RANDOM PERMUTATIONS FIX A WORST CASE FOR CYCLIC COORDINATE DESCENT∗

CHING-PEI LEE† AND STEPHEN J. WRIGHT‡

Abstract. 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 family of quadratic functions for which CCD is significantly slower than for RCD; a recent paper of Sun and Ye has explored the poor behavior of CCD on this family. The RPCD approach performs well on this family, and this paper explains this good behavior with a tight analysis.

Key words. Coordinate descent, randomization, permutations

AMS subject classifications. 65F10, 90C25, 68W20

1. Introduction. We consider coordinate descent (CD) algorithms applied to the following problem:

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

The basic (component-wise) coordinate descent framework is shown in Algorithm 1. Here, we denote

$$\nabla_i f(x) = [\nabla f(x)]_i, \quad e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T,$$

where the single nonzero in $e_i$ appears in position $i$. Each outer cycle (indicated by index $\ell$) is called an “epoch,” with each epoch consisting of $n$ iterations (indexed by $j$). The counter $k$ 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:

- CCD (Cyclic CD): $i(\ell, j) = j$.
- RCD (Randomized CD, also known as Stochastic CD): $i(\ell, j)$ is chosen uniformly at random from $\{1, 2, \ldots, n\}$ — a sampling-with-replacement approach.
- RPCD (Random-Permutations Cyclic CD): At the start of epoch $\ell$, we choose a random permutation of $\{1, 2, \ldots, n\}$, denoted by $\pi_\ell$. Index $i(\ell, j)$ is chosen to be the $j$th entry in $\pi_\ell$. This approach represents sampling without replacement, within each cycle.
Algorithm 1 Coordinate Descent

Set $x^1 \in \mathbb{R}^n$;
for $\ell = 0, 1, 2, \ldots$ do
  for $j = 1, 2, \ldots, n$ 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

(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 $|\nabla_i f(x^k)|$.)

1.1. Linear Convergence Results for CD. 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 the paper [14] for a discussion of the state of the art in early 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

$$f(x^{(\ell+1)n}) - f^* \leq \rho(f(x^{\ell n}) - f^*)$$

for some $\rho \in (0, 1)$, where $f^*$ is the optimal value of (1). When (3) holds, a reduction in function error by a factor of $\epsilon$ can be attained in at most $|\log \epsilon|/(1 - \rho)$ iterations. Thus we sometimes refer to $1/(1 - \rho)$ as the “complexity” of an algorithmic variant.

We also focus on the case of $f$ convex quadratic, that is,

$$f(x) = \frac{1}{2} x^T A x,$$

where $A$ is symmetric positive semidefinite.

We preface the 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

$$|\nabla_i f(x + t e_i) - \nabla_i f(x)| \leq L_i |t|, \quad \text{for all } x \in \mathbb{R}^n \text{ and } t \in \mathbb{R}.$$

We have

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

The standard Lipschitz constant $L$ is defined so that

$$\|\nabla f(x + d) - \nabla f(x)\| \leq L \|d\|, \quad \text{for all } x, d \in \mathbb{R}^n.$$

(Here and throughout, we use $\| \cdot \|$ to denote $\| \cdot \|_2$.) For reasonable choices of the constants in (5), (6), and (7), the following bounds are satisfied:

$$1 \leq \frac{L}{L_{\text{max}}} \leq n.$$
These can be illustrated in the case of \( f \) convex quadratic (4), for which we can set
\[
\mu = \lambda_{\text{min, nz}}(A), \quad L_i = A_{ii}, \quad i = 1, 2, \ldots, n; \quad L_{\text{max}} = \max_{i=1,2,\ldots,n} A_{ii}, \quad L = \|A\|_2,
\]
where \( \lambda_{\text{min, nz}}(\cdot) \) denotes the minimum nonzero eigenvalue. For such functions, the upper bound is achieved in (8) by \( A = 11^T \) (where \( 1 = (1, 1, \ldots, 1)^T \)), for which \( L_i = 1, i = 1, 2, \ldots, n, L_{\text{max}} = 1, \) and \( L = n \).

The following property of Lojasiewicz [7] is useful in proving linear convergence:
\[
\|\nabla f(x)\|_2^2 \geq 2\mu[f(x) - f^*], \quad \text{for some } \mu > 0.
\]
This property holds for \( f \) strongly convex (with modulus of strong convexity \( \mu \)), and for the case in which \( f \) grows quadratically with distance from a non-unique minimizing set, as in the “optimal strong convexity” condition of [6, (1.2)]. It also holds generically for convex quadratic programs, even when the Hessians are singular. Further, condition (10) holds for the functional form considered by Luo and Tseng [8, 9], which is
\[
f(x) = g(Ex), \quad \text{where } E \in \mathbb{R}^{m \times n} \text{ and } g : \mathbb{R}^m \rightarrow \mathbb{R} \text{ strongly convex},
\]
without any conditions on \( E \). (For a proof, see Appendix B.) Property (10) is called the “Polyak-Lojasiewicz condition” in [5].

Luo and Tseng [8] prove linear convergence for a function of the form (11), where they require \( E \) to have no zero columns. They obtain expressions for the constant \( \rho \) in (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 [9], the same authors analyze a family of methods (including CD) for more general functions that satisfy a local error bound of the form \( \|x - P(x)\| \leq \chi \|\nabla f(x)\| \) holds (where \( P(x) \) is the projection of \( x \) onto the solution set of (1) and \( \chi \) is some constant). Again, their analysis is not clear about how the constant \( \rho \) of (3) depends on the properties of \( f \).

A family of linear convergence results is proved in [1, Theorem 3.9] for the case in which \( f \) is strongly convex (immediately extendable to the case in which \( f \) satisfies the condition (10)). For constant stepsizes \( \alpha_k \equiv \alpha \leq (1/L_{\text{max}}) \), convergence of the form (3) holds with
\[
\rho = 1 - \frac{\mu}{(2\alpha)(1 + nL_{\text{max}}^2\alpha^2)}.
\]
In particular, for \( \alpha = 1/L \), we have \( \rho = 1 - \mu/(2L(n + 1)) \). The rate constant \( \rho \) is optimized by \( \alpha = 1/(\sqrt{n}L) \), for which \( \rho = 1 - \mu/(\sqrt{n}L) \). For the case in which \( f \) is a convex quadratic (4) 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), [1, (3.23)] show that \( \rho = 1 - \mu/(2L_{\text{max}}(1 + n^2L^2/\mu^2)) \) in expression (3). Paradoxically, as noted by Sun and Ye [12], 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 [12] to
\[
\rho = 1 - \max \left\{ \frac{\mu L_{\text{min}}}{nLL_{\text{avg}}}, \frac{\mu L_{\text{min}}}{L^2(2 + \log n/\pi)^2}, \frac{\mu L_{\text{min}}}{n^2L_{\text{avg}}^2} \right\},
\]
For the random-permutations cyclic version RPCD, the convergence theory in [1] 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.

Turning to the sampling-with-replacement variant RCD, it follows from the convergence theory of Nesterov [10] that (3) holds in expectation over the i.i.d. uniformly random choices of indices $i(\ell, j)$ with

\begin{equation}
\rho \leq \left(1 - \frac{\mu}{nL_{\text{max}}}\right)^n \approx 1 - \frac{\mu}{L_{\text{max}}}.
\end{equation}

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, include empirical-risk-minimization functions that arise in data analysis, the computational cost of one steepest-descent step is comparable to the cost of one epoch of Algorithm 1 (see [14]). Standard analysis of steepest descent shows that fixed-steplength variants applied to functions satisfying (10) have linear convergence of the form (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 $L_{\text{max}}$. These bounds suggest that CCD may tend to track the performance of SD, while RCD could be significantly better is the ratio $L/L_{\text{max}}$ is large, that is, toward the upper end of its range in (8). The phenomenon of large values of $L/L_{\text{max}}$ is captured well by convex quadratic problems (4) in which the Hessian $A$ has a large contribution from $11^T$. Such matrices were used in computations by one of the authors in mid-2015 (see [13]; reported briefly in [14]). These tests observed that on such matrices, the RCD was indeed much faster than CCD (and also SD). It was observed also that the performance of RPCD was as fast as that of RCD; it did not track CCD as the existing worst-case analysis would suggest. Later work, reported in [15], identified the matrix

\begin{equation}
A := \delta I + (1 - \delta) 11^T,
\end{equation}

(where $1 = (1, 1, \ldots, 1)^T$) as being the archetype of a problem with large ratio $L/L_{\text{max}}$. This matrix has one dominant eigenvalue $\delta + n(1 - \delta)$ with eigenvector $1$, with the other $(n - 1)$ eigenvalues equal to $\delta$. (This matrix also has the property, vital to the analysis, that $P^TAP = A$ for all permutation matrices $P$.) In [15], the RPCD variant was shown to be significantly superior to the CCD. Independently, [12] studied this same matrix (15), using analysis to explain the practical advantage of RCD over CCD, and showing that the performance of CCD approaches its worst-case theoretical bound. Our focus in this paper is to analyze the performance of RPCD on the same matrix.

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 are almost always involve a sampling-without-replacement scheme. The reasons are clear: Convergence analysis is much easier for sampling with replacement, and implementations are more efficient (involving less data movement and better use of hierarchical memory structures) for sampling without replacement. 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 a randomized scheme over deterministic scheme.

Our analysis also suggests superior performance of sampling-without-replacement (RPCD) over sampling-with-replacement (RCD) on this problem. Although the implementations do not show a large difference between these two randomized variants, the results suggest that some tightening of the RCD analysis may be possible.

2. Computations on Convex Quadratics. We consider the application of CCD and RPCD to the convex quadratic problem (4). This problem has solution $x^* = 0$ with optimal objective $f^* = 0$. We assume that the matrix $A$ is diagonally scaled so that

$$A_{ii} = 1, \text{ for } i = 1, 2, \ldots, n. \tag{16}$$

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

$$x^{k+1} = x^k - \frac{1}{A_{ii}} (Ax^k)_i e_i = x^k - (Ax^k)_i e_i, \text{ with } k = \ell n + j \text{ and } i = i(\ell, j). \tag{17}$$

Some variants of CD methods applied (4) 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 follows:

$$x^{(\ell+1)n} = -(L + D)^{-1}(L^T x) = C x^{\ell n}, \text{ where } C := -(L + D)^{-1}L^T. \tag{18}$$

By applying the formula (18) recursively, we obtain

$$x^{\ell n}_{\text{CCD}} = C^\ell x^0, \quad f(x^{\ell n}_{\text{CCD}}) = \frac{1}{2} (x^0)^T C^\ell C x^0. \tag{19}$$

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

$$\rho_{\text{CCD}}(A, x^0) := \lim_{\ell \to \infty} \left( f(x^{\ell n}_{\text{CCD}})/f(x^0) \right)^{1/\ell}. \tag{20}$$

To obtain a bound on this quantity, we denote the eigenvalues of $C$ by $\gamma_i$, $i = 1, 2, \ldots, n$, and recall that the spectral radius $\rho(C)$ is $\max_{i=1, 2, \ldots, n} |\gamma_i|$. Since $A$ is positive definite, we have $\rho(C) < 1$ [3, Theorem 11.2.3]. We have from Gelfand’s formula [2] that

$$\rho(C) = \lim_{\ell \to \infty} \|C^\ell\|^{1/\ell}, \tag{21}$$
We can obtain a bound on \( \rho_{\text{CCD}}(A, x^0) \) in terms of \( \rho(C) \) as follows:

\[
\rho_{\text{CCD}}(A, x^0) := \lim_{\ell \to \infty} \left( \left( \frac{f(x_{n\ell}^0)}{f(x^0)} \right)^{1/\ell} \right) \\
= \lim_{\ell \to \infty} \left( x_0^T (C^\ell)^T A C^\ell x_0 / x_0^T A x_0 \right)^{1/\ell} \\
= \lim_{\ell \to \infty} \left( \|A^{1/2} C^\ell x_0\|_2^2 / \|A^{1/2} x_0\|_2^2 \right)^{1/\ell} \\
= \lim_{\ell \to \infty} \left( \|(A^{1/2} C^\ell A^{-1/2})(A^{1/2} x_0)\|_2^2 / \|A^{1/2} x_0\|_2^2 \right)^{1/\ell} \\
\leq \lim_{\ell \to \infty} \left( \|A^{1/2} C^\ell A^{-1/2}\|_2^2 \right)^{1/\ell} \\
\leq \lim_{\ell \to \infty} \text{cond}(A)^{1/\ell} \|C^\ell\|_2^{2/\ell} = \rho(C)^2.
\]

We can write RPCD by using a permutation matrix \( P_\ell \) to represent the permutation \( \pi_\ell \) on epoch \( \ell \). We split the matrix \( P_\ell^T A P_\ell \), and define the operator \( C_\ell \) as follows:

\[
P_\ell^T A P_\ell = L_\ell + D_\ell + L_\ell^T, \quad C_\ell := -(L_\ell + D_\ell)^{-1} L_\ell^T.
\]

After \( \ell \) cycles, we have

\[
x_{n\ell,\text{RPCD}} = P_\ell C_\ell P_\ell^T P_{\ell-1} C_{\ell-1} P_{\ell-2}^T \ldots P_1 C_1 P_1^T x^0,
\]

yielding a function value of

\[
f(x_{n,\text{RPCD}}) = \frac{1}{2} (x^0)^T (P_1 C_1^T P_1^T \ldots P_\ell C_\ell^T P_\ell^T A P_\ell C_\ell P_\ell^T \ldots P_1 C_1 P_1^T) x^0.
\]

If we could take the expected value of this quantity over all random permutations \( P_1, P_2, \ldots, P_\ell \), and over the initial point \( x^0 \), 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, this expectation becomes tractable for \( A \) defined by \((15)\)). We can however take the expectation with respect to \( x^0 \) only. Since the elements of \( x^0 \) are distributed according to \( N(0, 1) \), we have

\[
\mathbb{E}_{x^0} f(x_{n,\text{RPCD}}) = \frac{1}{2} \text{trace} \left( P_1 C_1^T P_1^T \ldots P_\ell C_\ell^T P_\ell^T A P_\ell C_\ell P_\ell^T \ldots P_1 C_1 P_1^T \right).
\]

Figure 1 shows relative 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, \( \mathbb{E}_{x^0} f(x_{n,\text{RPCD}}) / \mathbb{E}_{x^0} f(x^0) \) (see \((26)\) for \( \mathbb{E}_{x^0} f(x_{n,\text{CCD}}) \); similar formulas apply for \( \mathbb{E}_{x^0} f(x_{n,\text{RPCD}}) \) 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.

**2.2. Permutation-Invariant \( A \).** In trying to find the simplest instance of a matrix \( A \) for which the superiority of randomization is observed, we identified the case in which \( A \) is invariant under all symmetric permutations. This is exactly the case in which \( A \) has the form \((15)\). 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 on \( \delta \) ensures that \( A \) has the following properties:
Fig. 1: CCD and RPCD on convex quadratic objective, for log-uniform eigenvalue distribution.

- 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, $P^T AP = A$ for any $n \times n$ permutation matrix $P$;
- $L/L_{\text{max}}$ achieves its maximum value of $n$, opening a wide gap between the worst-case theoretical behavior of CCD and RCD.

Figure 2 shows results for CCD and RPCD, on a matrix $A$ of the permutation-invariant form (15), with $\delta = .05$. The leading eigenvalues are $95.05, .05, \ldots, .05$ and its leading eigenvector is $1 = (1, 1, \ldots, 1)^T$. As in Figure 1, we plot the expected relative function value on the vertical axis. Note in this case the clear superiority of RPCD over CD. Figure 3 shows results for the CCD, RPCD, and RCD variants on the matrix $A$ from (15) with $n = 100$ and $\delta = .05$. 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 $N(0, 1)$. The similarity between Figures 2 and 3 suggests that the practical performance of these methods can be captured accurately by taking expectations over the initial point $x^0$.

We now derive expressions for the epoch iteration matrix $C$ of Section 2.1 for the specific case of the permutation-invariant matrix (15). This is needed for the analysis of RPCD on this matrix.
By applying the splitting (18) to (15), we have

\begin{equation}
D = I, \quad L = (1 - \delta)E,
\end{equation}

where

\[
E = \begin{bmatrix}
0 & 0 & 0 & \ldots & 0 & 0 \\
1 & 0 & 0 & \ldots & 0 & 0 \\
1 & 1 & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 & 1 & 1 & \ldots & 1 & 0
\end{bmatrix}.
\]

Thus, defining

\begin{equation}
\bar{L} := -(L + D)^{-1},
\end{equation}

we have

\begin{equation}
\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}
\end{equation}

\begin{equation}
C = (1 - \delta)\bar{L}E^T.
\end{equation}
Fig. 3: CCD, RPCD, and RCD on convex quadratic objective, where \( A \) is the matrix (15) with \( n = 100 \) and \( \delta = .05 \).

Writing \( \bar{L} \) explicitly, we have

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

We have from (29b) 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

$$C_{ij} = (1 - \delta) \sum_{\ell=1}^{i-1} L_{i\ell}$$

$$= (1 - \delta) [(1 - \delta)(\delta^{i-2} + \delta^{i-3} + \ldots + \delta + 1) - 1]$$

$$= (1 - \delta) \left[ (1 - \delta) \frac{1 - \delta^{i-1}}{1 - \delta} - 1 \right]$$

$$= -(1 - \delta)\delta^{i-1}.$$

For the complementary case $i \geq j$, we have

$$C_{ij} = (1 - \delta) \sum_{\ell=1}^{j-1} L_{i\ell}$$

$$= (1 - \delta) [(1 - \delta)(\delta^{j-2} + \delta^{j-3} + \ldots + 1) - 1]$$

$$= (1 - \delta)^2 \delta^{i-j} \frac{1 - \delta^{i-1}}{1 - \delta}$$

$$= (1 - \delta)\delta^{i-j}(1 - \delta^{i-1})$$

$$= (1 - \delta)(\delta^{i-j} - \delta^{i-1}).$$

To summarize, we have

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

2.3. Convergence Rates of CCD and RCD. Here, we examine the theoretical convergence rate of CCD on the quadratic function with Hessian (15) by using the results of Sun and Ye [12].

Recalling the rate (13) from [12, Proposition 3.1], and substituting the following quantities for (15):

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

we find that

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

(We use $\rho_{\text{CCD}}(\delta, x^0)$ in place of $\rho_{\text{CCD}}(A, x^0)$, to emphasize the dependence of $A$ in (15) on the parameter $\delta$.) By making the mild assumption that $\delta \leq 3/4$, this expression simplifies to

$$\rho_{\text{CCD}}(\delta, x^0) \leq 1 - \frac{\delta}{n(1 - \delta) + \delta}.$$

On the other hand, Sun and Ye show the following lower bound on $\rho_{\text{CCD}}(\delta, x^0)$ (obtained by substituting from (31) into [12, Theorem 3.1]):

$$\rho_{\text{CCD}}(\delta, x^0) \geq \left( 1 - \frac{2\delta \pi^2}{n(n(1 - \delta) + \delta)} \right)^2.$$
RANDOM PERMUTATIONS FIX A WORST CASE FOR CYCLIC COORDINATE DESCENT

Fig. 4: Convergence of $f$ for CCD and RPCD applied to (4), (15), with $\delta = .001$ and $n = 10, 20, 40, 80$. Convergence rate of CCD deteriorates as $n$ grows, as predicted by the estimates, while the convergence rate of RPCD is independent of $n$.

By combining (32) and (33), we see that for small values of $\delta/n$, the average epoch-wise decrease in error is $\rho_{\text{CCD}}(\delta, x^0) = 1 - c\delta/n^2$, 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 [11], [16, p. 464], and [4, Theorem 3.44]. This dependency on $n$ is further confirmed in the numerical result of running CCD for $A$ with the same $\delta$ but different $n$ shown in Figure 4a.

For RCD, we have by substituting the values in (31) into (14) that the expected per-epoch improvement in error is given by

$$\rho_{\text{RCD}}(\delta, \text{predicted}) := \left(1 - \frac{\mu}{nL_{\text{max}}}\right)^n = \left(1 - \frac{\delta}{n}\right)^n = 1 - \delta + \delta^2 n - \frac{1}{2n} + O(\delta^3).$$

Comparison with the rate obtained for CCD shows that RCD has $O(n^2)$ time better complexity, and unlike CCD, the rate does not depend strongly on $n$.

Figure 4b 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.

2.4. Convergence Rate of RPCD. Since $A$ in (15) is invariant under symmetric permutations, the matrices $L$ and $D$ are the same for all $P^TAP$, where $P$ is any permutation matrix. When considering RPCD applied to this problem, we have in the notation of (23) that $C_\ell \equiv C$ for all $\ell$. The expression (24) simplifies as follows:

$$x^{\ell n}_{\text{RPCD}} = P_\ell C P^T_\ell P_{\ell-1} C P^T_{\ell-1} \cdots P_1 C P^T_1 x^0.$$

The function values are

$$f(x^{\ell n}_{\text{RPCD}}) = \frac{1}{2} (x^0)^T (P_1 C P^T_{\ell} \cdots P_1 C P^T_{\ell} A P_{\ell} C P^T_{\ell} \cdots P_1 C P^T_1) x^0.$$
We now analyze the expected value of the function (36) obtained after \( \ell \) epochs of RPCD, where \( A \) has the form (15). 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 follows:

\[
\bar{A}(t) = \mathbb{E}_{P_\ell-t+1, \ldots, P_\ell} \left( P_\ell-t+1 C^T P_{\ell-t+1}^T \cdots P_t C^T P_t^T A P_t C P_t^T \cdots P_{\ell-t+1} C P_{\ell-t+1}^T \right),
\]

and note that \( \bar{A}(0) = A \) and (by comparison with (36)) that

\[
\mathbb{E} f(x^{\ell}_{\text{RPCD}}) = \frac{1}{2} \mathbb{E}_{x^0} \left[ (x^0)^T \bar{A}(\ell) x^0 \right].
\]

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

\[
\bar{A}(t) = \mathbb{E}_{P_\ell} \left( P_{\ell-t+1} C^T P_{\ell-t+1}^T \bar{A}(t-1) P_{\ell-t+1} C P_{\ell-t+1}^T \right)
= \mathbb{E}_{P} \left( P C^T P^T \bar{A}(t-1) P C P^T \right).
\]

(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 \( \eta_t I + \nu_t 11^T \), for some positive coefficients \( \eta_t \) and \( \nu_t \). Note that for \( \bar{A}(t) \) of this form, we have that \( P^T \bar{A}(t) P = \bar{A}(t) \), a property that is crucial to our analysis. We derive a stationary iteration between the successive pairs \((\eta_{t-1}, \nu_{t-1})\) and \((\eta_t, \nu_t)\), and reveal the convergence properties of RPCD by analyzing the \( 2 \times 2 \) matrix that relates successive pairs.

We start with a technical lemma.

**Lemma 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 \( \Pi \), we have

\[
B := \mathbb{E}_P [P Q P^T] = \tau_1 I + \tau_2 11^T,
\]

where

\[
\tau_2 = \frac{1^T Q 1 - \text{trace}(Q)}{n(n-1)}, \quad \tau_1 = \frac{\text{trace}(Q)}{n} - \tau_2.
\]

**Proof.** For any \( P \in \Pi \), if \( P \) shifts the \( i \)th position to the \( j \)th position, then \( (P Q P^T)_{jj} = Q_{ii} \). Since the probability of taking any permutation from \( \Pi \) is identical, we have that

\[
\mathcal{P}((P Q P^T)_{jj} = Q_{ii}) = \frac{1}{n}, \quad \forall i, j \in \{1, \ldots, n\}
\]

(where \( \mathcal{P}(\cdot) \) 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 \( i \)th and the \( j \)th entries to the \( k \)th and the \( l \)th positions, respectively, that is,

\[
(P Q P^T)_{kl} = Q_{ij}.
\]
Note that we always have that \( i \neq j \Rightarrow k \neq l \) because permutations are bijections from and to \( \{1, \ldots, n\} \). Thus, there are \( (n-2)! \) permutations in \( \Pi \) with the property \((41)\). Under the same reasoning as before, each off-diagonal entry of \( B \) is the average of all off-diagonal entries of \( Q \).

Finally, we obtain \((40)\) by noting that \( B_{ii} = \tau_1 + \tau_2 \), while \( B_{ij} = \tau_2 \) for \( i \neq j \). \( \square \)

We have immediately from Lemma 1 that

\[
\mathbb{E}_P(P^T C^T C P) = d_1 I + d_2 \mathbf{1}\mathbf{1}^T, \quad \mathbb{E}_P(P^T C^T \mathbf{1}\mathbf{1}^T C P) = m_1 I + m_2 \mathbf{1}\mathbf{1}^T,
\]

where

\[
\begin{align*}
  d_2 &= \frac{1^T C^T C 1 - \text{trace}(C^T C)}{n(n-1)} = \frac{\|C\|_F^2 - \|C\|_2^2}{n(n-1)} \\
  d_1 &= \frac{\text{trace}(C^T C)}{n} - d_2 = \frac{\|C\|_F^2}{n} - d_2 \\
  m_2 &= \frac{(1^T C 1)^2 - (1^T C)(C^T 1)}{n(n-1)} = \frac{(1^T C 1)^2 - \|C^T 1\|_2^2}{n(n-1)} \\
  m_1 &= \frac{(1^T C)(C^T 1) - m_2}{n} = \frac{\|C^T 1\|_2^2}{n} - m_2
\end{align*}
\]

Note that for \((43c)\) and \((43d)\) we used the property \( \text{trace}(AB) = \text{trace}(BA) \).

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

**Theorem 2.** Consider solving \((4)\) with the matrix \( A \) defined in \((15)\) using RPCD. For \( \tilde{A}^{(t)} \) defined in \((38)\), with \( \tilde{A}^{(0)} = A \), we have

\[
\tilde{A}^{(t)} = \eta_t I + \nu_t \mathbf{1}\mathbf{1}^T
\]

with \( \eta_0, \nu_0 = (\delta, 1 - \delta) \) and

\[
\begin{bmatrix}
  \eta_{t+1} \\
  \nu_{t+1}
\end{bmatrix} = M \begin{bmatrix}
  \eta_t \\
  \nu_t
\end{bmatrix} = M^{t+1} \begin{bmatrix}
  \delta \\
  1 - \delta
\end{bmatrix}, \quad \forall t \geq 0,
\]

where

\[
M := \begin{bmatrix}
  d_1 & m_1 \\
  d_2 & m_2
\end{bmatrix},
\]

and \( d_1, d_2, m_1, m_2 \) are defined in \((43)\).

**Proof.** We first prove \((44)\) by induction. By \((15)\), it holds at \( t = 0 \). Now assume it holds for \( t = k \), for some \( \eta_k \) and \( \nu_k \), then for \( k + 1 \) we have from \((38)\)

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

Because \( \tilde{A}^{(k)} \) is in the form \((44)\), it is invariant to row and column permutations, that is, \( P^T \tilde{A}^{(k)} P = \tilde{A}^{(k)} \) for all \( P \in \Pi \). Hence,

\[
\begin{align*}
\tilde{A}^{(k+1)} &= \mathbb{E}_P[P C^T \tilde{A}^{(k)} C P^T] \\
&= \eta_k \mathbb{E}_P[P C^T C P^T] + \nu_k \mathbb{E}_P[P C^T \mathbf{1}\mathbf{1}^T C P^T] \\
&= \eta_k (d_1 I + d_2 \mathbf{1}\mathbf{1}^T) + \nu_k (m_1 I + m_2 \mathbf{1}\mathbf{1}^T) \\
&= (\eta_k d_1 + \nu_k m_1) I + (\eta_k d_2 + \nu_k m_2) \mathbf{1}\mathbf{1}^T,
\end{align*}
\]

\[(47)\]
Theorem 3. Consider solving (4) with the matrix $A$ defined in (15) using RPCD. Suppose that the elements of $x^0$ are i.i.d. $N(0,1)$. Then the expected value of $f$ after $\ell$ epochs of RPCD is $\sigma_\ell := n(\eta_\ell + \nu_\ell)/2$.

Proof. We have
\[
\mathbb{E} \left( f(x_{\text{RPCD}}^n) \right) = \frac{1}{2} \mathbb{E}_{x^0} (x^0)^T \bar{A}(\ell) x^0 \\
= \frac{1}{2} \eta_\ell \mathbb{E}_{x^0} (\|x^0\|_2^2) + \frac{1}{2} \nu_\ell \mathbb{E}_{x^0} ((1^T x^0)^2) \\
= \frac{1}{2} (\eta_\ell + \nu_\ell) n.
\]

Figure 5 plots the estimate obtained from this theorem against the value of $f(x_{\text{RPCD}}^n)$ obtained from (36) for particular random choices of $x^0$ and the permutation matrices $P_1, P_2, \ldots, P_\ell$. The close similarity between the two curves confirms that the analysis is tight.

The details of computing $d_1, d_2, m_1, m_2$ from (43) as functions of $\delta$ and $n$ are shown in Appendix A. The full expressions are quite complex, but we obtain simpler approximations for the case of $\delta$ with $0 < \delta \ll 1$ and large $n$. We can thus analyze the
asymptotic convergence of the sequence \( \{(\eta_t, \nu_t)\} \) by estimating the spectral radius \( \rho(M) \) as a function of \( \delta \).

The eigenvalues of \( M \) are solutions of the following quadratic:

\[
0 = \det(M - \lambda I) = (d_1 - \lambda)(m_2 - \lambda) - d_2m_1 = \lambda^2 - (d_1 + m_2)\lambda + d_1m_2 - d_2m_1,
\]

which are

\[
(48) \quad \lambda = \frac{1}{2} \left[ d_1 + m_2 \pm \sqrt{(d_1 + m_2)^2 - 4(d_1m_2 - d_2m_1)} \right].
\]

Let us consider the regime in which \( \delta \) is small and positive and \( n \) is somewhat larger than 1. The formulas (56) yield the following approximations for \( d_1, d_2, m_1, \) and \( m_2 \):

\[
\begin{align*}
d_1 &= 1 - 2\delta - 2\frac{\delta}{n} + O(\delta^2), \\
m_2 &= \frac{2\delta^3}{n^2} + O\left(\frac{\delta^4}{n^2}\right) + O\left(\frac{\delta^3}{n^3}\right), \\
d_2 &= 1 - 2\delta - 2\frac{\delta}{n} + 4\frac{\delta}{n} + O(\delta^2), \\
m_1 &= \frac{\delta^2}{n} + O\left(\frac{\delta^3}{n^2}\right) + O\left(\frac{\delta^4}{n}\right).
\end{align*}
\]

We thus have

\[
(49) \quad d_1 + m_2 = 1 - 2\delta - 2\frac{\delta}{n} + O(\delta^2),
\]

and

\[
\begin{align*}
d_1m_2 - d_2m_1 &= \left(1 - 2\delta - 2\frac{\delta}{n} + O(\delta^2)\right) \left(\frac{2\delta^3}{n^2} + O\left(\frac{\delta^4}{n^2}\right) + O\left(\frac{\delta^3}{n^3}\right)\right) \\
&\quad - \left(1 - 2\delta - 2\frac{\delta}{n} + 4\frac{\delta}{n} + O(\delta^2)\right) \left(\frac{\delta^2}{n} + O\left(\frac{\delta^3}{n^2}\right) + O\left(\frac{\delta^4}{n}\right)\right) \\
&\quad = \left(\frac{2\delta^3}{n^2} + O\left(\frac{\delta^4}{n^2}\right) + O\left(\frac{\delta^3}{n^3}\right)\right) - \left(\frac{\delta^2}{n} + O\left(\frac{\delta^3}{n^2}\right) + O\left(\frac{\delta^4}{n}\right)\right) \\
&\quad = -\frac{\delta^2}{n} + O\left(\frac{\delta^2}{n^2}\right) + O\left(\frac{\delta^3}{n}\right).
\end{align*}
\]

Since for \( \delta \) sufficiently small and \( n \) sufficiently large, we have

\[
d_1 + m_2 > 0, \quad (d_1 + m_2)^2 \gg |d_1m_2 - d_2m_1|,
\]

it follows from (48) that the eigenvalues of \( M \) can be approximated as follows:

\[
\lambda_1 \approx d_1 + m_2 - \frac{(d_1m_2 - d_2m_1)}{d_1 + m_2}, \quad \lambda_2 \approx \frac{(d_1m_2 - d_2m_1)}{d_1 + m_2}.
\]

By substituting from (49) and (50), we obtain

\[
\lambda_1 = 1 - 2\delta - 2\frac{\delta}{n} + O(\delta^2), \quad \lambda_2 \approx -\frac{\delta^2}{n}.
\]
Elementary calculation shows that the matrix of eigenvectors has the approximate form

\[ V \approx \begin{bmatrix} 1 & 0 \\ a & 1 \end{bmatrix}, \quad \text{for some } a \approx 1, \text{ whose inverse is } V^{-1} \approx \begin{bmatrix} 1 & 0 \\ -a & 1 \end{bmatrix}. \]

We thus have that

\[ M^t \approx V \left[ (1 - 2\delta - \frac{2\delta}{n})^t \right] \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} V^{-1} = \left[ (1 - 2\delta - \frac{2\delta}{n})^t \right] \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \]

so it follows from Theorem 2 that

\[ \left[ \frac{\eta_t}{\nu_t} \right] = M^t \left[ \begin{array}{c} \delta \\ 1 - \delta \end{array} \right] \approx \left[ (1 - 2\delta - \frac{2\delta}{n})^t \right] \begin{bmatrix} \delta \\ 0 \end{bmatrix}, \quad t = 1, 2, \ldots. \]

Thus from Theorem 3, we have

\[ (51) \quad \sigma_0 = \frac{n}{2}, \quad \sigma_t = \frac{n}{2} \left( 1 - 2\delta - \frac{2\delta}{n} \right)^t \delta \leq \frac{n}{2} (1 - 2\delta)^t \delta, \quad t = 1, 2, \ldots. \]

Note in particular that the decrease on the first iteration is rather dramatic — a factor of approximately \( \delta \). The factor-of-\( \delta \) decrease in error over the first iteration for RPCD is empirically observed in this experiment. The same behavior occurs in all three variants of CD. This phenomenon for CCD is predicted in [12, Theorem 3.1], but the reason that the same happens in RCD remains unclear.

On later iterations \((51)\) indicates that an asymptotic linear rate of \(1 - 2\delta\) cuts in for RPCD. This rate gives a complexity improvement of \(O(n^2)\) over that of CCD. On the other hand, this complexity is also about half that of RCD shown in \((34)\) (a rate of approximately \(1 - 2\delta\) for RPCD and \(1 - \delta\) for RCD). This result conforms with the observation frequently made that RPCD is usually faster than RCD in practice. In our experiments reported in Figure 3 and Table 1, we observe that indeed RPCD is slightly better than RCD, but the difference is not as large as predicted from the analysis. The empirical value of \(\rho_{RCD}(\delta, x^0)\) is closer to \(1 - 2\delta\), indicating that some tightening may be available for the RCD analysis on this problem.

2.5. 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 \(\delta\). For the empirical rates \(\rho_{CCD}(\delta, x^0), \rho_{RCD}(\delta, x^0), \text{ and } \rho_{RPCD}(\delta, x^0)\), we used such formulas as \((20)\), but we took the average decrease factor \emph{only over the last 10 epochs}, so as to capture the asymptotic rates. We used the termination criterion \(f(x^n) - f^* \leq 10^{-8}\). For the theoretical predictions, we used \(\rho(C)^2\) for CCD (as suggested by \((22)\)), the formula \(\rho_{RCD}(\delta, \text{predicted}) = (1 - \delta/n)^n\) for RCD (see \((34)\)), and \(\rho(M)\) for RPCD (as suggested by Theorems 2 and 3 and Gelfand’s formula). We note from this table that the theoretical predictions for CCD and RPCD are quite sharp, even for values of \(\delta\) that are not particularly small. For RCD, the empirical results are somewhat better than predicted by the theory. RPCD has the best practical and theoretical asymptotic convergence of the three variants, with the advantage increasing as \(\delta\) increases.

3. Conclusions. Recent work has shown that the problem \((4)\) with Hessian matrix \((15)\) 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 random-permutations cyclic coordinate descent that sharply predicts
Table 1: Observed and predicted per-epoch convergence rates for CCD, RCD, and RPCD, for various values of $\delta$. ($n = 100$ in all experiments.)

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$\rho_{\text{CCD}}(\delta, x^0)$</th>
<th>$\rho(C)^2$</th>
<th>$\rho_{\text{RCD}}(\delta, x^0)$</th>
<th>$\rho_{\text{RPCD}}(\delta, \text{predicted})$</th>
<th>$\rho_{\text{RPCD}}(\delta, x^0)$</th>
<th>$\rho(M)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>0.9344</td>
<td>0.9342</td>
<td>0.3472</td>
<td>0.4479</td>
<td>0.1158</td>
<td>0.1162</td>
</tr>
<tr>
<td>0.50</td>
<td>0.9924</td>
<td>0.9924</td>
<td>0.4560</td>
<td>0.6058</td>
<td>0.3255</td>
<td>0.3289</td>
</tr>
<tr>
<td>0.33</td>
<td>0.9972</td>
<td>0.9971</td>
<td>0.5674</td>
<td>0.7185</td>
<td>0.4964</td>
<td>0.4994</td>
</tr>
<tr>
<td>0.20</td>
<td>0.9988</td>
<td>0.9988</td>
<td>0.7060</td>
<td>0.8186</td>
<td>0.6633</td>
<td>0.6635</td>
</tr>
<tr>
<td>0.10</td>
<td>0.9995</td>
<td>0.9995</td>
<td>0.8196</td>
<td>0.9048</td>
<td>0.8173</td>
<td>0.8164</td>
</tr>
<tr>
<td>0.03</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9443</td>
<td>0.9704</td>
<td>0.9407</td>
<td>0.9412</td>
</tr>
</tbody>
</table>

It remains a challenge to analyze the distinctions between CCD, RCD, and RPCD in more general settings, and to discover and understand other instances that bring out differences in performance between these three variants.

REFERENCES

Appendix A. Estimating Terms in the Recurrence Matrix.

Here we first give detailed computations for the following quantities, for the matrix $A$ given in (15) and the corresponding value of $C$ defined in (27) and (29):

$$(1^T C 1)^2, \quad C 1, \quad \|C 1\|^2, \quad \|C^T 1\|^2, \quad \|C\|_F^2.$$ 

We then use these quantities to compute $d_1, d_2, m_1,$ and $m_2$ from (43). We assume throughout that $n \geq 3$.

For $1^T C 1$, we have from (27) and (29) that

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

where $u$ and $v$ are in $\mathbb{R}^n$, with

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

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

We are mostly interested in small positive values of $\delta$, so we have

$$1^T C 1 = - \sum_{i=1}^{n} (n - i) \delta^{n-i}$$

$$= - \sum_{i=1}^{n} (n \sum_{t=0}^{n-i-1} \delta^t - \sum_{i=1}^{n} i \delta^{n-i})$$

$$= - \sum_{i=1}^{n} (1 - \delta^n \frac{1 - \delta^{n-i}}{1 - \delta} - \sum_{i=1}^{n} \sum_{t=0}^{n-i} \delta^t)$$

$$= -n + \sum_{i=1}^{n} (1 - \delta^{n-i+1}) + O(\delta^n)$$

$$= - \sum_{i=1}^{n} \delta^i + O(\delta^n)$$

$$= - \frac{\delta}{1 - \delta} + O(\delta^n)$$

$$= -\delta - \delta^2 + O(\delta^3).$$

Therefore,

$$(52) \quad (1^T C 1)^2 = \delta^2 + 2\delta^3 + O(\delta^4).$$

We now use (30) to compute the other quantities required for evaluation of (43).
We have

$$(C_1)_i = \sum_{j=1}^{i} C_{ij} + \sum_{j=i+1}^{n} C_{ij}$$

$$= (1 - \delta) \sum_{j=1}^{i} (\delta^{i-j} - \delta^{i-1}) - (1 - \delta) \sum_{j=i+1}^{n} \delta^{i-1}$$

$$= (1 - \delta) \sum_{j=1}^{i} \delta^{i-j} - (1 - \delta)n\delta^{i-1}$$

$$= (1 - \delta^i) - n(1 - \delta)\delta^{i-1}$$

$$= 1 - n\delta^{i-1} + (n - 1)\delta^i.$$

We thus obtain

$$\|C_1\|^2 = \sum_{i=1}^{n} \left[ 1 - 2n\delta^{i-1} + 2(n - 1)\delta^i + n^2\delta^{2i-2} - 2n(n - 1)\delta^{2i-1} + (n - 1)^2\delta^{2i} \right]$$

$$= n - 2n \sum_{i=1}^{n} \delta^{i-1} + 2(n - 1)\delta \sum_{i=1}^{n} \delta^{i-1} + n^2 \sum_{i=1}^{n} (\delta^2)^{i-1}$$

$$- 2n(n - 1)\delta \sum_{i=1}^{n} (\delta^2)^{i-1} + O(\delta^2 n^2)$$

$$= n - \frac{2n}{1 - \delta} + \frac{2(n - 1)\delta}{1 - \delta} + n^2 - 2n(n - 1)\delta + O(\delta^2 n^2)$$

$$= n - 2n(1 + \delta) + 2(n - 1)\delta + n^2 - 2n(n - 1)\delta + O(\delta^2 n^2)$$

$$= n(n - 1) + \delta(-2n^2 + 2n - 2) + O(\delta^2 n^2).$$

We next have

$$(C^T_1)_j = \sum_{i=1}^{j-1} C_{ij} + \sum_{i=j}^{n} C_{ij}$$

$$= -(1 - \delta) \sum_{i=1}^{j-1} \delta^{i-1} + (1 - \delta) \sum_{i=j}^{n} (\delta^{i-j} - \delta^{i-1})$$

$$= -(1 - \delta) \sum_{i=1}^{n} \delta^{i-1} + (1 - \delta) \sum_{t=0}^{n-j} \delta^t$$

$$= -(1 - \delta^n) + (1 - \delta) \frac{1 - \delta^{n-j+1}}{1 - \delta}$$

$$= \delta^n - \delta^{n-j+1}. $$
Therefore, we have

\[
\|C^T 1\|^2_2 = \sum_{j=1}^{n} (\delta^n - \delta^{n-j+1})^2 \\
= \sum_{j=1}^{n} \delta^{2n-2j+2} + O(\delta^{n+1}) \\
= \delta^2 \sum_{j=1}^{n} (\delta^2)^{n-j} + O(\delta^{n+1}) \\
= \delta^2 \frac{1}{1 - \delta^2} + O(\delta^{n+1}) \\
= \delta^2 + O(\delta^4).
\]

(54)

For \(\|C\|^2_F\), we obtain

\[
\frac{1}{(1 - \delta^2)^2} \|C\|^2_F \\
= \sum_{j=1}^{n} \left\{ \sum_{i=1}^{j-1} \delta^{2i-2} + \sum_{i=j}^{n} (\delta^{2i-2j} - 2\delta^{2i-j-1} + \delta^{2i-2}) \right\} \\
= \sum_{j=1}^{n} \left\{ \sum_{i=1}^{j-1} \delta^{2i-2} + \sum_{i=j}^{n} \delta^{2i} - 2\delta^{j-1} \sum_{t=0}^{n-j} \delta^{2t} + \delta^{j-2} \sum_{i=1}^{n-j} \delta^{2i} \right\} \\
= \sum_{j=1}^{n} \left\{ \frac{1 - (\delta^2)^{j-1}}{1 - \delta^2} + \left(1 - 2\delta^{j-1} + \delta^{2j-2}\right) \frac{1 - \delta^{2n-2j+2}}{1 - \delta^2} \right\} \\
= \frac{1}{1 - \delta^2} \sum_{j=1}^{n} \left\{ 2 - (\delta^2)^{j-1} - 2\delta^{j-1} + \delta^{2j-2} - \delta^{2n-2j+2} + 2\delta^{2n-j+1} - \delta^{2n} \right\} \\
= \frac{1}{1 - \delta^2} \left\{ 2n - (1 + \delta^2) - 2(1 + \delta^2) + (1 + \delta^2) - \delta^2 + O(\delta^3) \right\} \\
= \frac{1}{1 - \delta^2} \left\{ 2n - 2 - 2\delta - 3\delta^2 \right\} + O(\delta^3)
\]

so that

\[
\|C\|^2_F = \frac{1 - \delta}{1 + \delta} (2n - 2 - 2\delta - 3\delta^2) + O(\delta^3) \\
= (1 - \delta)(1 - \delta + \delta^2)(2n - 2 - 2\delta - 3\delta^2) + O(n\delta^3) \\
= (1 - 2\delta + 2\delta^2)(2n - 2 - 2\delta - 3\delta^2) + O(n\delta^3) \\
= (2n - 2) - \delta(4n - 2) + \delta^2(4n - 3) + O(n\delta^3).
\]

(55)

From the formulas (43) together with (52), (53), (54), and (55), we have the
Thus by strong convexity, we have
\[
\text{set for (1) has the form } \{ x \mid f(x) \geq f(t^*) + \frac{\sigma}{2} \| x - P(x) \| \|^2 \geq f^* + \frac{\sigma \sigma_{\min,nz}}{2} \| x - P(x) \|^2 \}
\]
Meanwhile we have by convexity of $f$ that
\[ f^* \geq f(x) + \nabla f(x)^T (P(x) - x), \]
so that
\[ f(x) - f^* \leq \| \nabla f(x) \| \| P(x) - x \| \leq \| \nabla f(x) \| \left( \frac{2}{\sigma_{\text{min, nz}}^2} \right)^{1/2} (f(x) - f^*)^{1/2}. \]
Dividing both sides by $(f(x) - f^*)^{1/2}$ we obtain
\[ \| \nabla f(x) \| \left( \frac{2}{\sigma_{\text{min, nz}}^2} \right)^{1/2} \geq (f(x) - f^*)^{1/2} \Rightarrow \| \nabla f(x) \|^2 \geq \frac{\sigma_{\text{min, nz}}^2}{2} (f(x) - f^*), \]
which has the form (10).