When Cyclic Coordinate Descent Outperforms Randomized Coordinate Descent

Authored by:

Asuman Ozdaglar
Mert Gurbuzbalaban
Pablo A. Parrilo
Nuri Vanli

Abstract

The 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 CCD relative to RCD, which 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.

1 Paper Body

We consider solving smooth convex optimization problems using the coordinate descent (CD) method. 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 ik 2 {1, 2, . . . , n} is selected and the decision vector is updated to approximately minimize the objective function in the ik -th coordinate [3, 4]. The CD method can be deterministic or randomized depending on the choice of the update coordinates. If the coordinate indices ik are chosen in a cyclic manner from the set {1, 2, . . . , n}, then the method is called the cyclic coordinate descent (CCD) method. When ik is

sampled uniformly from the set {1, 2, . . . , n}, the resulting method is called the randomized coordinate descent (RCD) method.1 The CD method has a long history in optimization and its convergence has been studied extensively in 80s and 90s (cf. [5, 12, 13, 18]). It has seen a resurgence of recent interest because of its applicability and superior empirical performance in machine learning and large-scale data analysis [7, 8]. 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 [11, 21, 22] for problems with non-smooth terms), and Beck and Tetruashvili [1], which provided rate estimates for block coordinate gradient descent method that yields rate results for CCD with exact minimization for quadratic problems. Tighter rate estimates (with respect to [1]) for CCD are then presented in [23]. These rate estimates suggest that CCD can be slower than RCD (precisely O(n2) times slower for quadratic

1 Note that there are other coordinate selection rules as well (such as the Gauss-Southwell rule [17]). However, in this paper, we focus on cyclic and randomized rules.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

problems, 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 (e.g., see numerical examples in [1, 24]). This gap was investigated in [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 problem2 min

x2Rn

1 T x Ax, 2

(1)

where A is a positive definite matrix. We consider two problem classes: i) A is a 2-cyclic matrix, whose formal definition is given in Definition 4.4, but an equivalent and insightful definition is the bipartiteness of the graph induced by the matrix A D, where D is the diagonal part of A; ii) A is an M-matrix, i.e., the off-diagonal entries of A are nonpositive. These matrices arise in a large number of applications such as in inference in attractive Gaussian-Markov random fields [14] and in minimization of quadratic forms of graph Laplacians (for which A = D W , where W Pdenotes the weighted adjacency matrix of the graph and D is

the diagonal matrix given by Di,i = j Wi,j), for example in spectral partitioning [6] and semisupervised learning [2]. We build on the seminal work of Young [27] and Varga [25] 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 for 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 that of RCD. We provide numerical experiments in Section 5 and concluding remarks in Section 6. Notation: For a matrix H, we let Hi denote its ith row and Hi,j denote its entry at the ith row and jth column. For a vector x, we let xi denote its ith entry. Throughout the paper, we reserve superscripts for iteration counters of iterative algorithms and use x? to denote the optimal solution of problem (1). For a vector x, kxk denotes its Euclidean norm and for a matrix H, ——H—— denotes its operator norm. For matrices, and ? 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 Method

Starting from an initial point x0 2 Rn , the coordinate descent (CD) method, at each iteration k, picks a coordinate of x, say ik , and updates the decision vector by performing exact minimization along the ik th coordinate, which for problem (1) yields xk+1 = xk

1 Aik ,ik

Ai k x k e i k ,

k = 0, 1, 2, . . . ,

(2)

where eik is the unit vector, whose ik th 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 ik , along the gradient component direction (with the particular step size of Ai 1,i). The coordinate index ik can be selected according to a deterministic or randomized k k rule: 2

For ease of presentation, we consider minimization of 12 xT Ax, yet our results directly extend for problems of the type 12 xT Ax bT x for any b 6= 0.

2

? When ik is chosen using the cyclic rule with order 1, . . . , n (i.e., ik = k (mod 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

LT ,

where D is the diagonal part of A and L is the strictly lower triangular part of A. Then, over each epoch ‘ 0 (where an epoch is defined to be consecutive n iterations), the CCD iterations given in (2) can be written as (‘+1)n

xCCD

= C x‘n CCD ,

where

C = (D

L)

1

LT .

(3)

Note that the epoch in (3) is equivalent to one iteration of the Gauss-Seidel (GS) method applied to the first-order optimality condition of (1), i.e., applied to the linear system Ax = 0 [26]. ? When ik is chosen at random among {1, . . . , n} with probabilities {p1 , . . . , pn } independently at each iteration k, the resulting algorithm is called the randomized coordinate descent (RCD) method3 . Given the kth iterate generated by the RCD algorithm, i.e., xkRCD , we have ? ? k Ek xk+1 SD 1 A xkRCD , (4) RCD — xRCD = I

where S = diag(p1 , . . . , pn) contains the coordinate sampling probabilities and the conditional expectation Ek is taken over the random variable ik given xkRCD . Using the nested property of the expectations, the RCD iterations in expectation over each epoch ‘ 0 satisfy (‘+1)n

ExRCD

3

= R Ex‘n RCD

with R := I

SD

1

A

n

.

(5)

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 [25]. 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 Rate of Converge for Iterative Algorithms

Consider an iterative algorithm with update rule x(‘+1)n = Cx‘n (e.g., the CCD algorithm). The reduction in the distance to the optimal solution of the

iterates generated by this algorithm after ‘ epochs is given by x‘n x? C ‘ (x0 x?) = . (6) 0 ? ——x x —— ——x0 x? ——

Note that the right hand side of (6) can be as large as C ‘ , hence in the worst-case, the average decay of distance at each epoch of this algorithm is C ‘ 1/‘

1/‘

. Over any finite epochs ‘

1, we

1/‘

have C ?(C) and C ! ?(C) as ‘ ! 1 by Gelfand?s formula. Thus, we define the asymptotic worst-case convergence rate of an iterative algorithm (with iteration matrix C) as follows ! x‘n x? 1 Rate(CCD) := lim sup log = log (?(C)) . (7) ‘!1 x0 2Rn ‘ ——x0 x? —— ‘

‘

We emphasize that this notion has been used extensively for studying the performance of iterative methods such as GS and Jacobi methods [5, 18, 25, 27]. Note that according to our definition in (7), larger rate means faster algorithm and we will use these terms interchangably in throughout the paper. 3

sd

3

Analogously, for a randomized algorithm with expected update rule Ex(‘+1)n = R Ex‘n (e.g., the RCD algorithm), we consider the asymptotic convergence of the expected iterate error E(x‘n) x? and define the asymptotic worst-case convergence rate as ! E(x‘n) x? 1 Rate(RCD) := lim sup log = log (?(R)) , (8) ‘!1 x0 2Rn ‘ ——x0 x? —— Note that in (8), we use the distance of the expected iterates Ex‘n x? as our convergence criterion. One can also use the expected distance (or the squared distance) of the iterates E x‘n x? as the convergence criterion, which is a stronger convergence criterion than the one in (8). This follows since E x‘n x? Ex‘n x? by Jensen?s inequality and any convergence rate on ‘n ? E x x immediately implies at least the same convergence rate on Ex‘n x? as well. Since we consider the reciprocal case, i.e., obtain a convergence rate on Ex‘n x? and show that it is slower than that of CCD, our results naturally imply that the convergence rate on E x‘n x? 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 1, consider the 2n ? 2n symmetric matrix ? 1 0 0n?n A = I L LT , where L = 2 n?n , (9) n 1n?n 0n?n

and 1n?n is the n ? n matrix with all entries equal to 1 and 0n?n is the n ? 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 LT), it is easy to check that 1 1/n and 1 + 1/n are eigenvalues of A with the corresponding T T 11?n] . The remaining 2n 2 eigenvalues of A are eigenvectors [11?n 11?n] and [11?n equal to 1. The iteration matrix of the CCD algorithm when applied to the problem

in (1) with the matrix (9) can be found as ? 1 0 1 C = n?n n12 n?n . 0n?n n3 1n?n

The eigenvalues of C are all zero except the eigenvalue of 1/n2 with the corresponding eigenvector [n11?n , 11?n]T . Therefore, ?(C) = 1/n2 and Rate(CCD) = log(?(C)) = 2 log n. On the other hand, the spectral radius of the expected iteration matrix of RCD can be found as ? ?n 1 min (A) ?(R) = 1 1 , min (A) = n n

which yields Rate(RCD) =

log(?(R)) ? log n. Thus, we conclude Rate(CCD) Rate(RCD)

2,

for all n

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)) for which the asymptotical 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 4.1. A is a symmetric positive definite matrix whose smallest eigenvalue is ? and the diagonal entries of A are 1. 4

If A is a positive semidefinite matrix, then our results will still hold, where ? is the smallest non-zero eigenvalue of A and x? is the projection of x0 onto the null space of A. Moreover, given any positive definite matrix A with diagonals D 6= I, the diagonal entries of the preconditioned matrix D 1/2 AD 1/2 are 1. Therefore, Assumption 4.1 is mild. The relationship between the smallest eigenvalue of the original matrix and the preconditioned matrix are as follows. Let ¿ 0 and Lmax denote the smallest eigenvalue and the largest diagonal entry of the original matrix, respectively. Then, the smallest eigenvalue of the preconditioned matrix satisfies ? /Lmax . Remark 4.2. For the RCD algorithm, the coordinate index ik 2 {1, . . . , n} (at iteration k) can be chosen using different probability distributions {p1 , . . . , pn }. Two common choices of distributions A are pi = n1 , for all i 2 {1, . . . , n} and pi = PN i,iA [15]. Since by Assumption 4.1, the diagonal J=1

j,j

entries of A are 1, both of these distributions reduces to pi = n1 , for all i 2 {1, . . . , n}. 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. This worst-case rate has been presented in many works in the literature for strongly convex optimization problems [15, 26]. The proof is deferred to Appendix. Lemma 4.3. Suppose Assumption 4.1 holds. Then, the spectral radius of the

expected iteration matrix R of the RCD algorithm (defined in (5)) is given by ? ? ?n ?(R) = 1 . (10) n 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 rate of CCD is more than two times faster than that of RCD. Definition 4.4 (2-Cyclic Matrix). A matrix H is 2-cyclic if there exists a permutation matrix P such that ? 0 B1 T P HP = D + , (11) B2 0 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 (11). Then, the graph induced by the matrix H D is bipartite. The definition in (11) is first introduced in [27], where it had an alternative name: Property A. A generalization of this property is later introduced by Varga to the class of p-cyclic matrices [25] where p 2 can be arbitrary. We next introduce the following definition that will be useful in Theorem 4.12 and explicitly identify the class of matrices that satisfy this definition in Lemma 4.6. Definition 4.5 (Consistently Ordered Matrix). For a matrix H, let H = HD HL HU be its decomposition such that HD is a diagonal matrix, HL (and HU) is a strictly lower (and upper) triangular matrix. If the eigenvalues of the matrix ?HL + ?HU HD are independent of ? for any 2 R and ? 6= 0, then H is said to be consistently ordered. Lemma 4.6. [27, Theorem 4.5] A matrix H is 2-cyclic if and only if there exists a permutation matrix P such that P HP T is consistently ordered. In the next theorem, we characterize the convergence rate of CCD algorithm applied to a 2-cyclic matrix. Since ?(R) 1 ? by Lemma 4.3, the following theorem indicates that the spectral radius of the CCD iteration matrix is smaller than ?2 (R). Theorem 4.7. Suppose Assumption 4.1 holds and A is a consistently ordered 2-cyclic matrix. Then, the spectral radius of the CCD algorithm is given by ?(C) = (1

2

?) .

Remark 4.8. Note that our motivating example in (9) is an example of a consistently ordered 2-cyclic matrix, for which Theorem 4.7 is applicable. In particular, for the example in (9), we can apply Theorem 4.7 with ? = 1 1/n leading to ?(C) = 1/n2 , which exactly coincides with our previous computations of ?(C) in Section 3.2. We also give an example in Appendix F, for which CCD is twice faster than RCD for any arbitrary initialization with probability one. 5

The following corollary states that the asymptotic worst-case 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 4.7 and definitions (7)-(8). Corollary 4.9. Suppose Assumption 4.1 holds and A is a consistently ordered 2-cyclic matrix. Then, the asymptotic worst-case rates of CCD and RCD satisfy Rate(CCD) = 2?n , Rate(RCD)

where ?n :=

log(1 ?) . n log 1 n?

(12)

In the following remark, we highlight several properties of the constant ?n . Remark 4.10. ?n is a monotonically increasing function of n over the interval

[1, 1), where ?1 = 1 ?) and limn!1 ?n = log(1 ¿ 1. Furthermore, lim?!0+ ?n = 1. ? We emphasize that the CCD method applied to 1 is equivalent to the Gauss-Seidel (GS) algorithm applied to the linear system Ax = 0 and when A is a 2-cyclic matrix, the GS algorithm is twice as fast as the Jacobi algorithm [25, 27]. Hence, when A is a 2-cyclic matrix and ? is sufficiently small, the RCD method is approximately as fast as the Jacobi algorithm. 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. Definition 4.11 (M-matrix). A real matrix A with Ai,j ? 0 for all i 6= j is an M-matrix if A has the decomposition A = sI B such that B 0 and s ?(B). We emphasize that M-matrices arise in a variety of applications such as in belief propagation over Gaussian graphical models [14] and in distributed control of positive systems [20]. Furthermore, graph Laplacians are M-matrices, therefore solving linear systems with M-matrices (or equivalently solving (1) for an M-matrix A) arise in a variety of applications for analyzing random walks over graphs as well as distributed optimization and consensus problems over graphs (cf. [10] 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 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 that of RCD for irreducible M-matrices. Theorem 4.12. Suppose Assumption 4.1 holds, A is an irreducible M-matrix and n 2. Then, the iteration matrix of the CCD algorithm C = (I L) 1 LT satisfies the following inequality (1

?)2 ? ?(C) ?

1 ? , 1+?

(13)

where the inequality on the left holds with equality if and only if A is a consistently ordered matrix. An immediate consequence of Theorem 4.12 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 4.13. The standard CCD method follows the standard cyclic order (1, 2, . . . , n) as described in Section 2. However, we can construct a CCD method that follows an alternative deterministic order by considering a permutation ? of {1, 2, . . . , n}, and choosing the coordinates according to the order (?(1), ?(2), . . . , ?(n)) instead. For any given order ?, (1) can be reformulated as follows min

x? 2Rn

1 T x A? x ? , 2 ?

where

A? := P? AP?T

and x? = P? x,

where P? is the corresponding permutation matrix of ?. Supposing that Assumption 4.1 holds, the corresponding CCD iterations for this problem can

be written as follows x(‘+1)n = C? x‘n ? ? ,

where C? = (I 6

L?)

1

LT?

and L? = P? LP? .

If A is an irreducible M-matrix and satisfies Assumptions 4.1, then so does A? . Consequently, Theorem 4.12 yields the same upper and lower bounds (in (13)) on ?(C?) as well, i.e., the spectral radius of the iteration matrix of CCD with any cyclic order ? satisfies (1

?)2 ? ?(C?) ?

1 ? , 1+?

(14)

where the inequality on the left holds with equality if and only if A? is a consistently ordered matrix. Therefore, if a consistent order ? ? exists, then the CCD method with the consistent order ? ? attains the smallest spectral radius (or equivalently, the fastest asymptotic worst-case convergence rate) among the CCD methods with any cyclic order. Remark 4.14. The irreducibility of A is essential to derive the lower bound in (13) of Theorem 4.12. However, the upper bound in (13) holds even when A is a reducible matrix. We next compare the spectral radii bounds for CCD (given in Theorem 4.12) and RCD (given in Lemma 4.3). Since ? ¿ 0, the right-hand side of (13) can be relaxed to (1 ?)2 ? ?(C) ¡ 1 ?. 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 4.15. Suppose Assumption 4.1 holds, A is an irreducible M-matrix and n 2. Then, the asymptotic worst-case rates of CCD and RCD satisfy 1 ¡ ?n ¡

Rate(CCD) ? 2?n , Rate(RCD)

where

?n :=

log(1 ?) , n log 1 n?

(15)

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 4.16. Suppose Assumption 4.1 holds, A is an irreducible M-matrix and n 2. Then, we have Rate(CCD) lim = 2. ?!0+ Rate(RCD)

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), where A is an n ? n matrix defined as follows ? 1 0 0 T A = I L L , where L = (16) n n 1 0 , ? n 2 2 and 1 n2 ? n2 is the n2 ? n2 matrix with all entries equal to 1. Here, it can be easily checked that A is a consistently

ordered, 2-cyclic matrix. By Theorem 4.7 and Corolloary 4.9, the asymptotic worst-case convergence rate of CCD on this example is 2?n = 2

log(1 ?) log(0.5) ? 2.77 ? = 1 n log 1 n 50 log 1 200

(17)

times faster than that of RCD. This is illustrated in Figure 1 (left), 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 ‘). On the other hand, when the matrix A is not consistently ordered, according to Theorem 4.12, 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 permutation matrix P and replace A with AP := P AP T in the optimization problem (1) (This is equivalent to solving the system AP x = 0) so that AP is not consistently ordered (We generate P randomly and compute AP). Figure 1 (right) 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. 7

Consistent Ordering, Worst-Case Initialization

Inconsistent Ordering, Worst-Case Initialization

?4

?6

x? ——

?2

?4

?6

?8

log ——x‘

?2

x? ——

0

log ——x‘

0

?8

?10

?10

CCD RCD Expected RCD

?12

?14

1

2

3

4

CCD RCD Expected RCD

?12

5 6 7 Number of Epochs ‘

8
9
?14
10
1
2
3
4
5 6 7 Number of Epochs ‘
8
9
10

Figure 1: Distance to the optimal solution of the iterates of CCD and RCD for the cyclic matrix in (16) (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) figure corresponds to the consistent (inconsistent) ordering for the same quadratic optimization problem. M-Matrix, Worst-Case Initialization

M-Matrix, Random Initialization

?4
?4 ?
?2
——x‘ x? —— ——x0 x? ——
?2
?
0
——x‘ x? —— ——x0 x? ——
0
?
? log
?6
log
?6
?8
?8
?10
?12
?10
CCD RCD Expected RCD 0
20
40 60 Number of Epochs ‘
80
?12
100
CCD RCD Expected RCD 0
20
40 60 Number of Epochs ‘

80
100

Figure 2: Distance to the optimal solution of the iterates of CCD and RCD for the M-matrix matrix in (18) for the worst-case initialization (left figure) and a random initialization (right figure). We next consider the case, where A is an irreducible positive definite M-matrix. In particular, we consider the matrix A = (1 +)I 1n?n , (18) 1 where 1n?n is the n ? n matrix with all entries equal to 1 as before and = 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 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 4.12 and Corolloary 4.15.

6

Conclusion

In this paper, we compare the CCD and RCD methods for quadratic problems, whose Hessian is a 2-cyclic matrix or an M-matrix. We show by a novel analysis that for these classes of quadratic problems, CCD is always faster than RCD in terms of the asymptotic worst-case rate. We also give a characterization of the best cyclic order to use in the CCD algorithm for these classes of problems and show that with the best cyclic order, CCD enjoys more than a twice faster asymptotic worst-case rate with respect to RCD. We also provide numerical experiments that show the tightness of our results.

2 References

[1] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. SIAM Journal on Optimization, 23(4):2037?2060, 2013. [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, 7:2399?2434, 2006.

8

[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. PrenticeHall, Inc., 1989. [6] F. R. K. Chung. Spectral Graph Theory. American Mathematical Society, 1997. [7] J. Friedman, T. Hastie, H. H?fling, and R. Tibshirani. Pathwise coordinate optimization. The Annals of Applied Statistics, 1(2):302?332, 2007. [8] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1?22, 2010. [9] J. F. C. Kingman. A convexity property of positive matrices. The Quarterly Journal of Mathematics, 12(1):283?284, 1961. [10] S. J. Kirkland and M. Neumann. Group inverses of M-matrices and their applications. CRC Press, 2012. [11] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods.

Mathematical Programming, 152(1):615?642, 2015. [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(1):7?35, 1992. [13] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. Annals of Operations Research, 46(1):157?178, 1993. [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, 7:2031?2064, 2006. [15] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, 22(2):341?362, 2012. [16] Roger D. Nussbaum. Convexity and log convexity for the spectral radius. Linear Algebra and its Applications, 73(Supplement C):59 ? 122, 1986. [17] J. Nutini, M. Schmidt, I. H. Laradji, M. Friedlander, and H. Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, pages 1632?1641, 2015. [18] J. Ortega and W. Rheinboldt. Iterative Solution of Nonlinear Equations in Several Variables. Society for Industrial and Applied Mathematics, 2000. [19] R. J. Plemmons. M-matrix characterizations.I?nonsingular m-matrices. Linear Algebra and its Applications, 18(2):175 ? 188, 1977. [20] A. Rantzer. Distributed control of positive systems. ArXiv:1203.0047, 2014. [21] P. Richt?rik and M. Tak??c. Parallel coordinate descent methods for big data optimization. Mathematical Programming, 156(1):433?484, 2016. [22] A. Saha and A. Tewari. On the nonasymptotic convergence of cyclic coordinate descent methods. SIAM Journal on Optimization, 23(1):576?601, 2013. [23] R. Sun and M. Hong. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In Advances in Neural Information Processing Systems 28, pages 1306?1314. 2015. [24] R. Sun and Y. Ye. Worst-case Complexity of Cyclic Coordinate Descent: O(n2) Gap with Randomized Version. ArXiv:1604.07130, 2016. [25] R. S. Varga. Matrix iterative analysis. Springer Science & Business Media, 2009. [26] S. J. Wright. Coordinate descent algorithms. Mathematical Programming, 151(1):3?34, 2015. [27] D. M. Young. Iterative solution of large linear systems. Academic Press, 1971.

9