IMA Journal of Numerical Analysis (2019) **39**, 1246–1275
doi:10.1093/imanum/dry040
Advance Access publication on 30 July 2018

Random permutations fix a worst case for cyclic coordinate descent

Ching-pei Lee and Stephen J. Wright*
Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, USA
ching-pei@cs.wisc.edu *Corresponding author: swright@cs.wisc.edu

[Received on 25 May 2017; revised on 4 March 2018]

Variants of the coordinate descent approach for minimizing a nonlinear function are distinguished in part by the order in which coordinates are considered for relaxation. Three common orderings are cyclic (CCD), in which we cycle through the components of x in order; randomized (RCD), in which the component to update is selected randomly and independently at each iteration; and random-permutations cyclic (RPCD), which differs from CCD only in that a random permutation is applied to the variables at the start of each cycle. Known convergence guarantees are weaker for CCD and RPCD than for RCD, though in most practical cases, computational performance is similar among all these variants. There is a certain type of quadratic function for which CCD is significantly slower than for RCD; a recent paper by Sun & Ye (2016, Worst-case complexity of cyclic coordinate descent: $O(n^2)$ gap with randomized version. *Technical Report*. Stanford, CA: Department of Management Science and Engineering, Stanford University. arXiv:1604.07130) has explored the poor behavior of CCD on functions of this type. The RPCD approach performs well on these functions, even better than RCD in a certain regime. This paper explains the good behavior of RPCD with a tight analysis.

Keywords: coordinate descent; randomization; permutations.

1. Introduction

The basic (component wise) coordinate descent framework for the smooth unconstrained optimization problem

$$\min f(x), \quad \text{where } f : \mathbb{R}^n \to \mathbb{R} \text{ is smooth and convex}, \tag{1.1}$$

is shown in Algorithm 1. Here we denote

$$\nabla_i f(x) = \big[\nabla f(x)\big]_i, \quad e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}, \tag{1.2}$$

where the single nonzero in e_i appears in position i. Each outer cycle (indicated by index ℓ) is called an 'epoch', with each epoch consisting of n iterations (indexed by j). The counter $k = \ell n + j$ keeps track of the total number of iterations. At each iteration, component $i(\ell, j)$ of x is selected for updating; a step is taken along the negative gradient direction in this component only.

There are several variants within this simple framework. One important source of variation is the choice of coordinate $i = i(\ell, j)$. Three popular choices are as follows:

- Cyclic coordinate descent (CCD): $i(\ell, j) = j + 1$.
- Randomized coordinate descent (RCD), also known as stochastic coordinate descent: $i(\ell, j)$ is chosen uniformly at random from $\{1, 2, \ldots, n\}$—sampling with replacement.

Algorithm 1 Coordinate descent

Choose $x^0 \in \mathbb{R}^n$;
for $\ell = 0, 1, 2, \ldots$ **do**
  for $j = 0, 1, 2, \ldots, n-1$ **do**
    Define $k = \ell n + j$;
    Choose index $i = i(\ell, j) \in \{1, 2, \ldots, n\}$;
    Choose $\alpha_k > 0$;
    $x^{k+1} \leftarrow x^k - \alpha_k \nabla_i f(x^k) e_i$;
  end for
end for

- Random-permutations cyclic coordinate descent (RPCD): At the start of epoch ℓ, we choose a random permutation of $\{1, 2, \ldots, n\}$, denoted by $\pi_{\ell+1}$. Index $i(\ell, j)$ is chosen to be the $(j+1)$th entry in $\pi_{\ell+1}$. This approach represents sampling without replacement, within each epoch.

(Other ways to choose $i(\ell, j)$ include weighted forms of RCD in which $i(\ell, j)$ is selected from a nonuniform distribution, and a Gauss–Southwell form in which $i(\ell, j)$ is the component that maximizes $|\nabla_i f(x^k)|$.)

When f is a convex quadratic function, and when α_k in Algorithm 1 is chosen to minimize f exactly along each coordinate direction, these variants are simply different variants of the Gauss–Seidel approach for solving the equivalent system of linear equations.

The coordinate descent approach is enjoying renewed popularity because of its usefulness in data analysis applications. Its convergence properties have come under renewed scrutiny. We refer to Wright (2015b) for a discussion of the state of the art as of 2015 but make a few additions and updates here, with an emphasis on results concerning linear convergence of the function values, by which we mean epoch wise convergence of the form

$$f\left(x^{(\ell+1)n}\right) - f^* \leqslant \rho\left(f(x^{\ell n}) - f^*\right) \quad \text{for some } \rho \in (0, 1), \tag{1.3}$$

where ρ is typically much closer to 1 than to 0 and f^* is the optimal value of (1.1). For randomized methods, we consider a corresponding expression *in expectation:*

$$\mathbb{E}\left[f\left(x^{(\ell+1)n}\right) - f^*\right] \leqslant \rho \mathbb{E}\left[f\left(x^{\ell n}\right) - f^*\right], \tag{1.4}$$

where the expectation is taken over all random variables encountered in the algorithm. When (1.3) holds, a reduction in function error by a factor of ε can be attained in approximately $|\log \varepsilon|/(1-\rho)$ epochs. We sometimes refer to $1/(1-\rho)$ as the 'complexity' of an algorithm for which (1.3) or (1.4) holds.

1.1 *Characterizing the objective*

We preface a discussion of linear convergence rates with some definitions of certain constants associated with f. We assume for simplicity that the domain of f is the full space $\mathbb{R}^n$. The component Lipschitz constants L_i, $i = 1, 2, \ldots, n$ satisfy

$$\left|\nabla_i f(x + te_i) - \nabla_i f(x)\right| \leqslant L_i |t| \quad \forall\, x \in \mathbb{R}^n \text{ and } t \in \mathbb{R}. \tag{1.5}$$

We have

$$L_{\max} := \max_{i=1,2,\ldots,n} L_i, \quad L_{\min} := \min_{i=1,2,\ldots,n} L_i, \quad L_{\text{avg}} := \sum_{i=1}^{n} L_i/n. \tag{1.6}$$

The standard Lipschitz constant L is defined so that

$$\left\|\nabla f(x + d) - \nabla f(x)\right\| \leqslant L\|d\| \quad \forall\, x, d \in \mathbb{R}^n. \tag{1.7}$$

(Here and throughout we use $\|\cdot\|$ to denote $\|\cdot\|_2$.) For reasonable choices of the constants in (1.5), (1.6) and (1.7) the following bounds are satisfied:

$$1 \leqslant \frac{L}{L_{\max}} \leqslant n. \tag{1.8}$$

The following property of Łojasiewicz (1963) is useful in proving linear convergence:

$$\left\|\nabla f(x)\right\|^2 \geqslant 2\mu\left[f(x) - f^*\right] \quad \text{for some } \mu > 0. \tag{1.9}$$

This property holds for f strongly convex (with modulus of strong convexity μ), and for the case in which f grows quadratically with distance from a nonunique minimizing set, as in the 'optimal strong convexity' condition of Liu & Wright (2015, (1.2)). It also holds generically for convex quadratic programs, even when the Hessians are singular. Further, condition (1.9) holds for the functional form considered by Luo & Tseng (1992, 1993), which is

$$f(x) = g(Ex), \quad \text{where } E \in \mathbb{R}^{m\times n} \text{ and } g : \mathbb{R}^m \to \mathbb{R} \text{ strongly convex}, \tag{1.10}$$

without any conditions on E. (For a proof, see Appendix C.) In Karimi *et al*. (2016), property (1.9) is called the Polyak–Łojasiewicz condition.

In this paper our focus is on the case of f convex quadratic, that is,

$$f(x) = \tfrac{1}{2}x^{\mathrm{T}}Ax, \quad \text{where } A \text{ is symmetric positive semidefinite}. \tag{1.11}$$

For this function the values of L_i, L, $L_{\max}$ and μ are

$$\mu = \lambda_{\min,\text{nz}}(A), \quad L_i = A_{ii},\ i = 1, 2, \ldots, n; \quad L_{\max} = \max_{i=1,2,\ldots,n} A_{ii}; \quad L = \|A\|_2; \tag{1.12}$$

where $\lambda_{\min,\mathrm{nz}}(\cdot)$ denotes the minimum nonzero eigenvalue. For such functions the upper bound in (1.8) is achieved by $A = \mathbf{1}\mathbf{1}^{\mathrm{T}}$ (where $\mathbf{1} = (1,1,\ldots,1)^{\mathrm{T}}$), for which $L_i = 1$, $i = 1,2,\ldots,n$; $L_{\max} = 1$; and $L = n$.

We have not included a linear term in (1.11), but note that there is no loss of generality in doing so. If we were to consider instead

$$f(x) = \tfrac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x = \tfrac{1}{2}(x-x^*)^{\mathrm{T}}A(x-x^*) - \tfrac{1}{2}b^{\mathrm{T}}A^{-1}b, \quad \text{where } x^* = A^{-1}b$$

(note that x^* is the minimizer of this function), the main results of Sections 2 and 3 would continue to hold, except that in several theorems the initial iterate x^0 would be replaced by $x^0 - x^*$, and $f(x)$ is replaced by $f(x) - f(x^*)$.

1.2 *Linear convergence results for CD variants*

Luo & Tseng (1992) prove linear convergence for a function of the form (1.10), where they require E to have no zero columns. They obtain expressions for the constant ρ in (1.3) for two variants of CD—a Gauss–Southwell variant and an 'almost cyclic' rule—but these constants are difficult to characterize in terms of fundamental properties of f. In Luo & Tseng (1993), the same authors analyse a family of methods (including CD) for more general functions that satisfy a local error bound of the form $\|x - P(x)\| \leqslant \chi\|\nabla f(x)\|$ (where $P(x)$ is the projection of x onto the solution set of (1.1) and χ is some constant). Again, their analysis is not clear about how the constant ρ of (1.3) depends on the properties of f.

A family of linear convergence results is proved in Beck & Tetruashvili (2013, Theorem 3.9) for the case in which f is strongly convex (immediately extendable to the case in which f satisfies the condition (1.9)). For constant stepsizes $\alpha_k \equiv \alpha \leqslant 1/L_{\max}$, convergence of the form (1.3) holds with

$$\rho \leqslant 1 - \frac{\mu}{(2/\alpha)(1 + nL^2\alpha^2)}. \tag{1.13}$$

In particular, for $\alpha = 1/L$, we have $\rho \leqslant 1 - \mu/(2L(n+1))$. The upper bound on ρ is optimized by steplength $\alpha = 1/(\sqrt{n}L)$, for which $\rho \leqslant 1 - \mu/(\sqrt{n}L)$. For the case in which f is a convex quadratic (1.11) and an exact line search is performed at each iteration (that is, $\alpha_k = 1/A_{ii}$, where $i = i(\ell, j)$ is the index to be updated in iteration j of Algorithm 1), Beck & Tetruashvili (2013, (3.23)) show that $\rho \leqslant 1 - \mu/(2L_{\max}(1 + n^2L^2/\mu^2))$ in expression (1.3). Paradoxically, as noted by Sun & Ye (2016), use of the exact steplength leads to a considerably slower rate bound than the conservative fixed choices. The bound for this case is improved in Sun & Ye (2016) to

$$\rho \leqslant 1 - \max\left\{\frac{\mu L_{\min}}{nLL_{\mathrm{avg}}}, \frac{\mu L_{\min}}{L^2(2+\log n/\pi)^2}, \frac{\mu L_{\min}}{n^2L_{\mathrm{avg}}^2}\right\}. \tag{1.14}$$

For the random-permutations cyclic version RPCD, the convergence theory in Beck & Tetruashvili (2013) can be applied without modification to attain the bounds given above. As we discuss below, however, the practical performance of RPCD is sometimes much better than these bounds would suggest.

Convergence of the sampling-with-replacement variant RCD for strongly convex unconstrained problems was analysed by Nesterov (2012). It follows from the convergence theory of Nesterov (2012, Theorem 2) that (1.4) holds over the i.i.d. uniformly random choices of indices $i(\ell, j)$ with

$$\rho \leqslant \left(1 - \frac{\mu}{nL_{\max}}\right)^n \approx 1 - \frac{\mu}{L_{\max}}. \tag{1.15}$$

A different convergence rate is proved in Nesterov (2012, Theorem 5), namely,

$$\mathbb{E}\big(f(x^k) - f^*\big) \leqslant C\left(1 - \frac{2\mu}{n(L_{\max} + \mu)}\right)^k, \tag{1.16}$$

for some constant C depending on the initial point. This is an R-linear expression, obtained from Q-linear convergence of the modified function $f(x) - f(x^*) + \sum_i L_i(x_i - x_i^*)^2/2$, where x^* is the (unique) solution of (1.1). It indicates a complexity of approximately $|\log \varepsilon|(L_{\max} + \mu)/(2\mu)$.

An important benchmark in studying the convergence rates of coordinate descent is the steepest-descent (SD) method, which takes a step from x^k along all coordinates simultaneously, in the direction $-\nabla f(x^k)$. For some important types of functions, including empirical-risk-minimization functions that arise in data analysis, the computational cost of one SD step is comparable to the cost of one epoch of Algorithm 1 (see Wright, 2015b). Standard analysis of SD shows that fixed-steplength variants applied to functions satisfying (1.9) have linear convergence of the form (1.3) (with one iteration of SD replacing one epoch of Algorithm 1) with $\rho = 1 - \mu/L$. This worst-case complexity is not improved qualitatively by using exact line searches.

In comparing convergence rates between CCD and SD (on the one hand) and RCD (on the other hand), we see that the former tend to depend on L while the latter depends on $L_{\max}$. These bounds suggest that CCD may tend to track the performance of SD, while RCD could be significantly better if the ratio $L/L_{\max}$ is large, that is, toward the upper end of its range in (1.8). The phenomenon of large values of $L/L_{\max}$ is captured well by convex quadratic problems (1.11) in which the Hessian A has a large contribution from $\mathbf{1}\mathbf{1}^T$. Such matrices were used in computations by one of the authors in 2015 (see Wright, 2015a; reported briefly in Wright, 2015b). These tests showed that on such matrices, RCD was indeed much faster than CCD (and also SD). The performance of RPCD was as fast as that of RCD; it did not track CCD as the obvious worst-case analysis would suggest. Later work, reported in Wright (2015c), identified the matrix

$$A := \delta I + (1 - \delta)\mathbf{1}\mathbf{1}^{\mathrm{T}}, \quad \text{where } \delta \in \big(0, n/(n-1)\big) \tag{1.17}$$

(where $\mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{T}}$) as being the archetype of a problem with large ratio $L/L_{\max}$. This matrix has one dominant eigenvalue $\delta + n(1 - \delta)$ with eigenvector $\mathbf{1}$, with the other $(n - 1)$ eigenvalues equal to δ. (This matrix also has $P^T A P = A$ for all permutation matrices P—a property that greatly simplifies the analysis of RPCD variants, as we see below.) In Wright (2015c) the RPCD variant was shown to be significantly superior to CCD in computational tests. Independently, Sun & Ye (2016) studied this same matrix (1.17), using analysis to explain the practical advantage of RCD over CCD and showing that the performance of CCD approaches its worst-case theoretical bound. RPCD is also studied in Sun & Ye (2016, Proposition 3.4, Section C.2), the results suggesting similar behavior for RPCD and RCD on problem (1.11), (1.17). However, these results are based on upper bounds on the quantity $\|\mathbb{E}(x^k)\|$.

By Jensen's inequality this quantity provides a lower bound for $\mathbb{E}f(x^k)$ and also for $\mathbb{E}\|x^k\|^2$, but not an upper bound. (The latter is the focus of this paper.)

The matrix (1.17) is also studied in Arjevani *et al*. (2016), which investigates the tightness of the worst-case theoretical Q-linear convergence rate for RCD applied to the problem (1.11), (1.17) proved in Nesterov (2012). This paper shows a lower bound for $\mathcal{O}(\|\mathbb{E}[x^k]\|)$ for RCD, but not for the expected objective value.

1.3 *Motivation and outline*

Our focus in this paper is to analyse the performance of RPCD for minimizing (1.11) with A defined in (1.17). Our interest in RPCD is motivated by computational practice. Much has been written about randomized optimization algorithms (particularly stochastic gradient and coordinate descent) in recent years. The analysis usually applies to sampling-with-replacement versions, but the implementations almost always involve a sampling-without-replacement scheme. The reasons are clear: convergence analysis is much more straightforward for sampling with replacement while for sampling *without* replacement, implementations are more efficient, involving less data movement. Moreover, it has long been folklore in the machine-learning community that sampling-without-replacement schemes perform better in practice. In this paper we take a step toward bringing the analysis into line with the practice, by giving a tight analysis of the sampling-without-replacement scheme RPCD, on a special but important function that captures perfectly the advantages of randomized schemes over a deterministic scheme.

In Section 2, we derive tools for analysing epoch wise convergence of CD variants on convex quadratic problems (1.11), focusing on the permutation-invariant matrix (1.17) and recalling results for the CCD and RCD variant in this case (obtained from Sun & Ye, 2016 and Nesterov, 2012). Section 3 contains our results for RPCD applied to (1.11) with the permutation-invariant matrix (1.17), characterizing its convergence rate in terms of a two-parameter recurrence. The relationship of this two-parameter sequence to the expected function value at the end of each epoch is described in Theorem 3.3. Our main result, Theorem 3.4, gives bounds on these two parameters in terms of δ (the parameter that defines (1.17)) and epoch number. These bounds indicate that the convergence rate of RPCD matches that of RCD and both are much faster than CCD on the problem defined by (1.11) and (1.17). We also note that a slightly tighter bound on the asymptotic behavior of the two-variable recurrence can be obtained from the spectral radius of the 2×2 matrix governing this recurrence, in a regime in which δ is close to zero. We derive an estimate of this spectral radius in (3.15), using results from Appendix B. Theorem 3.5 explores the behavior of the randomized methods on the very first iteration, showing that a significant decrease can be expected just on this one iteration. (Similar results can be expected for the cyclic variant CCD, as we remark in comments following Theorem 3.5.)

Empirical verification of our analysis of RPCD and computational comparisons with CCD and RCD are presented in Section 4. The theoretical results are confirmed nicely in all cases. We conclude with some discussions in Section 5.

2. Convergence of CD variants on convex quadratics

We consider the application of CCD and RPCD to the convex quadratic problem (1.11). This problem has solution $x^* = 0$ with optimal objective $f^* = 0$. We assume that the matrix A is diagonally scaled so that

$$A_{ii} = 1 \quad \text{for } i = 1, 2, \ldots, n. \tag{2.1}$$

Under this assumption the step of Algorithm 1 with exact line search will have the form

$$x^{k+1} = x^k - \frac{1}{A_{ii}}(Ax^k)_i e_i = x^k - (Ax^k)_i e_i, \quad \text{with } k = \ell n + j \text{ and } i = i(\ell, j). \tag{2.2}$$

Some variants of CD methods applied to (1.11) can be viewed as Gauss–Seidel methods applied to the system $Ax = 0$. Cyclic CD corresponds to standard Gauss–Seidel, whereas RCD and RPCD are variants of randomized Gauss–Seidel.

2.1 *CCD and RPCD convergence rates: general A*

Writing $A = L + D + L^T$, where L is strictly lower triangular and D is the diagonal, one epoch of the CCD method can be written as

$$x^{(\ell+1)n} = -(L + D)^{-1}(L^{\mathrm{T}} x^{\ell n}) = C x^{\ell n}, \quad \text{where } C := -(L + D)^{-1} L^{\mathrm{T}}. \tag{2.3}$$

By applying formula (2.3) recursively we obtain the following expression for the iterate generated after ℓ epochs of CCD:

$$x^{\ell n}_{\mathrm{CCD}} = C^\ell x^0, \quad f\left(x^{\ell n}_{\mathrm{CCD}}\right) = \frac{1}{2}(x^0)^T (C^{\mathrm{T}})^\ell A C^\ell x^0. \tag{2.4}$$

The average improvement in f per epoch is obtained from the formula

$$\rho_{\mathrm{CCD}}(A, x^0) := \lim_{\ell \to \infty} \left(f\left(x^{\ell n}_{\mathrm{CCD}}\right)/f(x^0)\right)^{1/\ell}. \tag{2.5}$$

To obtain a bound on this quantity we denote the eigenvalues of C by γ_i, $i = 1, 2, \ldots, n$, and recall that the spectral radius $\rho(C)$ is $\max_{i=1,2,\ldots,n} |\gamma_i|$. Since A is positive definite we have $\rho(C) < 1$ (Golub & Van Loan, 2012, Theorem 11.2.3). We have from Gelfand's formula (Gelfand, 1941) that

$$\rho(C) = \lim_{\ell \to \infty} \|C^\ell\|^{1/\ell}. \tag{2.6}$$

We can obtain a bound on $\rho_{\mathrm{CCD}}(A, x^0)$ in terms of $\rho(C)$:

$$\begin{aligned}
\rho_{\mathrm{CCD}}(A, x^0) &:= \lim_{\ell \to \infty} \left(f\left(x^{\ell n}_{\mathrm{CCD}}\right)/f(x^0)\right)^{1/\ell} \\
&= \lim_{\ell \to \infty} \left(x_0^T (C^T)^\ell A C^\ell x_0 / x_0^{\mathrm{T}} A x_0\right)^{1/\ell} \\
&= \lim_{\ell \to \infty} \left(\|A^{1/2} C^\ell x_0\|_2^2 / \|A^{1/2} x_0\|_2^2\right)^{1/\ell} \\
&= \lim_{\ell \to \infty} \left(\|(A^{1/2} C^\ell A^{-1/2})(A^{1/2} x_0)\|_2^2 / \|A^{1/2} x_0\|_2^2\right)^{1/\ell} \\
&\leqslant \lim_{\ell \to \infty} \left(\|A^{1/2} C^\ell A^{-1/2}\|_2^2\right)^{1/\ell} \\
&\leqslant \lim_{\ell \to \infty} \mathrm{cond}(A)^{1/\ell} \|C^\ell\|^{2/\ell} = \rho(C)^2.
\end{aligned} \tag{2.7}$$

We can describe each epoch of RPCD algebraically by using a permutation matrix P_l to represent the permutation π_l on epoch $l-1$. We split the matrix $P_l^{\mathrm{T}} A P_l$ and define the operator C_l as

$$P_l^{\mathrm{T}} A P_l = L_l + D_l + L_l^{\mathrm{T}}, \quad C_l := -\left(L_l + D_l\right)^{-1} L_l^{\mathrm{T}}. \tag{2.8}$$

The iterate generated after ℓ epochs of RPCD is

$$x_{\mathrm{RPCD}}^{\ell n} = P_\ell C_\ell P_\ell^{\mathrm{T}} P_{\ell-1} C_{\ell-1} P_{\ell-1}^{\mathrm{T}} \ldots P_1 C_1 P_1^{\mathrm{T}} x^0. \tag{2.9}$$

(Note that in epoch $l-1$ the elements of x are permuted according to the permutation matrix P_l, then operated on with C_l, then the permutation is reversed with P_l^{T}.) The function value after ℓ epochs is

$$f\left(x_{\mathrm{RPCD}}^{\ell n}\right) = \frac{1}{2}\left(x^0\right)^{\mathrm{T}} \left(P_1 C_1^{\mathrm{T}} P_1^{\mathrm{T}} \ldots P_\ell C_\ell^{\mathrm{T}} P_\ell^{\mathrm{T}} A P_\ell C_\ell P_\ell^{\mathrm{T}} \ldots P_1 C_1 P_1^{\mathrm{T}}\right) x^0. \tag{2.10}$$

If we could take the expected value of this quantity over all random permutations $P_1, P_2, \ldots, P_\ell$ we would have good expected-case bounds on the convergence of RPCD. This expectation is quite difficult to manipulate in general (though, as we see below, it is not so difficult for A defined by (1.17)). When the elements of x^0 are distributed according to $N(0,1)$ we have

$$\mathbb{E}_{x^0} f\left(x_{\mathrm{RPCD}}^{\ell n}\right) = \frac{1}{2}\mathrm{trace}\left(P_1 C_1^{\mathrm{T}} P_1^{\mathrm{T}} \ldots P_\ell C_\ell^{\mathrm{T}} P_\ell^{\mathrm{T}} A P_\ell C_\ell P_\ell^{\mathrm{T}} \ldots P_1 C_1 P_1^{\mathrm{T}}\right). \tag{2.11}$$

Figure 1 shows typical computational results of the CCD and RPCD variants of Algorithm 1 in the case in which the eigenvalues of A follow a log-uniform distribution, with $\kappa(A) \approx 10^4$. The eigenvectors form an orthogonal matrix with random orientation. Here we plot the relative *expected* values with respect to x^0 of the f on the vertical axis, that is, $\mathbb{E}_{x^0}(f(x^{\ell n}))/\mathbb{E}_{x^0}(f(x^0))$ (see (2.11) for $\mathbb{E}_{x^0} f(x_{\mathrm{RPCD}}^{\ell n})$; similar formulas apply for $\mathbb{E}_{x^0} f(x_{\mathrm{CCD}}^{\ell n})$ and $\mathbb{E}_{x^0} f(x^0)$). This figure captures the typical relative behavior of CCD and RPCD for 'benign' distributions of eigenvalues: there is little difference in performance between the two variants.

2.2 *CD variants applied to permutation-invariant A*

In our search for the simplest instance of a matrix A for which the superiority of randomization is observed, we arrived at the matrix (1.17). As mentioned above the eigenvalues of A are

$$\delta + n(1-\delta), \delta, \delta, \ldots, \delta, \quad \text{where } \delta \in (0, n/(n-1)).$$

The restriction in (1.17) ensures that A has the following properties:

- symmetric and positive definite;
- unit diagonals: $A_{ii} = 1$, $i = 1, 2, \ldots, n$;
- invariant under symmetric permutations of the rows and columns, that is, $P^{\mathrm{T}} A P = A$ for any $n \times n$ permutation matrix P;
- $L/L_{\max}$ is close to its maximum value of n when δ is small, opening a wide gap between the worst-case theoretical behaviors of CCD and RCD.

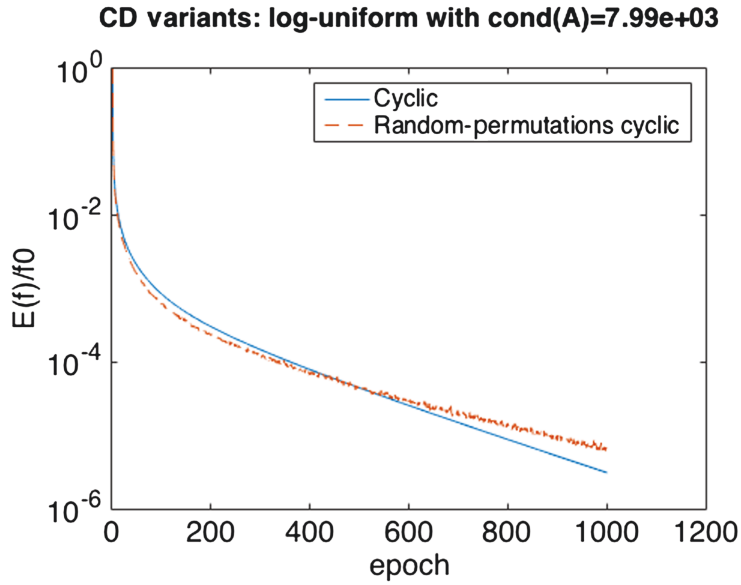
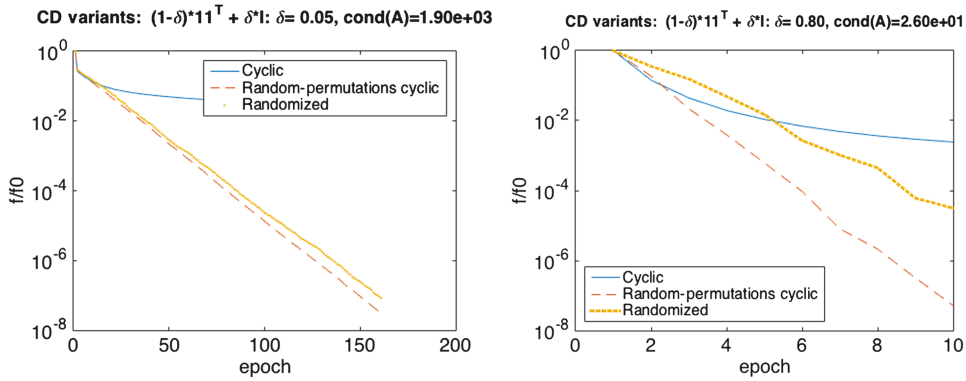


FIG. 1. CCD and RPCD on convex quadratic objective, for log-uniform eigenvalue distribution.



(a) $\delta = .05$ (b) $\delta = .8$

FIG. 2. CCD, RPCD and RCD on convex quadratic objective, with A defined by (1.17) with $n = 100$ and various δ.

Figure 2 shows results for the CCD, RPCD and RCD variants on the matrix A from (1.17) with $n = 100$ and two different values of δ. Here the vertical axis shows actual function values (not expected values) relative to $f(x^0)$, for some particular x^0 whose elements are drawn i.i.d. from $N(0, 1)$. For both values of δ, both randomized variants are much faster than CCD. For the larger value of δ RPCD has a clear advantage over RCD. Our analysis below supports these empirical observations.

We now derive expressions for the epoch iteration matrix C of Section 2.1 for the specific case of the permutation-invariant matrix (1.17). This is needed for the analysis of RPCD on this matrix. By applying the splitting (2.3) to (1.17) we have

$$D = I, \quad L = (1-\delta)E, \quad \text{where } E = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 & 0 \end{bmatrix}. \tag{2.12}$$

Thus, defining

$$\bar{L} := -(L+D)^{-1}, \tag{2.13}$$

we have

$$\bar{L}_{ij} = \begin{cases} -1 & \text{if } i = j, \\ (1-\delta)\delta^{i-j-1} & \text{if } i > j, \\ 0 & \text{if } i < j, \end{cases} \tag{2.14a}$$

$$C = (1-\delta)\bar{L}E^{\mathrm{T}}. \tag{2.14b}$$

Writing $\bar{L}$ explicitly we have

$$\bar{L} = \begin{bmatrix} -1 & 0 & 0 & 0 & \dots & 0 \\ (1-\delta) & -1 & 0 & 0 & \dots & 0 \\ (1-\delta)\delta & (1-\delta) & -1 & 0 & \dots & 0 \\ (1-\delta)\delta^2 & (1-\delta)\delta & (1-\delta) & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ (1-\delta)\delta^{n-2} & (1-\delta)\delta^{n-3} & (1-\delta)\delta^{n-4} & \dots & \dots & -1 \end{bmatrix}.$$

We have from (2.14b) and the properties of E and $\bar{L}$ that

$$C_{ij} = (1-\delta)\sum_{\ell=1}^{n} \bar{L}_{i\ell}E_{j\ell} = (1-\delta)\sum_{\ell=1}^{\min(i,j-1)} \bar{L}_{i\ell}.$$

Thus, for $i < j$ we have

$$\begin{aligned} C_{ij} &= (1-\delta)\sum_{\ell=1}^{i}\bar{L}_{i\ell} \\ &= (1-\delta)\left[(1-\delta)(\delta^{i-2}+\delta^{i-3}+\cdots+\delta+1)-1\right] \\ &= (1-\delta)\left[(1-\delta)\frac{1-\delta^{i-1}}{1-\delta}-1\right] \\ &= -(1-\delta)\delta^{i-1}. \end{aligned}$$

For the complementary case $i \geqslant j$ we have

$$\begin{aligned} C_{ij} &= (1-\delta)\sum_{\ell=1}^{j-1}\bar{L}_{i\ell} \\ &= (1-\delta)\left[(1-\delta)(\delta^{i-2}+\delta^{i-3}+\cdots+\delta^{i-j})\right] \\ &= (1-\delta)^2\delta^{i-j}(\delta^{j-2}+\delta^{j-3}+\cdots+1) \\ &= (1-\delta)^2\delta^{i-j}\frac{1-\delta^{j-1}}{1-\delta} \\ &= (1-\delta)\delta^{i-j}(1-\delta^{j-1}) \\ &= (1-\delta)(\delta^{i-j}-\delta^{i-1}). \end{aligned}$$

To summarize we have

$$C_{ij} = \begin{cases} -(1-\delta)\delta^{i-1} & \text{for } i < j, \\ (1-\delta)(\delta^{i-j}-\delta^{i-1}) & \text{for } i \geqslant j. \end{cases} \tag{2.15}$$

2.3 *Convergence rates of CCD and RCD on the permutation-invariant A*

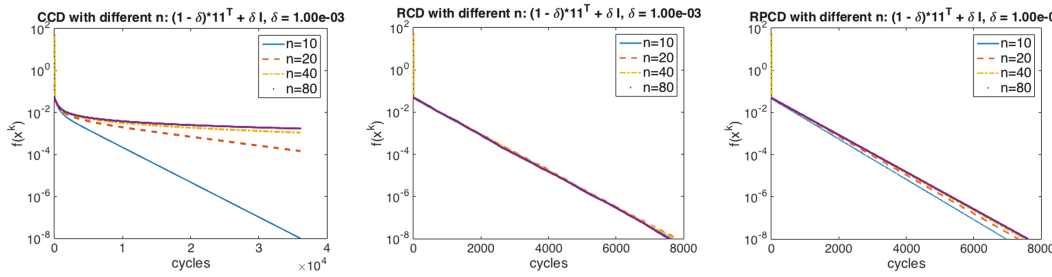
Here we examine the theoretical convergence rate of CCD on the quadratic function with Hessian (1.17) by using the results of Sun & Ye (2016).

Recalling the rate (1.14) from Sun & Ye (2016, Proposition 3.1) and substituting the following quantities for (1.17):

$$L = n(1-\delta)+\delta, \quad L_{\min} = 1, \quad L_{\text{avg}} = 1, \quad \mu = \delta, \tag{2.16}$$

we find that

$$\rho_{\text{CCD}}(\delta, x^0) \leqslant 1 - \max\left\{\frac{\delta}{n(n(1-\delta)+\delta)}, \frac{\delta}{(n(1-\delta)+\delta)^2(2+\log n/\pi)^2}, \frac{\delta}{n^2}\right\}.$$



(a) CCD (b) RCD (c) RPCD

FIG. 3. Convergence of f for CCD, RCD and RPCD applied to (1.11), (1.17), with $\delta = 0.001$ and $n = 10, 20, 40, 80$. Convergence rate of CCD deteriorates as n grows, as predicted, while the convergence rates of RCD and RPCD are independent of n.

(We use $\rho_{\mathrm{CCD}}(\delta, x^0)$ in place of $\rho_{\mathrm{CCD}}(A, x^0)$ to emphasize the dependence of A in (1.17) on the parameter δ.) By making the mild assumption that $\delta \leqslant 3/4$, this expression simplifies to

$$\rho_{\mathrm{CCD}}(\delta, x^0) \leqslant 1 - \frac{\delta}{n\big(n(1-\delta)+\delta\big)}. \tag{2.17}$$

On the other hand, Sun and Ye show the following lower bound on $\rho_{\mathrm{CCD}}(\delta, x^0)$ (obtained by substituting from (2.16) into Sun & Ye, 2016, Theorem 3.1):

$$\rho_{\mathrm{CCD}}(\delta, x^0) \geqslant \left(1 - \frac{2\delta\pi^2}{n\big(n(1-\delta)+\delta\big)}\right)^2. \tag{2.18}$$

By combining (2.17) and (2.18) we see that for small values of δ/n the average epoch wise decrease in error is $\rho_{\mathrm{CCD}}(\delta, x^0) = 1 - c\delta/n^2$, for some moderate value of c. Classical numerical analysis for Gauss–Seidel derives similar dependency on n^2 for this case from the eigenvalues of A, D and L; see Samarskii & Nikolaev (1989), Young (1971, p. 464) and Hackbusch (2016, Theorem 3.44). This dependency on n is confirmed empirically, by running CCD for A with the same δ but different n, as shown in Fig. 3(a).

For RCD we have, by substituting the values in (2.16) into (1.15), that the expected per-epoch improvement in error is given by

$$\rho_{\mathrm{RCD}}(\delta, \text{predicted}) \leqslant \left(1 - \frac{\mu}{nL_{\max}}\right)^n = \left(1 - \frac{\delta}{n}\right)^n \approx 1 - \delta + \mathcal{O}(\delta^2). \tag{2.19}$$

This result suggests that complexity of RCD is $\mathcal{O}(n^2)$ times better than CCD for small δ and that its rate does not depend strongly on n. This independence of n is confirmed empirically by Fig. 3(b). Expression

(1.16) suggests a slightly better complexity for RCD of roughly $|\log \varepsilon|(1+\delta)/(2\delta)$ epochs, rather than $|\log \varepsilon|/\delta$ epochs, corresponding to replacing $1-\delta$ in (2.19) by

$$\left(1 - \frac{2\delta}{n(1+\delta)}\right)^n \approx 1 - \frac{2\delta}{1+\delta}. \tag{2.20}$$

A kind of lower bound on the per-iterate improvement of RCD on the problem (1.11), (1.17) can be found by setting $x_i = (-1)^i$, $i = 1, 2, \ldots, n$ with n even. It can be shown that the function values for this x and the next RCD iterate x^+ are

$$f(x) = \frac{1}{2}\delta n, \quad f(x^+) = \frac{1}{2}\delta(n-\delta) = \left(1 - \frac{\delta}{n}\right) f(x),$$

regardless of the component of x chosen for updating. This expression reveals a one-iteration improvement that matches the upper bound (1.15). However, as with some other lower-bound examples, the longer-term behavior of the iteration sequence is more difficult to predict. This same example provides a Q-linear rate in the quantity $f(x) - f(x^*) + \sum_i L_i (x_i - x_i^*)^2/2$ of $(1 - 2\delta/(n(1+\delta)))$, exactly matching the upper bound of (1.16), (2.20), proving that the R-linear rate of (1.16) is also tight, in a sense. A lower bound on $\|\mathbb{E}(x^k)\|$ is proved in Arjevani *et al*. (2016), but this does not translate into a lower bound of the expected function value.

Figure 3(c) shows that RPCD too has a convergence rate independent of n on this matrix. (The performances of RPCD and RCD are quite similar on the problems graphed.) The convergence rate of CCD deteriorates with n, according to the predictions above.

3. Convergence of RPCD for the permutation-invariant A

We now analyse the expected convergence behavior of RPCD on the convex quadratic problem (1.11) with permutation-invariant Hessian A defined by (1.17). We start by deriving a two-parameter recurrence that captures the behavior of the method over each epoch and by deriving an estimate for the expected convergence of $f(x^{\ell n})$ to zero, as a function of these parameters. In our main results we analyse the rate of convergence of this sequence of parameter pairs to zero.

3.1 *A two-parameter recurrence*

Since A in (1.17) is invariant under symmetric permutations the matrices L and D are the same for all $P^T A P$, where P is any permutation matrix. When considering RPCD applied to this problem we have in the notation of (2.8) that $C_\ell \equiv C$ for all ℓ. The expression (2.9) simplifies to

$$x^{\ell n}_{\text{RPCD}} = P_\ell C P_\ell^{\mathrm{T}} P_{\ell-1} C P_{\ell-1}^{\mathrm{T}} \ldots P_1 C P_1^{\mathrm{T}} x^0. \tag{3.1}$$

The function values are

$$f\left(x^{\ell n}_{\text{RPCD}}\right) = \frac{1}{2}\left(x^0\right)^{\mathrm{T}} \left(P_1 C^{\mathrm{T}} P_1^{\mathrm{T}} \ldots P_\ell C^{\mathrm{T}} P_\ell^{\mathrm{T}} A P_\ell C P_\ell^{\mathrm{T}} \ldots P_1 C P_1^{\mathrm{T}}\right) x^0. \tag{3.2}$$

We now analyse the expected value of the function (3.2) obtained after ℓ epochs of RPCD, where A has the form (1.17). Expectation is taken over the independent permutation matrices $P_\ell, P_{\ell-1}, \ldots, P_1$ in

succession, followed finally by expectation over x^0. We define $\bar{A}^{(t)}$, $t = 0, 1, 2, \ldots, \ell$ as

$$\bar{A}^{(t)} := \mathbb{E}_{P_{\ell-t+1},\ldots,P_\ell}\left(P_{\ell-t+1}C^{\mathrm{T}}P_{\ell-t+1}^{\mathrm{T}}\ldots P_\ell C^{\mathrm{T}}P_\ell^{\mathrm{T}}AP_\ell CP_\ell^{\mathrm{T}}\ldots P_{\ell-t+1}CP_{\ell-t+1}^{\mathrm{T}}\right),$$

and note that $\bar{A}^{(0)} = A$ and (by comparison with (3.2)) that

$$\mathbb{E}f\left(x_{\mathrm{RPCD}}^{\ell n}\right) = \frac{1}{2}\mathbb{E}_{x^0}\left[(x^0)^{\mathrm{T}}\bar{A}^{(\ell)}x^0\right]. \tag{3.3}$$

We have the following recursive relationship between successive terms in the sequence of $\bar{A}^{(t)}$ matrices:

$$\bar{A}^{(t)} = \mathbb{E}_{P_{\ell-t+1}}\left(P_{\ell-t+1}C^{\mathrm{T}}P_{\ell-t+1}^{\mathrm{T}}\bar{A}^{(t-1)}P_{\ell-t+1}CP_{\ell-t+1}^{\mathrm{T}}\right) = \mathbb{E}_P\left(PC^{\mathrm{T}}P^{\mathrm{T}}\bar{A}^{(t-1)}PCP^{\mathrm{T}}\right). \tag{3.4}$$

(We can drop the subscript on $P_{\ell-t+1}$ since the permutation matrices at each stage are i.i.d.) We will show by a recursive argument that each $\bar{A}^{(t)}$ has the form $\eta_t I + \nu_t \mathbf{1}\mathbf{1}^{\mathrm{T}}$, for some parameters η_t and ν_t. Note that for $\bar{A}^{(t)}$ of this form we have $P^T\bar{A}^{(t)}P = \bar{A}^{(t)}$, a property that is crucial to our analysis. We derive a stationary iteration between the successive pairs (η_{t-1}, ν_{t-1}) and (η_t, ν_t) and reveal the convergence properties of RPCD by analysing the 2×2 matrix that relates successive pairs.

We start with a technical lemma.

LEMMA 3.1 Given any matrix $Q \in \mathbb{R}^{n\times n}$ and permutation matrix P selected uniformly at random from the set of all permutations Π we have

$$B := \mathbb{E}_P[PQP^{\mathrm{T}}] = \tau_1 I + \tau_2 \mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{3.5}$$

where

$$\tau_2 = \frac{\mathbf{1}^{\mathrm{T}}Q\mathbf{1} - \mathrm{trace}(Q)}{n(n-1)}, \quad \tau_1 = \frac{\mathrm{trace}(Q)}{n} - \tau_2. \tag{3.6}$$

Proof. For any $P \in \Pi$ if P shifts the ith position to the jth position then $(PQP^{\mathrm{T}})_{jj} = Q_{ii}$. Since the probability of taking any permutation from Π is identical we have

$$\mathscr{P}\left((PQP^{\mathrm{T}})_{jj} = Q_{ii}\right) = \frac{1}{n} \quad \forall\, i, j \in \{1, \ldots, n\}$$

(where $\mathscr{P}(\cdot)$ denotes probability). Therefore, each diagonal entry B is the average over all diagonal entries of Q:

$$B_{jj} = \frac{\sum_{i=1}^n Q_{ii}}{n}, \quad j = 1, 2, \ldots, n.$$

Consider permutations that shift the ith and the jth entries to the kth and the lth positions, respectively, that is,

$$(PQP^{\mathrm{T}})_{kl} = Q_{ij}. \tag{3.7}$$

Note that we always have $i \neq j \Rightarrow k \neq l$ because permutations are bijections from and to $\{1, \ldots, n\}$. Thus, there are $(n-2)!$ permutations in Π with the property (3.7). Under the same reasoning as before, each off-diagonal entry of B is the average of all off-diagonal entries of Q:

$$B_{kl} = \frac{\sum_{1 \leqslant i,j \leqslant n, i \neq j} Q_{ij}}{n(n-1)}, \quad k, l \in \{1, 2, \ldots, n\},\ k \neq l.$$

Finally, we obtain (3.6) by noting that $B_{ii} = \tau_1 + \tau_2$, while $B_{ij} = \tau_2$ for $i \neq j$. □

We have immediately from Lemma 3.1 that

$$\mathbb{E}_P(P^{\mathrm{T}} C^{\mathrm{T}} C P) = d_1 I + d_2 \mathbf{1}\mathbf{1}^{\mathrm{T}}, \quad \mathbb{E}_P(P^{\mathrm{T}} C^{\mathrm{T}} \mathbf{1}\mathbf{1}^{\mathrm{T}} C P) = m_1 I + m_2 \mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{3.8}$$

where

$$d_2 = \frac{\mathbf{1}^{\mathrm{T}} C^{\mathrm{T}} C \mathbf{1} - \mathrm{trace}(C^{\mathrm{T}} C)}{n(n-1)} = \frac{\|C\mathbf{1}\|_2^2 - \|C\|_F^2}{n(n-1)}, \tag{3.9a}$$

$$d_1 = \frac{\mathrm{trace}(C^{\mathrm{T}} C)}{n} - d_2 = \frac{\|C\|_F^2}{n} - d_2, \tag{3.9b}$$

$$m_2 = \frac{(\mathbf{1}^{\mathrm{T}} C \mathbf{1})^2 - (\mathbf{1}^{\mathrm{T}} C)(C^{\mathrm{T}} \mathbf{1})}{n(n-1)} = \frac{(\mathbf{1}^{\mathrm{T}} C \mathbf{1})^2 - \|C^{\mathrm{T}} \mathbf{1}\|_2^2}{n(n-1)}, \tag{3.9c}$$

$$m_1 = \frac{(\mathbf{1}^{\mathrm{T}} C)(C^{\mathrm{T}} \mathbf{1})}{n} - m_2 = \frac{\|C^{\mathrm{T}} \mathbf{1}\|_2^2}{n-1} - \frac{(\mathbf{1}^{\mathrm{T}} C \mathbf{1})^2}{n(n-1)}. \tag{3.9d}$$

Note that for (3.9c) and (3.9d) we used the property $\mathrm{trace}(AB) = \mathrm{trace}(BA)$.

The following theorem reveals the relationship between successive matrices in the sequence $\bar{A}^{(0)}, \bar{A}^{(1)}, \ldots$.

THEOREM 3.2 Consider solving (1.11) with the matrix A defined in (1.17) using RPCD. For $\bar{A}^{(t)}$ defined in (3.4), with $\bar{A}^{(0)} = A$, we have

$$\bar{A}^{(t)} = \eta_t I + \nu_t \mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{3.10}$$

where $(\eta_0, \nu_0) = (\delta, 1-\delta)$ and

$$\begin{bmatrix} \eta_{t+1} \\ \nu_{t+1} \end{bmatrix} = M \begin{bmatrix} \eta_t \\ \nu_t \end{bmatrix} = M^{t+1} \begin{bmatrix} \delta \\ 1-\delta \end{bmatrix} \quad \forall\, t \geqslant 0, \tag{3.11}$$

where

$$M := \begin{bmatrix} d_1 & m_1 \\ d_2 & m_2 \end{bmatrix}, \tag{3.12}$$

and d_1, d_2, m_1, m_2 are defined in (3.9).

Proof. We first prove (3.10) by induction. By (1.17) it holds at $t = 0$. Now assume it holds for $t = k$, for some η_k and ν_k, then for $k+1$ we have from (3.4),

$$\bar{A}^{(k+1)} = \mathbb{E}_P\big[P C^{\mathrm{T}} P^{\mathrm{T}} \bar{A}^{(k)} P C P^{\mathrm{T}}\big].$$

Because $\bar{A}^{(k)}$ is in the form (3.10) it is invariant to row and column permutations, that is, $P^T\bar{A}^{(k)}P = \bar{A}^{(k)}$ for all $P \in \Pi$. Hence,

$$\begin{aligned}\bar{A}^{(k+1)} &= \mathbb{E}_P\left[PC^{\mathrm{T}}\bar{A}^{(k)}CP^{\mathrm{T}}\right]\\ &= \eta_k\mathbb{E}_P[PC^{\mathrm{T}}CP^{\mathrm{T}}] + \nu_k\mathbb{E}_P[PC^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}}CP^{\mathrm{T}}]\\ &= \eta_k(d_1I + d_2\mathbf{1}\mathbf{1}^{\mathrm{T}}) + \nu_k(m_1I + m_2\mathbf{1}\mathbf{1}^{T})\\ &= (\eta_k d_1 + \nu_k m_1)I + (\eta_k d_2 + \nu_k m_2)\mathbf{1}\mathbf{1}^{\mathrm{T}},\end{aligned} \tag{3.13}$$

giving the result. □

We obtain a result for the expected value of f after ℓ epochs by taking the expectation as in (3.3), showing that the sequence of expected function values at the end of each epoch is governed by the behavior of the sequence $\{(\eta_\ell, \nu_\ell)\}$.

THEOREM 3.3 Consider solving (1.11) with the matrix A defined in (1.17) using RPCD. Then, using the notation of Theorem 3.2, we have

$$\mathbb{E}_{P_1,P_2,\ldots,P_\ell} f(x^{\ell n}) = \frac{1}{2}\left(\eta_\ell\|x^0\|^2 + \nu_\ell(\mathbf{1}^{\mathrm{T}}x^0)^2\right) \leqslant \left\|\begin{bmatrix}\eta_\ell\\ \nu_\ell\end{bmatrix}\right\| \max\left(\|x^0\|^2, (\mathbf{1}^{\mathrm{T}}x^0)^2\right).$$

Proof. The result is obtained by taking expectations with respect to $P_\ell, P_{\ell-1}, \ldots, P_1$ in (3.2) and using the definition of $\bar{A}^{(t)}$ (with $t = \ell$) together with Theorem 3.2. □

Figure 4 plots the expected value from Theorem 3.3 against the value of $f(x^{\ell n}_{\mathrm{RPCD}})$ obtained from (3.2) for particular random choices of x^0 and the permutation matrices $P_1, P_2, \ldots, P_\ell$, showing that the estimate in this one instance tracks the expected value closely. (This behavior is typical.)

3.2 *Convergence of the two-parameter recurrence*

It is evident from Theorems 3.2 and 3.3 (and Gelfand's formula) that the asymptotic convergence of the expected value of f is governed by $\rho(M)$, which, because of definitions (3.12) and (3.9), is a function of δ and n. In Fig. 2(b) and Table 1, we see that this rate is significantly better than those obtained for RCD and CCD when δ is not too close to zero (that is, $\delta \geqslant 0.2$). In this section we estimate the convergence rate of RPCD for δ close to zero, showing that in this regime it is close to the rate of approximately $1 - 2\delta$ obtained by RCD (2.20), and much faster than the rate of CCD discussed in (2.17) and (2.18), which is $1 - c\delta/n^2$, for some modest value of c.

We start with a rigorous bound on the convergence rate for the sequence $\{(\eta_\ell, \nu_\ell)\}$, without resorting to eigenvalue calculations for M. The details of bounding the elements of M (d_1, d_2, m_1 and m_2 from (3.9)) as functions of δ and n are shown in Appendix A. For the case of $\delta \in [0, 0.4]$ and $n \geqslant 10$ the formulas (A.11) yield the following bounds:

$$\begin{aligned}0 \leqslant d_1 &\leqslant 1 - 2\delta + 3.6\delta^2,\\ |m_2| &\leqslant 0.05\delta^2,\\ 0 \leqslant d_2 &\leqslant 1 - 2\delta + 3.2\delta^2,\\ |m_1| &\leqslant 0.15\delta^2.\end{aligned}$$

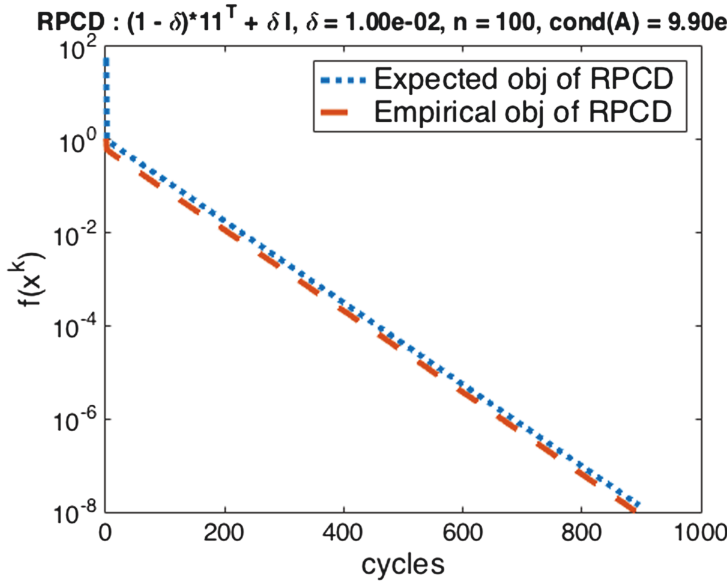


FIG. 4. Empirical objective value and expected objective value of RPCD.

TABLE 1 *Observed and predicted per-epoch convergence rates for CCD, RCD and RPCD, for various values of* δ ($n = 100$ *in all experiments)*

δ	0.80	0.50	0.33	0.20	0.10	0.03
$\rho_{\mathrm{CCD}}(\delta, x^0)$	0.9340	0.9924	0.9971	0.9988	0.9995	0.9998
$\rho(C)^2$	0.9342	0.9924	0.9971	0.9988	0.9995	0.9999
$\rho_{\mathrm{RCD}}(\delta, x^0)$	0.3146	0.4764	0.5945	0.7059	0.8287	0.9428
$\rho_{\mathrm{RCD}}(\delta, \text{predicted})$	0.4095	0.5123	0.6081	0.7161	0.8336	0.9434
$\rho_{\mathrm{RPCD}}(\delta, x^0)$	0.1054	0.3306	0.4929	0.6615	0.8178	0.9415
$\rho(M)$	0.1162	0.3289	0.4994	0.6635	0.8164	0.9412

By appealing to Theorems 3.2 and 3.3 we obtain our main result.

THEOREM 3.4 Consider solving (1.11), (1.17) with $\delta \in [0, 0.4]$ and $n \geqslant 10$ using RPCD. Then, using the notation of Theorem 3.2, we have

$$|\eta_\ell| \leqslant (1 - 2\delta + 4\delta^2)^{\ell-1}\delta, \quad |\nu_\ell| \leqslant (1 - 2\delta + 4\delta^2)^{\ell-1}\delta \quad \forall\, \ell \geqslant 1. \tag{3.14}$$

Thus, we have the following bound on the convergence of the expected value of the function:

$$\mathbb{E}_{P_1,P_2,\ldots,P_\ell} f(x^{\ell n}) \leqslant \frac{1}{2}(1 - 2\delta + 4\delta^2)^{\ell-1}\left(\|x^0\|^2 + (\mathbf{1}^{\mathrm{T}}x^0)^2\right)\delta \quad \forall\, \ell \geqslant 1.$$

Proof. Since $(\eta_0, \nu_0) = (\delta, 1-\delta)$ we have from (3.11) and using $\delta \in [0, 0.4]$ that

$$\begin{bmatrix}|\eta_1|\\|\nu_1|\end{bmatrix} \leqslant \begin{bmatrix}d_1 & |m_1|\\ d_2 & |m_2|\end{bmatrix}\begin{bmatrix}\delta\\(1-\delta)\end{bmatrix} \leqslant \begin{bmatrix}(1-2\delta+3.6\delta^2)\delta+0.15\delta^2(1-\delta)\\(1-2\delta+3.2\delta^2)\delta+0.05\delta^2(1-\delta)\end{bmatrix} \leqslant \begin{bmatrix}(1-1.85\delta+3.6\delta^2)\delta\\(1-1.95\delta+3.2\delta^2)\delta\end{bmatrix} \leqslant \begin{bmatrix}\delta\\\delta\end{bmatrix},$$

so that (3.14) holds for $\ell = 1$. Supposing that the bound holds for some value of $\ell \geqslant 1$ we have

$$\begin{aligned}\begin{bmatrix}|\eta_{\ell+1}|\\|\nu_{\ell+1}|\end{bmatrix} &\leqslant (1-2\delta+4\delta^2)^{\ell-1}\begin{bmatrix}(1-2\delta+3.6\delta^2) & 0.15\delta^2\\(1-2\delta+3.2\delta^2) & 0.05\delta^2\end{bmatrix}\begin{bmatrix}\delta\\\delta\end{bmatrix}\\ &\leqslant (1-2\delta+4\delta^2)^{\ell-1}\begin{bmatrix}(1-2\delta+3.75\delta^2)\delta\\(1-2\delta+3.25\delta^2)\delta\end{bmatrix}\\ &\leqslant (1-2\delta+4\delta^2)^{\ell}\begin{bmatrix}\delta\\\delta\end{bmatrix},\end{aligned}$$

verifying that the required bound still holds at $\ell+1$, thus proving (3.14).

The final claim follows directly from Theorem 3.3. □

This result indicates a global linear rate of at worst $1-2\delta+4\delta^2$, similar to the rate (2.20) obtained for RCD (identical to $\mathcal{O}(\delta)$) and much faster than the rate obtained for CCD in (2.17), (2.18).

By using slightly more refined estimates of the elements of M, which involve not strict upper bounds as in (A.11) but rather remainder terms containing higher powers of δ and/or $1/n$, we can obtain an estimate of $\rho(M)$. In Appendix B we obtain the following estimates of d_1, d_2, m_1 and m_2:

$$\begin{aligned}d_1 &= 1-2\delta-2\frac{\delta}{n}+2\delta^2+\mathcal{O}\left(\frac{\delta^2}{n}\right)+\mathcal{O}(\delta^3),\\ m_2 &= \mathcal{O}\left(\frac{\delta^2}{n}\right),\\ d_2 &= 1-\frac{2}{n}+\mathcal{O}(\delta),\\ m_1 &= \mathcal{O}\left(\frac{\delta^2}{n}\right).\end{aligned}$$

By substituting these estimates into (3.12) and calculating the spectral radius $\rho(M)$ as the largest root of the characteristic quadratic $\det(M-\lambda I)$ we obtain

$$\rho(M) = 1-2\delta-\frac{2\delta}{n}+2\delta^2+\mathcal{O}\left(\frac{\delta^2}{n}\right)+\mathcal{O}(\delta^3). \tag{3.15}$$

This asymptotic rate is identical to the rate for RCD (2.20) in the 1, δ and δ^2 terms, and is slightly better because of the presence of the $-2\delta/n$ term.

3.3 *The first iteration*

We noted in the numerical experiments (Figs 3 and 4) that the decrease in f over the first epoch of RPCD is rather dramatic. In fact, after just a single *iteration*, the function value was often of order δ for all three variants (CCD, RPCD and RCD). The following result supports this observation.

THEOREM 3.5 Consider solving (1.11) with the matrix A defined in (1.17) using RCD or RPCD with exact line search (2.2). Given any x^0 the expected function value after a single iteration satisfies

$$\mathbb{E}_i f(x^1) = \frac{1}{2}\delta\left(1-\frac{\delta}{n}\right)\|x^0\|^2 + \frac{1}{2}\delta(1-\delta)\left(1-\frac{2}{n}\right)(\mathbf{1}^{\mathrm{T}}x^0)^2 \leqslant \frac{1}{2}\delta\|x^0\|^2 + \delta f(x^0), \tag{3.16}$$

where i denotes the coordinate chosen for updating at the first iteration.

Proof. Note that i is chosen uniformly at random from $\{1, 2, \ldots, n\}$ for both RPCD and RCD. After one step of CD we have

$$\begin{aligned} x_i^1 &= x_i^0 - \left(x_i^0 + (1-\delta)\sum_{j\neq i} x_j^0\right) = -(1-\delta)\left(\sum_{j\neq i} x_j^0\right), \\ x_j^1 &= x_j^0 \quad \text{for } j \neq i. \end{aligned}$$

Thus, from (1.17) we have

$$\begin{aligned} f\left(x^1\right) &= \frac{1}{2}\left(x^1\right)^{\mathrm{T}} A x^1 \\ &= \frac{1}{2}\delta\left\|x^1\right\|^2 + \frac{1}{2}(1-\delta)\left(\sum_{j=1}^n x_j^1\right)^2 \\ &= \frac{1}{2}\delta\left[\sum_{j\neq i}\left(x_j^0\right)^2 + (1-\delta)^2\left(\sum_{j\neq i} x_j^0\right)^2\right] + \frac{1}{2}(1-\delta)\left[\sum_{j\neq i} x_j^0 - (1-\delta)\sum_{j\neq i} x_j^0\right]^2 \\ &= \frac{1}{2}\delta\sum_{j\neq i}\left(x_j^0\right)^2 + \left(\sum_{j\neq i} x_j^0\right)^2\left[\frac{1}{2}\delta(1-\delta)^2 + \frac{1}{2}\delta^2(1-\delta)\right] \\ &= \frac{1}{2}\delta\sum_{j\neq i}\left(x_j^0\right)^2 + \frac{1}{2}\delta(1-\delta)\left(\sum_{j\neq i} x_j^0\right)^2. \end{aligned} \tag{3.17}$$

Since

$$\mathbb{E}_i \sum_{j\neq i} \left(x_j^0\right)^2 = \frac{n-1}{n} \sum_{j=1}^{n} \left(x_j^0\right)^2 = \frac{n-1}{n}\|x^0\|^2,$$

$$\begin{aligned}\mathbb{E}_i \left(\sum_{j\neq i} x_j^0\right)^2 &= \mathbb{E}_i(\mathbf{1}^{\mathrm{T}}x^0 - x_i)^2 \\ &= \mathbb{E}_i \left((\mathbf{1}^{\mathrm{T}}x^0)^2 - 2x_i(\mathbf{1}^{\mathrm{T}}x^0) + x_i^2\right) \\ &= (\mathbf{1}^{\mathrm{T}}x)^2 - \frac{2}{n}(\mathbf{1}^{\mathrm{T}}x^0)^2 + \frac{1}{n}\|x^0\|^2 \\ &= \left(1 - \frac{2}{n}\right)(\mathbf{1}^{\mathrm{T}}x^0)^2 + \frac{1}{n}\|x^0\|^2,\end{aligned}$$

we have by taking expectation with respect to i in (3.17) that the equality in (3.16) holds.

For the inequality in (3.16) we have from

$$f(x^0) = \frac{1}{2}(x^0)^{\mathrm{T}}Ax^0 = \frac{1}{2}\delta\|x^0\|^2 + \frac{1}{2}(1-\delta)(\mathbf{1}^{\mathrm{T}}x^0)^2 \geqslant \frac{1}{2}(1-\delta)(\mathbf{1}^{\mathrm{T}}x^0)^2$$

that

$$\begin{aligned}\mathbb{E}_i f(x^1) &= \frac{1}{2}\delta\left(1 - \frac{\delta}{n}\right)\|x^0\|^2 + \frac{1}{2}\delta(1-\delta)\left(1 - \frac{2}{n}\right)(\mathbf{1}^{\mathrm{T}}x^0)^2 \\ &\leqslant \frac{1}{2}\delta\|x^0\|^2 + \frac{1}{2}\delta(1-\delta)(\mathbf{1}^{\mathrm{T}}x^0)^2 \\ &\leqslant \frac{1}{2}\delta\|x^0\|^2 + \delta f(x^0),\end{aligned}$$

as required. □

For CCD we have from (3.17) with $i = 1$ that

$$\begin{aligned}f(x^1) &= \frac{1}{2}\delta(\|x^0\|^2 - (x_1^0)^2) + \frac{1}{2}\delta(1-\delta)\left((\mathbf{1}^{\mathrm{T}}x^0) - x_1^0\right)^2 \\ &\leqslant \frac{1}{2}\delta\|x^0\|^2 + \frac{1}{2}\delta(1-\delta)\left[(\mathbf{1}^{\mathrm{T}}x^0)^2 - 2(x_1^0)(\mathbf{1}^{\mathrm{T}}x^0)\right] \\ &\leqslant \frac{1}{2}\delta\left[\|x^0\|^2 + (\mathbf{1}^{\mathrm{T}}x^0)^2 + 2\|x^0\|(\mathbf{1}^{\mathrm{T}}x^0)\right] \\ &= \frac{1}{2}\delta\left[\|x^0\| + (\mathbf{1}^{\mathrm{T}}x^0)\right]^2.\end{aligned}$$

If x^0 is independent of δ we have $f(x^1) = \mathcal{O}(\delta)$. However, there is no guarantee that $f(x^1)$ is substantially smaller than $f(x^0)$. If x^0 is chosen 'adversarially' in such a way that $|\mathbf{1}^T x^0| \ll \|x^0\|$, we may find that $f(x^1)$ is not much smaller than $f(x^0)$. For random choices of x^0, however, we would expect a significant decrease on the first iteration, similar to that observed for RPCD and RCD.

4. Computational results

Some comparisons between empirical rates and rates predicted from the analysis are shown in Table 1, for $n = 100$ and different values of δ. For the empirical rates $\rho_{\mathrm{CCD}}(\delta, x^0)$, $\rho_{\mathrm{RCD}}(\delta, x^0)$ and $\rho_{\mathrm{RPCD}}(\delta, x^0)$ we used formulas like (2.5), but we took the average decrease factor *over only the last 10 epochs*, so as to capture the asymptotic rates and discount the early iterations. We used the termination criterion $f(x^{\ell n}) - f^* \leqslant 10^{-8}$. For the theoretical predictions we used $\rho(C)^2$ for CCD (as suggested by (2.7)), the formula $\rho_{\mathrm{RCD}}(\delta, \text{predicted}) = (1 - 2\delta/(n(1+\delta)))^n$ for RCD (from (2.20)) and $\rho(M)$ for RPCD (from (3.15)). We note from this table that the theoretical predictions for CCD and RPCD are quite sharp, even for values of δ that are not particularly small. For RCD the empirical results are slightly better than predicted by the theory when δ is large. RPCD has the best practical and theoretical asymptotic convergence of the three variants, with the advantage increasing as δ increases.

5. Conclusions

Recent work has shown that problem (1.11) with Hessian matrix (1.17) is a case that reveals significant differences in performance between cyclic and randomized variants of coordinate descent. Here we provide an analysis of the performance of RPCD that sharply predicts the practical convergence behavior of this approach, showing an asymptotic convergence rate that at least matches (and is even slightly better than) that obtained by a random sampling-with-replacement scheme.

Empirically, it appears that convex quadratic instances that reveal differences between CCD, RCD and RPCD are quite limited in scope, with (1.17) being the canonical instance and the one whose analysis is most tractable. In work subsequent to this paper (Wright & Lee, 2017) we analyse the case of quadratic convex f in which the Hessian has the form $\delta I + (1-\delta)uu^T$, where $u \in \mathbb{R}^n$ is a vector whose elements have magnitude not too different from 1. By a diagonal transformation this matrix has the form $\delta I + (1-\delta)\mathbf{1}\mathbf{1}^T + \varepsilon D$, where D is diagonal with elements in the range $[0, 1]$ and $\varepsilon \geqslant 0$. Our analysis in Wright & Lee (2017) builds on the approach in this paper but is somewhat more complex; the exact two-variable recurrence of Theorem 3.2 becomes an approximate recurrence involving more terms.

Acknowledgements

We thank two referees and the Editor-in-Chief for their comments on earlier drafts, which caused us to improve the presentation and sharpen the results of the paper.

Funding

National Science Foundation (DMS-1216318 and IIS-1447449); Office of Naval Research (N00014-13-1-0129); Air Force Office of Scientific Research (FA9550-13-1-0138); Argonne National Laboratory (Subcontracts 3F-30222 and 8F-30039).

References

Arjevani, Y., Shalev-Shwartz, S. & Shamir, O. (2016) On lower and upper bounds in smooth and strongly convex optimization. *J. Mach. Learn. Res.*, **17**, 1–51.

Beck, A. & Tetruashvili, L. (2013) On the convergence of block coordinate descent type methods. *SIAM J. Optim.*, **23**, 2037–2060.

Gelfand, I. (1941) Normierte ringe. *Rech. Math. [Mat. Sbornik]*, **9**, 3–24.

Golub, G. H. & Van Loan, C. F. (2012) *Matrix Computations*, 4th edn. Baltimore, MD: Johns Hopkins University Press.

Hackbusch, W. (2016) *Iterative Solution of Large Sparse Systems of Equations*, 2nd edn. Switzerland: Springer International Publishing.

Hoffman, A. J. (1952) On approximate solutions of systems of linear inequalities. *J. Res. Natl. Bur. Stand.*, **49**, 263–265.

Karimi, H., Nutini, J. & Schmidt, M. (2016) Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016*, Lecture Notes in Computer Science, vol. 9851. Cham: Springer.

Liu, J. & Wright, S. J. (2015) Asynchronous stochastic coordinate descent: parallelism and convergence properties. *SIAM J. Optim.*, **25**, 351–376.

Łojasiewicz, S. (1963) Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aus Dérivées Partielles.* Éditions du Centre National de la Recherche Scientifique. pp. 87–89 (French).

Luo, Z.-Q. & Tseng, P. (1992) On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, **72**, 7–35.

Luo, Z.-Q. & Tseng, P. (1993) Error bounds and convergence analysis of feasible descent methods: a general approach. *Ann. Oper. Res.*, **46**, 157–178.

Nesterov, Y. E. (2012) Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, **22**, 341–362.

Samarskii, A. A. & Nikolaev, E. S. (1989) *Numerical Methods for Grid Equations*, vol. II. Iterative Methods. Basel: Birkhäuser.

Sun, R. & Ye, Y. (2016) Worst-case complexity of cyclic coordinate descent: $O(n^2)$ gap with randomized version. *Technical Report*. Stanford, CA: Department of Management Science and Engineering, Stanford University. arXiv:1604.07130.

Wright, S. J. (2015a) Computations with coordinate descent methods. Presentation, *Workshop on Challenges in Optimization for Data Science*. Available at https://pcombet.math.ncsu.edu/data2015/.

Wright, S. J. (2015b) Coordinate descent algorithms. *Math. Program. B*, **151**, 3–34.

Wright, S. J. (2015c) Coordinate descent methods. Colloquium, *Courant Institute of Mathematical Sciences*. December, 2015.

Wright, S. J. & Lee, C.-P. (2017) Analyzing random permutations for cyclic coordinate descent. *Technical Report*. Madison, WI: Department of Computer Sciences, University of Wisconsin-Madison. arXiv:1706:00908.

Young, D. M. (1971). *Iterative Solution of Large Linear Systems*. Orlando, FL: Academic Press.

Appendix

A. Estimating terms in the recurrence matrix M

Here we first find upper and (in some cases) lower bounds for the following quantities, for the matrix A given in (1.17) and the corresponding value of C defined in (2.12) and (2.14):

$$(\mathbf{1}^{\mathrm{T}}C\mathbf{1})^2, \quad \|C\mathbf{1}\|^2, \quad \|C^{\mathrm{T}}\mathbf{1}\|^2, \quad \|C\|_F^2. \tag{A.1}$$

We then use these quantities to obtain bounds on d_1, d_2, m_1 and m_2 from (3.9). We assume throughout that $n \geqslant 10$ and $\delta \in [0, 0.4]$.

For $\mathbf{1}^T C\mathbf{1}$, we have from (2.12) and (2.14) that

$$\mathbf{1}^{\mathrm{T}} C\mathbf{1} = (1-\delta)(\mathbf{1}^{\mathrm{T}}\bar{L})(E^{\mathrm{T}}\mathbf{1}) = (1-\delta)u^{\mathrm{T}}v,$$

where $u = \bar{L}^T\mathbf{1}$ and $v = E^T\mathbf{1}$ have the following components:

$$v_i = n - i, \quad i = 1, 2, \ldots, n$$

(from (2.12)) and

$$u_i = -1 + (1-\delta)\sum_{t=0}^{n-i-1}\delta^t = -1 + (1-\delta)\frac{1-\delta^{n-i}}{1-\delta} = -\delta^{n-i}, \quad i = 1, \ldots, n$$

(from (2.14a)). For $\delta \in [0, 0.4]$, we have

$$\begin{aligned}
0 \geqslant \mathbf{1}^{\mathrm{T}} C\mathbf{1} &= -(1-\delta)\sum_{i=1}^{n}(n-i)\delta^{n-i} \\
&= -(1-\delta)\sum_{i=1}^{n-1} i\delta^i \\
&= -(1-\delta)\sum_{i=1}^{n-1}\sum_{j=i}^{n-1}\delta^j \\
&\geqslant -(1-\delta)\sum_{i=1}^{n-1}\frac{\delta^i}{1-\delta} \\
&= -\sum_{i=1}^{n-1}\delta^i \\
&\geqslant -\frac{\delta}{1-\delta} \geqslant -2\delta.
\end{aligned}$$

Therefore, we have

$$0 \leqslant (\mathbf{1}^{\mathrm{T}} C\mathbf{1})^2 \leqslant 4\delta^2. \tag{A.2}$$

We next seek an upper bound for $\|C^{\mathrm{T}}\mathbf{1}\|_2^2$. We have from (2.15) that

$$\begin{aligned}(C^{\mathrm{T}}\mathbf{1})_j &= \sum_{i=1}^{j-1} C_{ij} + \sum_{i=j}^{n} C_{ij} \\ &= -(1-\delta)\sum_{i=1}^{j-1}\delta^{i-1} + (1-\delta)\sum_{i=j}^{n}(\delta^{i-j} - \delta^{i-1}) \\ &= -(1-\delta)\sum_{i=1}^{n}\delta^{i-1} + (1-\delta)\sum_{t=0}^{n-j}\delta^{t} \\ &= -(1-\delta^n) + (1-\delta)\frac{1-\delta^{n-j+1}}{1-\delta} \\ &= \delta^n - \delta^{n-j+1}.\end{aligned}$$

It follows that

$$\begin{aligned}\|C^{\mathrm{T}}\mathbf{1}\|_2^2 &= \sum_{j=1}^{n}(\delta^n - \delta^{n-j+1})^2 \\ &\leqslant \sum_{j=1}^{n}(\delta^{n-j+1})^2 \\ &= \sum_{j=1}^{n}\delta^{2j} \\ &\leqslant \frac{\delta^2}{1-\delta^2} \\ &\leqslant 1.34\delta^2. \end{aligned} \tag{A.3}$$

We now use (2.15) to compute bounds on the other quantities in (A.1). We have

$$\begin{aligned}(C\mathbf{1})_i &= \sum_{j=1}^{i} C_{ij} + \sum_{j=i+1}^{n} C_{ij} \\ &= (1-\delta)\sum_{j=1}^{i}(\delta^{i-j} - \delta^{i-1}) - (1-\delta)\sum_{j=i+1}^{n}\delta^{i-1} \\ &= (1-\delta)\sum_{j=1}^{i}\delta^{i-j} - (1-\delta)n\delta^{i-1} \\ &= (1-\delta^i) - n(1-\delta)\delta^{i-1} \\ &= 1 - n\delta^{i-1} + (n-1)\delta^i.\end{aligned}$$

We thus obtain

$$\begin{aligned}\|C\mathbf{1}\|_2^2 &= \sum_{i=1}^{n}\left[1-2n\delta^{i-1}+2(n-1)\delta^i+n^2\delta^{2i-2}-2n(n-1)\delta^{2i-1}+(n-1)^2\delta^{2i}\right]\\ &= n+[-2n+2(n-1)\delta]\sum_{i=1}^{n}\delta^{i-1}+\left[n^2-2n(n-1)\delta+(n-1)^2\delta^2\right]\sum_{i=1}^{n}(\delta^2)^{i-1}\\ &= n+[-2n+2(n-1)\delta]\,\frac{1-\delta^n}{1-\delta}+\left[n^2-2n(n-1)\delta+(n-1)^2\delta^2\right]\frac{1-\delta^{2n}}{1-\delta^2}. \qquad \text{(A.4)}\end{aligned}$$

Noting that $[-2n+2(n-1)\delta]<0$ and $[n^2-2n(n-1)\delta+(n-1)^2\delta^2]>0$ for the values of δ and n of interest, and using $2n\delta^n(1-\delta^2)\leqslant(2\delta^8)n\delta^2\leqslant 0.01n\delta^2$, we continue:

$$\begin{aligned}\|C\mathbf{1}\|_2^2 &\leqslant n+[-2n+2(n-1)\delta]\,\frac{1}{1-\delta}+\delta^n[2n-2(n-1)\delta]+\left[n^2-2n(n-1)\delta+(n-1)^2\delta^2\right]\frac{1}{1-\delta^2}\\ &\leqslant \frac{1}{1-\delta^2}\left[n(1-\delta^2)+[-2n+2(n-1)\delta]\,(1+\delta)+0.01n\delta^2+n^2-2n(n-1)\delta+(n-1)^2\delta^2\right]\\ &= \frac{1}{1-\delta^2}\left[n^2-2n^2\delta+2n\delta-n-2\delta+\delta^2\left[-n+2(n-1)+0.01n+(n-1)^2\right]\right]\\ &\leqslant \frac{1}{1-\delta^2}\left[n(n-1)(1-2\delta)+n^2\delta^2\right].\end{aligned}$$

Thus, dividing by $n(n-1)$ and using $\delta\in[0,0.4]$ to deduce that $(1-\delta^2)^{-1}\leqslant 1+1.5\delta^2$, we obtain

$$\begin{aligned}\frac{\|C\mathbf{1}\|_2^2}{n(n-1)} &\leqslant \frac{1}{1-\delta^2}\left[(1-2\delta)+\frac{n}{n-1}\delta^2\right]\\ &\leqslant (1+1.5\delta^2)[(1-2\delta)+1.12\delta^2]\\ &\leqslant (1-2\delta)+(1.5+1.12)\delta^2+2\delta^4\\ &\leqslant (1-2\delta)+(1.5+1.12+0.5)\delta^2\\ &\leqslant (1-2\delta)+3.2\delta^2. \qquad \text{(A.5)}\end{aligned}$$

For the corresponding lower bound, we pick up from (A.4) and again use $[-2n+2(n-1)\delta]<0$ and $[n^2-2n(n-1)\delta+(n-1)^2\delta^2]>0$, together with $[n^2-2n(n-1)\delta+(n-1)^2\delta^2]\leqslant n^2(1+\delta^2)$ and

$\delta^{2n}(1+\delta^2) \leqslant 0.001\delta^2$, to obtain

$$\begin{aligned}
\|C\mathbf{1}\|_2^2 &\geqslant n + [-2n + 2(n-1)\delta]\,\frac{1}{1-\delta} + \left[n^2 - 2n(n-1)\delta + (n-1)^2\delta^2\right]\frac{1}{1-\delta^2} - \delta^{2n}n^2(1+\delta^2) \\
&= \frac{1}{1-\delta^2}\left[n(1-\delta^2) + [-2n + 2(n-1)\delta]\,(1+\delta) + \left[n^2 - 2n(n-1)\delta + (n-1)^2\delta^2\right]\right] - 0.001n^2\delta^2 \\
&= \frac{1}{1-\delta^2}\left[(n^2 - n) + [2(n-1) - 2n - 2n(n-1)]\,\delta + \left[-n + 2(n-1) + (n-1)^2\right]\delta^2\right] - 0.001n^2\delta^2 \\
&= \frac{1}{1-\delta^2}\left[(n^2-n) + (-2n^2 + 2n - 2)\delta + (n^2 - n - 1)\delta^2\right] - 0.001n^2\delta^2 \\
&\geqslant \frac{1}{1-\delta^2}\left[n(n-1) - 2n(n-1)\delta - 2\delta + n(n-1)\delta^2 - \delta^2 - 0.001n^2\delta^2\right].
\end{aligned}$$

Thus, dividing by $n(n-1)$, we obtain

$$\begin{aligned}
\frac{\|C\mathbf{1}\|_2^2}{n(n-1)} &\geqslant \frac{1}{1-\delta^2}\left[1 - 2\delta + \delta^2 - \frac{2\delta + \delta^2}{n(n-1)} - \frac{0.001n^2}{n(n-1)}\delta^2\right] \\
&\geqslant \frac{1}{1-\delta^2}\left[1 - 2\delta + \delta^2 - \frac{2.5\delta}{n(n-1)} - 0.002\delta^2\right] \\
&\geqslant 1 - 2\delta + 0.998\delta^2 - \frac{3\delta}{n^2}. \qquad \text{(A.6)}
\end{aligned}$$

Note that this lower bound is strictly positive in the regime $\delta \in [0, 0.4]$ and $n \geqslant 10$.

For $\|C\|_F^2$, we obtain from (2.15) that

$$\begin{aligned}
\frac{1}{(1-\delta)^2}\|C\|_F^2 &= \sum_{j=1}^{n}\left\{\sum_{i=1}^{j-1}\delta^{2i-2} + \sum_{i=j}^{n}\left(\delta^{2i-2j} - 2\delta^{2i-j-1} + \delta^{2i-2}\right)\right\} \\
&= \sum_{j=1}^{n}\left\{\frac{1-\delta^{2j-2}}{1-\delta^2} + \left(1 - 2\delta^{j-1} + \delta^{2j-2}\right)\sum_{i=0}^{n-j}\delta^{2i}\right\} \\
&= \sum_{j=1}^{n}\left\{\frac{1-\delta^{2j-2}}{1-\delta^2} + \left(1 - 2\delta^{j-1} + \delta^{2j-2}\right)\frac{1-\delta^{2n-2j+2}}{1-\delta^2}\right\} \qquad \text{(A.7)} \\
&\leqslant \frac{1}{1-\delta^2}\sum_{j=1}^{n}\left\{\left(1-\delta^{2j-2}\right) + \left(1 - 2\delta^{j-1} + \delta^{2j-2}\right)\right\} \\
&\leqslant \frac{1}{1-\delta^2}\left\{2n - 2\frac{1-\delta^n}{1-\delta}\right\},
\end{aligned}$$

so that

$$\begin{aligned}\|C\|_F^2 &\leqslant \frac{1-\delta}{1+\delta}\left\{2n - 2\frac{1-\delta^n}{1-\delta}\right\}\\ &\leqslant \frac{1}{1+\delta}\left\{2n - 2n\delta - 2 + 2\delta^n\right\}\\ &\leqslant (1-\delta+\delta^2)\left\{2n - 2n\delta - 2 + 0.01\delta^2\right\} \qquad \text{(A.8)}\\ &= 2(n-1) + [-2(n-1) - 2n]\delta + [2(n-1) + 2n + 0.01]\delta^2 - [2n + 0.01]\delta^3 + 0.01\delta^4\\ &\leqslant 2(n-1) + [-4n + 2]\delta + 4n\delta^2,\end{aligned}$$

where in (A.8) we used $2\delta^n \leqslant 2(\delta^8)\delta^2 \leqslant 0.01\delta^2$. It therefore follows that

$$\frac{\|C\|_F^2}{n-1} \leqslant 2 - 4\delta - \frac{2\delta}{n-1} + \frac{4n}{n-1}\delta^2 \leqslant 2 - 4\delta - \frac{2\delta}{n} + 4.5\delta^2. \qquad \text{(A.9)}$$

It follows, using again $\delta \in [0, 0.4]$ and $n \geqslant 10$, that

$$\begin{aligned}\frac{\|C\|_F^2}{n-1} &\leqslant 2 - 4\delta - \frac{2\delta}{n} + 4.5\delta^2\\ &\leqslant 2 - 4\delta - \frac{6\delta}{n^2} + 0.998n\delta^2\\ &\leqslant 2\left(1 - 2\delta - \frac{3\delta}{n^2}\right) + 0.998n\delta^2\\ &\leqslant n\left(1 - 2\delta - \frac{3\delta}{n^2} + 0.998\delta^2\right) \leqslant \frac{\|C\mathbf{1}\|_2^2}{n-1},\end{aligned}$$

where we used (A.6) for the final inequality. It follows that

$$\|C\mathbf{1}\|_2^2 - \|C\|_F^2 \geqslant 0. \qquad \text{(A.10)}$$

From formulas (3.9) together with (A.2), (A.5), (A.6), (A.3), (A.9) and (A.10), and using $n \geqslant 10$, we have

$$0 \leqslant d_2 = \frac{\|C\mathbf{1}\|_2^2 - \|C\|_F^2}{n(n-1)} \leqslant \frac{\|C\mathbf{1}\|_2^2}{n(n-1)} \leqslant 1 - 2\delta + 3.2\delta^2, \qquad \text{(A.11a)}$$

$$
\begin{aligned}
0 \leqslant d_1 &= \frac{\|C\|_F^2}{n-1} - \frac{\|C\mathbf{1}\|_2^2}{n(n-1)} \\
&\leqslant \left(2 - 4\delta - \frac{2\delta}{n} + 4.5\delta^2\right) - \left(1 - 2\delta + 0.998\delta^2 - \frac{3\delta}{n^2}\right) \\
&\leqslant 1 - 2\delta - \frac{2\delta}{n} + \frac{3\delta}{n^2} + 3.6\delta^2 \\
&\leqslant 1 - 2\delta - \frac{\delta}{n}(2 - 3/n) + 3.6\delta^2 \\
&\leqslant 1 - 2\delta + 3.6\delta^2, \qquad \text{(A.11b)}
\end{aligned}
$$

$$
\begin{aligned}
|m_2| &= \left| \frac{(\mathbf{1}^{\mathrm{T}} C \mathbf{1})^2 - \|C^{\mathrm{T}}\mathbf{1}\|_2^2}{n(n-1)} \right| \\
&\leqslant \frac{\max\left((\mathbf{1}^{\mathrm{T}} C \mathbf{1})^2, \|C^{\mathrm{T}}\mathbf{1}\|^2\right)}{n(n-1)} \\
&\leqslant \frac{4\delta^2}{n(n-1)} \leqslant 0.05\delta^2, \qquad \text{(A.11c)}
\end{aligned}
$$

$$
\begin{aligned}
|m_1| &= \left| \frac{\|C^{\mathrm{T}}\mathbf{1}\|_2^2}{n-1} - \frac{(\mathbf{1}^{\mathrm{T}} C \mathbf{1})^2}{n(n-1)} \right| \\
&\leqslant \max\left(\frac{\|C^{\mathrm{T}}\mathbf{1}\|_2^2}{n-1}, \frac{(\mathbf{1}^{\mathrm{T}} C \mathbf{1})^2}{n(n-1)} \right) \qquad \text{(A.11d)} \\
&\leqslant \max\left(\frac{1.34}{9}, \frac{4}{90} \right) \delta^2 \leqslant 0.15\delta^2. \qquad \text{(A.11e)}
\end{aligned}
$$

B. Approximation of d_1, d_2, m_1 and m_2 for estimating $\rho(M)$

From (A.11c), (A.11d), (A.2) and (A.3), we have

$$
m_1 = \mathcal{O}\left(\frac{\delta^2}{n}\right), \quad m_2 = \mathcal{O}\left(\frac{\delta^2}{n^2}\right). \qquad \text{(B.1)}
$$

For the two terms d_1 and d_2, we first need better approximations of $\|C\mathbf{1}\|_2^2$ and $\|C\|_F^2$. From (A.4), we proceed with

$$
\begin{aligned}
\|C\mathbf{1}\|_2^2 &= n + (1 + \delta + \delta^2)\left[-2n + 2(n-1)\delta\right] + \left[n^2 - 2n(n-1)\delta + (n-1)^2\delta^2\right](1 + \delta^2) + \mathcal{O}\left(n^2\delta^3\right) \\
&= n(n-1) + \delta\left(-2n^2 + 2n - 2\right) + \delta^2\left(2n^2 - 2n - 1\right) + \mathcal{O}\left(n^2\delta^3\right). \qquad \text{(B.2)}
\end{aligned}
$$

For $\|C\|_F^2$, we obtain from (A.7) that

$$\begin{aligned}\frac{1}{(1-\delta)^2}\|C\|_F^2 &= \frac{1}{1-\delta^2}\sum_{j=1}^{n}\left\{2-(\delta^2)^{j-1}-2\delta^{j-1}+\delta^{2j-2}-\delta^{2n-2j+2}+2\delta^{2n-j+1}-\delta^{2n}\right\}\\ &= \frac{1}{1-\delta^2}\left\{2n-(1+\delta^2)-2(1+\delta+\delta^2)+(1+\delta^2)-\delta^2+\mathcal{O}(\delta^3)\right\}\\ &= \frac{1}{1-\delta^2}\left\{2n-2-2\delta-3\delta^2\right\}+\mathcal{O}(\delta^3),\end{aligned}$$

so that

$$\begin{aligned}\|C\|_F^2 &= \frac{1-\delta}{1+\delta}(2n-2-2\delta-3\delta^2)+\mathcal{O}(\delta^3)\\ &= (1-\delta)(1-\delta+\delta^2)(2n-2-2\delta-3\delta^2)+\mathcal{O}(n\delta^3)\\ &= (1-2\delta+2\delta^2)(2n-2-2\delta-3\delta^2)+\mathcal{O}(n\delta^3)\\ &= (2n-2)-\delta(4n-2)+\delta^2(4n-3)+\mathcal{O}(n\delta^3).\end{aligned} \tag{B.3}$$

We then have from (B.3) and (B.2) that

$$d_2 = \frac{\|C\mathbf{1}\|_2^2-\|C\|_F^2}{n(n-1)} = \frac{n(n-1)+\mathcal{O}\left(n^2\delta\right)-(2n-2)}{n(n-1)} = 1-\frac{2}{n}+\mathcal{O}\left(\delta\right), \tag{B.4a}$$

$$\begin{aligned}d_1 &= \frac{\|C\|_F^2}{n-1}-\frac{\|C\mathbf{1}\|_2^2}{n(n-1)}\\ &= 2-4\delta+4\delta^2+\frac{-2\delta}{n-1}-\left(1-2\delta+\frac{-2\delta}{n(n-1)}+2\delta^2\right)+\mathcal{O}\left(\frac{\delta^2}{n}\right)+\mathcal{O}(\delta^3)\\ &= 1-2\delta-\frac{2\delta}{n}+2\delta^2+\mathcal{O}\left(\frac{\delta^2}{n}\right)+\mathcal{O}(\delta^3).\end{aligned} \tag{B.4b}$$

C. Condition (1.9) for $g(Ex)$ with g strongly convex

Suppose that $f(x) = g(Ex)$, where g is strongly convex with modulus of convexity $\sigma > 0$, and $E \in \mathbb{R}^{m\times n}$. If $E = 0$, all x are optimal, so the claim (1.9) holds trivially. Otherwise, we have that $\sigma_{\min,\mathrm{nz}}$, the minimum nonzero singular value of E, is strictly positive.

By strong convexity of g, there exists a unique $t^* \in \mathbb{R}^m$ such that the solution set for (1.1) has the form $\{x \,|\, Ex = t^*\}$. Let $P(x)$ denote the projection of any vector $x \in \mathbb{R}^n$ onto this set. We have by Hoffman's lemma (Hoffman, 1952) that

$$\left\|x-P(x)\right\| \leqslant \sigma_{\min,\mathrm{nz}}^{-1}\left\|E\big(x-P(x)\big)\right\| = \sigma_{\min,\mathrm{nz}}^{-1}\|Ex-t^*\|.$$

Thus, by strong convexity, we have

$$f(x) = g(Ex) \geqslant g(t^*)+\frac{\sigma}{2}\left\|E(x-P(x))\right\|^2 \geqslant f^*+\frac{\sigma\sigma_{\min,\mathrm{nz}}^2}{2}\left\|x-P(x)\right\|^2. \tag{C.1}$$

Meanwhile we have by convexity of f that

$$f^* \geqslant f(x) + \nabla f(x)^{\mathrm{T}}\big(P(x) - x\big),$$

so that

$$f(x) - f^* \leqslant \big\|\nabla f(x)\big\|\,\big\|P(x) - x\big\| \leqslant \big\|\nabla f(x)\big\|\left(\frac{2}{\sigma\sigma_{\min,\mathrm{nz}}^2}\right)^{1/2}\big(f(x) - f^*\big)^{1/2}.$$

Dividing both sides by $(f(x) - f^*)^{1/2}$ we obtain

$$\big\|\nabla f(x)\big\|\left(\frac{2}{\sigma\sigma_{\min,\mathrm{nz}}^2}\right)^{1/2} \geqslant \big(f(x) - f^*\big)^{1/2} \;\Rightarrow \big\|\nabla f(x)\big\|^2 \geqslant \frac{\sigma\sigma_{\min,\mathrm{nz}}^2}{2}\big(f(x) - f^*\big),$$

which has the form (1.9).