-matrix as well. Consequently, $(I - L)^{-1} \geq 0$, which implies $C = (I - L)^{-1}L^T \geq 0$ by Lemma 4.8. Then, by Perron-Frobenius Theorem, there exists a real eigenvalue of C, denoted by λ, and the corresponding unit-norm eigenvector $z \geq 0$ satisfying $\lambda = \rho(C) \geq 0$ and

$$Cz = \lambda z.$$

Multiplying both sides of the above equality by $I - L$ from the left, we obtain

$$L^T z = \lambda(I - L)z,$$

and rearranging terms yields

$$(\lambda L + L^T)z = \lambda z. \tag{4.5}$$

Therefore, λ is an eigenvalue of the matrix $\lambda L + L^T$. We then observe that $\lambda L + L^T$ is an irreducible matrix as A is irreducible as the indices of the nonzero entries of both matrices are the same. Since $\lambda L + L^T$ is nonnegative and irreducible and z is nonnegative, then by Perron-Frobenius Theorem, z is the eigenvector corresponding to the spectral radius of $\lambda L + L^T$. Therefore,

$$\lambda = \rho(\lambda L + L^T) = \sqrt{\lambda}\,\rho\left(\sqrt{\lambda}L + \frac{1}{\sqrt{\lambda}}L^T\right). \tag{4.6}$$

In order to obtain a lower bound on the right-hand side of (4.6), we use Lemma 4.9 (note that $\lambda < 1$ by Definition 4.7) and conclude that

$$\lambda \geq \sqrt{\lambda}\,\rho\left(L + L^T\right), \tag{4.7}$$

with equality if and only if A is a consistently ordered matrix. Since $\lambda = \rho(C)$, (4.7) yields

$$\rho(C) \geq \rho^2\left(L + L^T\right) = \rho^2\left(I - A\right) = (1 - \mu)^2,$$

with equality if and only if A is a consistently ordered matrix, which concludes the proof of the lower bound in (4.4). In order to obtain an upper bound on $\rho(C)$, we turn our attention back to (4.5) and multiply both sides by z^T from the left. This yields

$$\lambda z^T L z + z^T L^T z = \lambda,$$

since $||z|| = 1$. Noting that $z^T L z = z^T L^T z$ and defining $\beta = z^T L z$, we obtain

$$\lambda = \frac{\beta}{1 - \beta}. \tag{4.8}$$

Since $\rho(L + L^T) = \rho(I - A) = 1 - \mu$, then for any $||y|| = 1$, we have $y^T(L + L^T)y \leq 1 - \mu$. Picking $y = z$ in this inequality yields $2\beta \leq 1 - \mu$ and combining this with (4.8) and noting $\lambda = \rho(C)$ imply the upper bound in (4.4). □

An immediate consequence of Theorem 4.10 is that for quadratic problems whose Hessian is an irreducible M-matrix, the best cyclic order that should be used in CCD can be characterized as follows.

REMARK 4. *The standard CCD method follows the standard cyclic order* $(1, 2, \ldots, n)$ *as described in Section 2. However, we can construct a CCD method that follows an alternative deterministic order*

by considering a permutation π of $\{1, 2, \ldots, n\}$, and choosing the coordinates according to the order $(\pi(1), \pi(2), \ldots, \pi(n))$ instead. For any given order π, (1.1) can be reformulated as follows

$$\min_{x_\pi \in \mathbb{R}^n} \frac{1}{2} x_\pi^T A_\pi x_\pi, \quad \text{where} \quad A_\pi := P_\pi A P_\pi^T \quad \text{and} \quad x_\pi = P_\pi\, x,$$

where P_π is the corresponding permutation matrix of π. Supposing that Assumption 1 holds, the corresponding CCD iterations for this problem can be written as follows

$$x_\pi^{(\ell+1)n} = C_\pi\, x_\pi^{\ell n}, \quad \text{where} \quad C_\pi = (I - L_\pi)^{-1} L_\pi^T \quad \text{and} \quad L_\pi = P_\pi L P_\pi.$$

If A is an irreducible M-matrix and satisfies Assumptions 1, then so does A_π. Consequently, Theorem 4.10 yields the same upper and lower bounds (in (4.4)) on $\rho(C_\pi)$ as well, i.e., the spectral radius of the iteration matrix of CCD with any cyclic order π satisfies

$$(1 - \mu)^2 \leq \rho(C_\pi) \leq \frac{1 - \mu}{1 + \mu}, \tag{4.9}$$

where the inequality on the left holds with equality if and only if A_π is a consistently ordered matrix. Therefore, if a consistent order π^* exists, then the CCD method with the consistent order π^* attains the smallest spectral radius (or equivalently, the fastest asymptotic worst-case convergence rate) among the CCD methods with any cyclic order.

REMARK 5. *The irreducibility of A is essential to derive the lower bound in (4.4) of Theorem 4.10. However, the upper bound in (4.4) holds even when A is a reducible matrix.*

We next compare the spectral radii bounds for CCD (given in Theorem 4.10) and RCD (given in Lemma 4.1). Since $\mu > 0$, the right-hand side of (4.4) can be relaxed to $(1 - \mu)^2 \leq \rho(C) < 1 - \mu$. A direct consequence of this inequality is the following corollary, which states that the asymptotic worst-case rate of CCD is strictly better than that of RCD at least by a factor that is strictly greater than 1.

COROLLARY 4.11. *Suppose Assumption 1 holds, A is an irreducible M-matrix and $n \geq 2$. Then, the asymptotic worst-case rate of CCD and RCD satisfies*

$$1 < \nu_n < \frac{\mathcal{R}(C)}{\overline{\mathcal{R}}(R)} \leq 2\nu_n, \quad \text{where} \quad \nu_n := \frac{\log(1 - \mu)}{n \log\left(1 - \frac{\mu}{n}\right)}, \tag{4.10}$$

and the inequality on the right holds with equality if and only if A is a consistently ordered matrix.

In the following corollary, we highlight that as the smallest eigenvalue of A goes to zero, the asymptotic worst-case rate of the CCD algorithm becomes twice the asymptotic worst-case rate of the RCD algorithm.

COROLLARY 4.12. *Suppose Assumption 1 holds, A is an irreducible M-matrix and $n \geq 2$. Then, we have*

$$\lim_{\mu \to 0^+} \frac{\mathcal{R}(C)}{\overline{\mathcal{R}}(R)} = 2.$$

Proof. By Theorem 4.10, we have the following worst-case asymptotic rate bounds for the CCD algorithm

$$-\log(1-\mu)+\log(1+\mu) \leq \mathcal{R}(C) \leq -2\log(1-\mu).$$

Dividing both sides of the above inequality by $-\log(1-\mu)$, we obtain

$$1-\frac{\log(1+\mu)}{\log(1-\mu)} \leq \frac{\mathcal{R}(C)}{-\log(1-\mu)} \leq 2.$$

Taking limit of both sides as $\mu \to 0^+$ yields

$$\lim_{\mu\to 0^+} \frac{\mathcal{R}(C)}{-\log(1-\mu)} = 2. \tag{4.11}$$

By Lemma 4.1, we have the following asymptotic worst-case rate for the RCD algorithm

$$\overline{\mathcal{R}}(R) = -n\log\left(1-\frac{\mu}{n}\right).$$

Dividing both sides of the above inequality by $-\log(1-\mu)$ and taking limit of both sides as $\mu \to 0^+$, we get

$$\lim_{\mu\to 0^+} \frac{\overline{\mathcal{R}}(R)}{-\log(1-\mu)} = 1. \tag{4.12}$$

Combining (4.11) and (4.12) concludes the proof. ∎

4.3. Convergence Rate of CCD for Non-frustrated Matrices. In this section, we define the class of non-frustrated matrices and present the convergence rate of the CCD algorithm applied to quadratic problems whose Hessian is a non-frustrated matrix.

4.3.1. Definition & Properties.

DEFINITION 4.13. *A real matrix A is called a non-frustrated matrix if A has the decomposition $A = I - B$ such that B does not contain any frustrated cycles, i.e., cycles with an odd number of negative edge weights.*

The class of non-frustrated matrices are highly related to the class of M-matrices as we highlight in the following lemma. It states that any non-frustrated matrix is sign-similar to an M-matrix.

LEMMA 4.14. *[7] Let $S(B)$ be the signed digraph of B, i.e., $S(B) = \text{sign}(B)$. If B is irreducible and all cycles of $S(B)$ are positive, then B is sign-similar to a nonnegative matrix, i.e., $B = DED^{-1}$, where $E \geq 0$ and D is a diagonal matrix with entries ± 1.*

4.3.2. Convergence Rates. Using Lemma 4.14 and Theorem 4.10, we show that the same convergence rate guarantees in Theorem 4.10 (for M-matrices) hold for the non-frustrated matrices as

well.

THEOREM 4.15. *Suppose Assumption 1 holds, A is an irreducible non-frustrated matrix and $n \geq 2$. Then, the iteration matrix of the CCD algorithm $C = (I - L)^{-1}L^T$ satisfies the following inequality*

$$(1 - \mu)^2 \leq \rho(C) \leq \frac{1 - \mu}{1 + \mu}, \tag{4.13}$$

where the inequality on the left holds with equality if and only if A is a consistently ordered matrix.

Proof. Since A is assumed to be an irreducible non-frustrated matrix, then by Definition 4.13 and Lemma 4.14, A is sign-similar to $\bar{A}$ (i.e., $A = D\bar{A}D^{-1}$ for some diagonal matrix D whose entries are ± 1.), where $\bar{A}$ is the comparison matrix of A defined as

$$\bar{A}_{i,j} = \begin{cases} A_{i,j} & \text{, if } i = j \\ -|A_{i,j}| & \text{, else.} \end{cases} \tag{4.14}$$

Let $\bar{A} = I - \bar{L} - \bar{L}^T$ be the decomposition of $\bar{A}$ such that $\bar{L}$ is a strictly lower triangular matrix. Then, by Theorem 4.10, we conclude that

$$(1 - \mu)^2 \leq \rho(\bar{C}) \leq \frac{1 - \mu}{1 + \mu},$$

where the inequality on the left holds with equality if and only if $\bar{A}$ is a consistently ordered matrix. To conclude the proof, we claim that C is sign-similar to $\bar{C}$, which follows since

$$\begin{aligned} C &= (I - L)^{-1}L^T \\ &= (D(I - \bar{L})D)^{-1}(D\bar{L}D)^T \\ &= D(I - \bar{L})^{-1}D^2\bar{L}^T D \\ &= D(I - \bar{L})^{-1}\bar{L}^T D \\ &= D\bar{C}D, \end{aligned}$$

where the equalities follow since D is a Householder matrix, i.e., $D = D^{-1}$ and $D^2 = I$. Hence, C is sign-similar to $\bar{C}$ and consequently $\rho(C) = \rho(\bar{C})$, which concludes the proof. ∎

5. Numerical Experiments. In this section, we compare the performance of CCD and RCD through numerical examples. First, we consider the quadratic optimization problem in (1.1), where A is an $n \times n$ matrix defined as follows

$$A = I - L - L^T, \quad \text{where} \quad L = \frac{1}{n}\begin{bmatrix} 0 & 0 \\ \mathbb{1}_{\frac{n}{2}\times\frac{n}{2}} & 0 \end{bmatrix}, \tag{5.1}$$

and $\mathbb{1}_{\frac{n}{2}\times\frac{n}{2}}$ is the $\frac{n}{2} \times \frac{n}{2}$ matrix with all entries equal to 1. Here, it can be easily checked that A is a consistently ordered 2-cyclic matrix. By Theorem 4.5 and Corolloary 4.6, the worst-case convergence

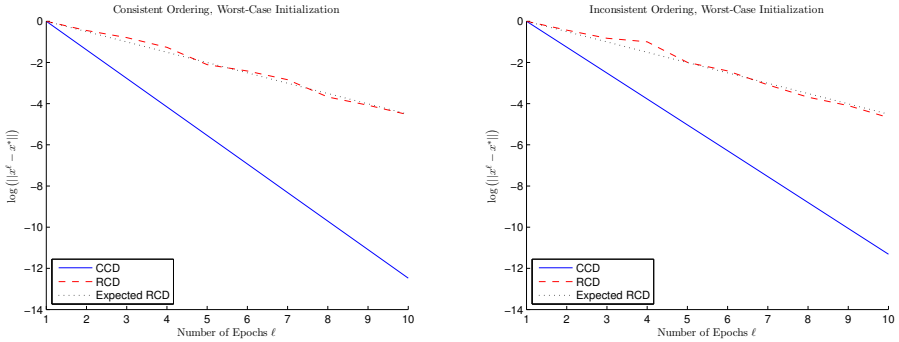

Fig. 4.1: Distance to the optimal solution of the iterates of CCD and RCD for the cyclic matrix in (5.1) (left figure) and a randomly permuted version of the same matrix (right figure) where the y-axis is on a logarithmic scale. The left (right) panel corresponds to the consistent (inconsistent) ordering for the same quadratic optimization problem.

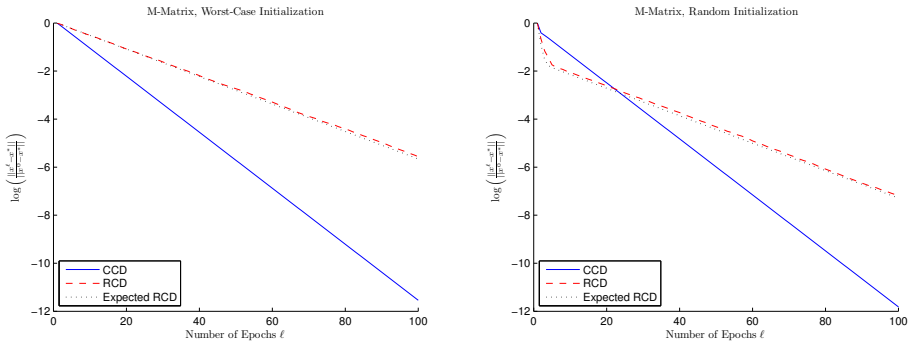

Fig. 4.2: Distance to the optimal solution of the iterates of CCD and RCD for the M-matrix matrix in (5.2) for the worst-case initialization (left figure) and a random initialization (right figure).

rate of CCD on this example is

$$2\nu_m = 2\frac{\log(1-\mu)}{m\log\left(1-\frac{\mu}{m}\right)} = \frac{\log(0.5)}{50\log\left(1-\frac{1}{200}\right)} \approx 2.77$$

times faster than the convergence rate of RCD asymptotically. This is illustrated on the left panel of Figure 4.1, where the distance to the optimal solution is plotted in a logarithmic scale over epochs. Note that even if our results our asymptotic, we see the same difference in performances on the early epochs (for small ℓ). On the other hand, when the matrix A is not consistently ordered, according to Theorem 4.10, CCD is still faster but the difference in the convergence rates decreases with respect to the consistent ordering case. To illustrate this, we need to generate an inconsistent ordering of the matrix A. For this goal, we generate a random permutation matrix P and replace A with $A_P := PAP^T$ in the optimization problem (1.1). The right panel in Figure 4.1 shows that for this inconsistent

ordering CCD is still faster compared to RCD, but not as fast (the slope of the decay of error line in blue marker is less steep) predicted by our theory.

We next consider the case that A is an irreducible positive definite M-matrix. In particular, we consider the matrix

$$A = (1+\delta)I - \delta \mathbb{1}_{n\times n}, \tag{5.2}$$

where $\mathbb{1}_{n\times n}$ is the $n \times n$ matrix with all entries equal to 1 as before and $\delta = \frac{1}{n+5}$. We set $n = 100$ and plot the performance of CCD and RCD methods for the quadratic problem defined by this matrix. In Figure 4.2, we compare the convergence rate of CCD and RCD for an initial point that corresponds to a worst-case (left figure) and for a random choice of an initial point (right figure). We conclude that the asymptotic rate of CCD is faster than that of RCD demonstrating our results in Theorem 4.10 and Corolloary 4.11.

6. Conclusion. In this paper, we compared CCD and RCD methods on a class of least squares problems. We showed by a novel analysis that on this class CCD is always faster than RCD in terms of the worst-case asymptotic rate. We also gave a characterization of the best cyclic order to follow in CCD algorithms on our class of problems showing that with the best cyclic order one can achieve more than twice acceleration compared to RCD. Finally, we provided numerical experiments and examples that show the tightness of our results.

REFERENCES

[1] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM Journal on Optimization, 23 (2013), pp. 2037–2060.

[2] M. Belkin, P. Niyogi, and V. Sindhwani, *Manifold regularization: A geometric framework for learning from labeled and unlabeled examples*, Journal of Machine Learning Research, 7 (2006), pp. 2399–2434.

[3] D. P. Bertsekas, *Nonlinear programming*, Athena Scientific, 1999.

[4] D. P. Bertsekas, *Convex Optimization Algorithms*, Athena Scientific, 2015.

[5] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Inc., 1989.

[6] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[7] G. M. Engel and H. Schneider, *Cyclic and diagonal products on a matrix*, Linear Algebra and its Applications, 7 (1973), pp. 301 – 335.

[8] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, *Pathwise coordinate optimization*, The Annals of Applied Statistics, 1 (2007), pp. 302–332.

[9] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, Journal of Statistical Software, 33 (2010), pp. 1–22.

[10] J. F. C. Kingman, *A convexity property of positive matrices*, The Quarterly Journal of Mathematics, 12 (1961), pp. 283–284.

[11] S. J. Kirkland and M. Neumann, *Group inverses of M-matrices and their applications*, CRC Press, 2012.

[12] Z.-Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, Journal of Optimization Theory and Applications, 72 (1992), pp. 7–35.

[13] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: a general approach*, Annals of Operations Research, 46 (1993), pp. 157–178.

[14] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, *Walk-sums and belief propagation in gaussian graphical models*, Journal of Machine Learning Research, 7 (2006), pp. 2031–2064.

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

[16] R. D. Nussbaum, *Convexity and log convexity for the spectral radius*, Linear Algebra and its Applications, 73 (1986), pp. 59 – 122.

[17] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, SIAM, 2000.

[18] R. J. Plemmons, *M-matrix characterizations.Inonsingular m-matrices*, Linear Algebra and its Applications, 18 (1977), pp. 175 – 188.

[19] A. Rantzer, *Distributed control of positive systems*, ArXiv:1203.0047, (2014).

[20] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Mathematical Programming, 156 (2016), pp. 433–484.

[21] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Mathematical Programming, 156 (2016), pp. 433–484.

[22] A. Saha and A. Tewari, *On the nonasymptotic convergence of cyclic coordinate descent methods*, SIAM Journal on Optimization, 23 (2013), pp. 576–601.

[23] R. Sun and M. Hong, *Improved iteration complexity bounds of cyclic block coordinate descent for convex problems*, in NIPS, 2015, pp. 1306–1314.

[24] R. Sun and Y. Ye, *Worst-case Complexity of Cyclic Coordinate Descent: $O(n^2)$ Gap with Randomized Version*, ArXiv:1604.07130, (2016).

[25] R. Tutunov, H. Bou-Ammar, and A. Jadbabaie, *Distributed SDDM solvers: Theory & applications*, arXiv:1508.04096, (2015).

[26] R. S. Varga, *Matrix iterative analysis*, Springer Science & Business Media, 2009.

[27] S. J. Wright, *Coordinate descent algorithms*, Mathematical Programming, 151 (2015), pp. 3–34.

[28] D. M. Young, *Iterative solution of large linear systems*, Academic Press, New York, NY, 1971.

Appendix A. Proof of Lemma 4.9.

Suppose the largest eigenvalue of B_α has a multiplicity of 1. Then,

$$\rho(B_\alpha) = \lim_{t\to\infty} [\mathrm{tr}\left((B_\alpha)^t\right)]^{1/t}. \tag{A.1}$$

In order to find the diagonal entries of $(B_\alpha)^t$, we consider the graph generated by the matrix B_α and define the weight of a walk as the product of the weights of the corresponding edges in the walk. We then observe that the ith diagonal of the matrix $(B_\alpha)^t$ can be written as the summation of weights of all closed walks of length t (from the ith node to itself). In particular, consider a valid closed walk w that contains edges $(i_s, i_{s+1})_{s=0}^{t-1}$ such that $i_0 = i_t = i$ and $[B_\alpha]_{i_s,i_{s+1}} > 0$ for all s. Then, we can define a symmetric walk w' with edges $(i_{s+1}, i_s)_{s=0}^{t-1}$ and the ith diagonal entry of $(B_\alpha)^t$ contains the weights of both w and w' as summands. Furthermore, the weight of the walk w can be written as $\phi_\alpha(w) = e^{c_w\alpha}\phi_0(w)$, for some integer c_w, where

$$\phi_0(w) = \prod_{s=0}^{t-1} [B_0]_{i_s,i_{s+1}}.$$

The weight of the symmetric walk w' is then found by $\phi_\alpha(w') = e^{-c_w\alpha}\phi_0(w)$ since B_0 is symmetric. Therefore, the ith diagonal entry of $(B_\alpha)^t$ can be found as follows

$$[(B_\alpha)^t]_{i,i} = \sum_{\text{all valid walks } w} \frac{e^{c_w\alpha} + e^{-c_w\alpha}}{2}\,\phi_0(w).$$

It is easy to observe that $\cosh(c_w\alpha) = \frac{e^{c_w\alpha}+e^{-c_w\alpha}}{2}$ is a strictly log-convex function of α for any $c_w \neq 0$. Thus, if there exists a walk w for which $c_w \neq 0$, then $\mathrm{tr}\left((B_\alpha)^t\right)$ is a strictly log-convex function of α since $\phi_0(w) > 0$ for all valid walks. On the other hand, $\mathrm{tr}\left((B_\alpha)^t\right)$ is constant in α if and only if $c_w = 0$ for all valid walks, which implies that the graph is bipartite since starting from an arbitrary node i it is not possible to return back to node i in odd number of steps. This together with (A.1) imply the statement of the lemma.

For the case the largest eigenvalue of B_α has a multiplicity of at least 2, we consider the matrix $\tilde{B}_\alpha(\epsilon) = B_\alpha + \epsilon I$, whose largest eigenvalue has a multiplicity of 1 for any $\epsilon > 0$. Using the same arguments as above, we can conclude that the statement of the lemma holds for any $\tilde{B}_\alpha(\epsilon)$ with $\epsilon > 0$ and taking the limit as $\epsilon \to 0^+$ concludes the proof of the lemma.

Appendix B. Example Achieving Lower and Upper Bounds. Consider solving the linear system $Ax = 0$ where A is defined as follows

$$A = \begin{bmatrix} 1 & -\delta \\ -\delta & 1 \end{bmatrix}$$

for some $\delta \in (0, 1)$. The CCD algorithm applied to this problem has the following iteration matrix

$$C = \begin{bmatrix} 0 & \delta \\ 0 & \delta^2 \end{bmatrix},$$

whereas the expected RCD iteration matrix is

$$R = \left(I - \frac{A}{2}\right)^2 = \begin{bmatrix} 1/2 & \delta/2 \\ \delta/2 & 1/2 \end{bmatrix}^2 = \frac{1}{4}\begin{bmatrix} 1+\delta^2 & 2\delta \\ 2\delta & 1+\delta^2 \end{bmatrix}.$$

The eigendecomposition of this matrix can be found as follows

$$R = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1+\delta}{2} & 0 \\ 0 & \frac{1-\delta}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^{-1}.$$

Therefore, after ℓ epochs the distance of the iterates generated by RCD starting from the initial point $x^0 = [a, b]^T$ becomes

$$\begin{aligned} \mathbb{E}\left|\left|x^\ell - x^*\right|\right| = \mathbb{E}\left|\left|x^\ell\right|\right| \geq \left|\left|\mathbb{E}x^\ell\right|\right| = \left|\left|R^\ell x^0\right|\right| &= \left|\left|\begin{bmatrix} \left(\frac{1+\delta}{2}\right)^\ell a \\ \left(\frac{1-\delta}{2}\right)^\ell b \end{bmatrix}\right|\right| \\ &= \sqrt{\left(\frac{1+\delta}{2}\right)^{2\ell} a^2 + \left(\frac{1-\delta}{2}\right)^{2\ell} b^2}. \\ &\geq \left(\frac{1+\delta}{2}\right)^\ell |a| \\ &\geq \delta^\ell |a|. \end{aligned}$$

Therefore, in order to achieve a solution in the ϵ-neighborhood of the optimal solution $x^* = 0$, i.e., to attain $\left|\left|x^\ell - x^*\right|\right| = \epsilon$, the RCD method requires

$$N_R(\epsilon) \geq \frac{\log \epsilon}{\log \delta} - \frac{\log |a|}{\log \delta}$$

epochs, for any $a \neq 0$.

On the other hand, for the CCD algorithm, we have

$$C^\ell = \begin{bmatrix} 0 & \delta^{2\ell-1} \\ 0 & \delta^{2\ell} \end{bmatrix},$$

and consequently the suboptimality of the iterates generated by the CCD algorithm is

$$\left|\left|C^\ell x_0\right|\right| = \delta^{2\ell}\sqrt{b^2 + \frac{1}{\delta^2}b^2}.$$

Therefore, in order to achieve a solution in the ϵ-neighborhood of the optimal solution $x^* = 0$, i.e., to attain $\left|\left|x^\ell - x^*\right|\right| = \epsilon$, the CCD method requires

$$N_C(\epsilon) = \frac{\log \epsilon}{2\log \delta} - \frac{\log\left(b^2 + \frac{1}{\delta^2}b^2\right)}{4\log \delta}$$

epochs.

Note that for small ϵ the first terms in the expression of $N_J(\epsilon)$ and $N_C(\epsilon)$ are dominant. In particular we have,

$$\lim_{\epsilon \to 0^+} \frac{N_R(\epsilon)}{N_C(\epsilon)} =\geq \frac{2 \log \delta}{\log \delta} = 2, \tag{B.1}$$

for any $a \neq 0$.