MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

Analyzing Random Permutations for Cyclic Coordinate Descent

STEPHEN J. WRIGHT AND CHING-PEI LEE

Abstract. We consider coordinate descent methods for minimization of convex quadratic functions, in which exact line searches are performed at each iteration. (This algorithm is identical to Gauss-Seidel on the equivalent symmetric positive definite linear system.) We describe a class of convex quadratic functions for which the random-permutations version of cyclic coordinate descent (RPCD) is observed to outperform the standard cyclic coordinate descent (CCD) approach on computational tests, yielding convergence behavior similar to the fully-random variant (RCD). A convergence analysis is developed to explain the empirical observations.

Coordinate descent, Gauss-Seidel, randomization, permutations

1. Introduction

The coordinate descent (CD) approach for solving the problem

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

follows the framework of Algorithm 1. We denote

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

where the single nonzero in e_i appears in position i. Epochs (indicated by the counter ℓ) encompass cycles of inner iterations (indicated by j). At each iteration k, one component of x is selected for updating; a steplength parameter α_k is applied to the negative gradient of f with respect to this component.

Received by the editor January 6, 2020.
2010 *Mathematics Subject Classification.* Primary 65F10; Secondary 90C25, 68W20.
This work was supported by NSF Awards IIS-1447449, 1628384, 1634597, and 1740707; ONR Award N00014-13-1-0129; AFOSR Award FA9550-13-1-0138, and Subcontracts 3F-30222 and 8F-30039 from Argonne National Laboratory; and Award N660011824020 from the DARPA Lagrange Program. Ching-pei Lee was at the University of Wisconsin-Madison when this work was originally submitted.

©XXXX American Mathematical Society

Algorithm 1 Coordinate Descent

```
Set Choose x^0 ∈ ℝ^n;
for ℓ = 0, 1, 2, ... do
  for j = 0, 1, 2, ..., n − 1 do
    Define k = ℓn + j;
    Choose index i = i(ℓ, j) ∈ {1, 2, ..., n};
    Choose α_k > 0;
    x^{k+1} ← x^k − α_k ∇_i f(x^k) e_i;
  end for
end for
```

The choice of coordinate $i = i(\ell, j)$ to be updated at inner iteration j of epoch ℓ differs between variants of CD, as follows:

- For "cyclic CD" (**CCD**), we choose $i(\ell, j) = j + 1$.
- For "fully randomized CD," also known as "stochastic CD," and abbreviated as **RCD**, we choose $i(\ell, j)$ uniformly at random from $\{1, 2, \ldots, n\}$ and independently at each iteration.
- For "random-permutations CD" (abbreviated as **RPCD**), we choose $\pi_{\ell+1}$ at the start of epoch ℓ to be a random permutation of the index set $\{1, 2, \ldots, n\}$ (chosen uniformly at random from the space of random permutations), then set $i(\ell, j)$ to be the $(j + 1)$th entry in $\pi_{\ell+1}$, for $j = 0, 1, 2, \ldots, n - 1$.

Note that x^{ln} denotes the value of x after l epochs.

We consider in this paper problems in which f is a strictly convex quadratic, that is

$$f(x) = \frac{1}{2} x^T A x, \tag{1.3}$$

with A symmetric positive definite. Even this restricted class of functions reveals significant diversity in convergence behavior between the three variants of CD described above. The minimizer of (1.3) is obviously $x^* = 0$. Although (1.3) does not contain a linear term, it is straightforward to extend our results to the case for problems of the form

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

by replacing x^0 in several places of our analysis with $x^0 - x^*$, where $x^* = A^{-1} b$ is the minimizer of this problem. We assume that the choice of α_k in Algorithm 1 is the *exact* minimizer of f along the chosen coordinate direction. The resulting approach is thus equivalent to the Gauss-Seidel method applied to the linear system $Ax = 0$. The variants CCD, RCD, RPCD can be interpreted as different cyclic / randomized variants of Gauss-Seidel for this system.

In the RPCD variant, we can express a single epoch as follows. Letting P be the permutation matrix corresponding to the permutation π on this epoch, we split the symmetrically permuted Hessian into strictly triangular and diagonal parts as follows:

$$P^T A P = L_P + \Delta_P + L_P^T, \tag{1.4}$$

where L_P is strictly lower triangular and Δ_P is diagonal. We then define

$$C_P := -(L_P + \Delta_P)^{-1} L_P^T, \tag{1.5}$$

so that the epoch indexed by $l-1$ can be written as follows:

$$x^{ln} = (P_l C_{P_l} P_l^T) x^{(l-1)n}, \tag{1.6}$$

where P_l denotes the matrix corresponding to permutation π_l. By recursing to the initial point x^0, we obtain after ℓ epochs that

$$x^{\ell n} = (P_\ell C_{P_\ell} P_\ell^T)(P_{\ell-1} C_{P_{\ell-1}} P_{\ell-1}^T) \dots (P_1 C_{P_1} P_1^T) x^0, \tag{1.7}$$

yielding a function value of

$$f(x^{\ell n}) = \frac{1}{2}(x^0)^T \left((P_1 C_{P_1}^T P_1^T) \dots (P_\ell C_{P_\ell}^T P_\ell^T) A (P_\ell C_{P_\ell} P_\ell^T) \dots (P_1 C_{P_1} P_1^T) \right) x^0. \tag{1.8}$$

We analyze convergence in terms of the expected value of f after ℓ epochs for any given x^0, with the expectation taken over the permutations $P_1, P_2, \dots, P_\ell$, that is,

$$\mathbb{E}_{P_1, P_2, \dots, P_\ell} f(x^{\ell n}). \tag{1.9}$$

1.1. **Previous Work.** Convergence of RCD is analyzed in [5], showing that when the objective is strongly convex, the method requires $O((nL_{\max}/\mu)|\log \hat{\epsilon}|)$ iterations to reach an objective function value that is within $\hat{\epsilon}$ of the optimal value, in expectation, when the coordinates are sampled in a uniform random manner, where μ is the modulus of strong convexity and $L_{\max}$ is the maximum coordinate-wise Lipschitz constant for the gradient. This rate can be improved to $O((nL_{\text{avg}}/\mu)|\log \hat{\epsilon}|)$ if the sampling probability for each coordinate is proportional to the coordinate-wise Lipschitz constants, where L_{avg} is the average of these constants. On the other hand, the best known convergence rate of CCD for convex quadratic problems, given by [10], has an iteration complexity for reaching an $\hat{\epsilon}$-accurate solution *deterministically* that can be $O(n^2)$ times slower than that for RCD in the worst case. The best worst-case convergence guarantees for RPCD so far are still identical to those for CCD. (The analyses for CCD assume only that each coordinate is processed exactly once per epoch, and are indifferent to the fact that the ordering of coordinates can change on each epoch, as in RPCD.) However, in practice it is sometimes observed that RPCD behaves in a manner more similar to RCD than CCD (see, for example, the experiments in [7] and the talks [11, 12]), and a rigorous explanation for the general convergence rate for the expected objective of RPCD over the random permutations has been difficult to obtain. Some trials have been conducted to tackle this problem. Recht and Ré [6] state a conjecture whose consequence is that RPCD converges faster than RCD on quadratic problems, but they prove the result only for some special random cases. Sun et al. [9] have analyzed the convergence speed of the distance between the expected iterate $\mathbb{E}[x^k]$ and the minimizer x^* for convex quadratic problems, but this cannot be translated to a result for the expected squared error $\mathbb{E}\|x^k - x^*\|^2$ nor the expected function suboptimality $\mathbb{E}(f(x^k) - f(x^*))$, which are much more informative quantities.[1]

Computational experience reported in [11, 12] showed that for most convex quadratic functions (1.3), the convergence behaviors of all variants of CD are similar. For example, when A is a matrix of the form $V\Sigma V^T$ where V is random orthogonal and Σ is a positive diagonal matrix whose diagonals (the eigenvalues of A) follow

[1] As an example of why a sequence $\{x^k\}$ for which $\mathbb{E}[x^k] = x^*$ does not give useful information about convergence rate, consider $x^k = x^* + r^k$, where r^k are drawn i.i.d. from $N(0, I)$. Such a sequence has $\mathbb{E}[x^k] = x^*$, yet it has $\mathbb{E}\|x^k - x^*\|^2 = 1$, so cannot be said to converge to x^* in expectation.

a log-uniform distribution, then CCD, RCD, and RPCD all converge at roughly the same rates, no matter how widely the eigenvalues are dispersed. However, these computational tests revealed a class of matrices A for which the variants had radically different performance: matrices of the form

$$A = \delta V \Sigma V^T + (1-\delta)\mathbf{1}\mathbf{1}^T, \tag{1.10}$$

for small positive values of δ, and $\mathbf{1} = (1, 1, \ldots, 1)^T$. For such matrices, the performance of RCD and RPCD is similar, but CCD converges much more slowly. In this paper, we explain much of this anomalous behavior by considering a matrix closely related to (1.10), and explaining the difference by means of a specialized analysis of RPCD.

The current paper is an extension of our paper [3] in which, motivated by the empirical observation above, we considered the special case of (1.10) in which $\Sigma = I$, that is,

$$A := \delta I + (1-\delta)\mathbf{1}\mathbf{1}^T, \quad \text{where } \delta \in (0, n/(n-1)). \tag{1.11}$$

It was proved by [10] that this matrix achieves worst-case convergence behavior for CCD. We showed in [3] that a factor of $O(n^2)$ fewer iterations are required by RPCD to achieve the same accuracy, and that the complexity of RPCD is similar to RCD in this case. Salient properties of the matrix (1.11) include the following.

(a) It has eigenvalue δ replicated $(n-1)$ times, and a single dominant eigenvalue $\delta + (1-\delta)n$, and
(b) it is invariant under symmetric permutations, that is, $P^T A P = A$ for all permutation matrices P.

The latter property makes the analysis of RPCD much more straightforward than for more general A of the form (1.10). Specifically, it follows from (1.5) that $C_P \equiv C = -(L+\Delta)^{-1}L^T$, where $A = L + \Delta + L^T$, that is, C_P is independent of P. For the matrix (1.11), the expression (1.6) thus simplifies to

$$x^{ln} = (P_l C P_l^T) x^{(l-1)n}, \quad l = 1, 2, 3, \ldots .$$

We refer to [3] for a more extensive discussion of prior related work on variants of coordinate descent. We note in particular that for general convex functions f, CCD has weaker convergence guarantees than for convex *quadratic* f, as analyzed in [1, 8, 4]. By contrast, the convergence results for RCD presented in [5] show no difference between quadratic and nonquadratic convex functions.

1.2. **Contributions.** In this work, we study the behavior of the RPCD variant of CD on problems of the form (1.3), where the coefficient matrix has the form

$$B_u := \delta I + (1-\delta) u u^T, \tag{1.12}$$

for some $u \in \mathbb{R}^n$. This paper focuses on the case in which the components of u are not too different in magnitude, and are all close to 1. Rather than working directly with (1.12), we work with a diagonally scaled version that has a form more tractable for analysis. By scaling (1.12) symmetrically with the matrix $U = \mathrm{diag}(u)$, we obtain

$$\begin{aligned} &A_\epsilon := \delta I + (1-\delta)\mathbf{1}\mathbf{1}^T + \epsilon D, \\ &\text{where } \delta \in (0, n/(n-1)),\ \epsilon \geq 0, \\ &D = \mathrm{diag}(d), \text{ with } \min_i d_i = 0 \text{ and } \max_i d_i = 1. \end{aligned} \tag{1.13}$$

(Details are given in Section 2.) Note that both forms (1.12) and (1.13) are generalizations of (1.11). They are both closely related to the more general form (1.10), in that (1.12) can be obtained from (1.10) by a symmetric scaling with $\Sigma^{-1/2}V$, while (1.13) has the form (1.10) with $V = I$ and $\Sigma = I + (\epsilon/\delta)D$. Thus, this paper provides a significantly more complete explanation of the anomalous convergence behavior involving matrices (1.10) than our earlier work.

For matrices of the form (1.13) in (1.3), this paper proves similar convergence behavior for RPCD to what was proved in [3] for the special case (1.11), in the regime defined by the following values of the parameters n, ϵ, and δ:

$$0 < \delta \leq \epsilon, \quad |\rho_1|\,\epsilon^2 < \delta \ll 1, \quad n\epsilon \leq 1. \tag{1.14}$$

where ρ_1 is a positive or negative quantity of modest size, magnitude not much larger than 1, and independent of n, ϵ, and δ. We prove that the convergence rate guarantee of RPCD for problems defined by (1.13) is similar to that of RCD, and much better than the rate bound for CCD. Specifically, we explain via analysis of a linear recurrence that captures the epoch-wise behavior of RPCD that the per-epoch objective improvement is bounded by a factor of approximately

$$1 - 1.4\delta, \tag{1.15}$$

which is similar to the corresponding factors of approximately $1 - \delta$ and $1 - 2\delta$ that are known for RCD (by different analyses), and significantly better than the factor of approximately $(1 - \delta/n^2)$ arising from worst-case theoretical guarantees for CCD. By the generalization of (1.11) to (1.13) (and thus (1.12)), we extend our understanding of the empirical behavior of RPCD, RCD, and CCD described at the beginning of Section 1.1.

1.3. **Remainder of the Paper.** In Section 2, we relate matrices of the forms (1.13) and (1.12), showing that the behavior of CD is similar on both. Section 3 presents our analysis for the behavior of RPCD on problem (1.3), (1.13). In particular, we define a sequence of matrices $\{\bar{A}_\epsilon^{(t)}\}$ such that given any initial guess x^0, the expected value of the objective $f(x^{tn})$ after the tth epoch is $\frac{1}{2}(x^0)^T \bar{A}_\epsilon^{(t)} x^0$. We then define a sequence of matrices $\{\hat{A}_\epsilon^{(t)}\}$ that dominates $\{\bar{A}_\epsilon^{(t)}\}$, and that can be parametrized compactly. We analyze convergence of the sequence $\{\hat{A}_\epsilon^{(t)}\}$ by means of a spectral analysis of the matrix that relates its parameters at successive values of t, and use it to develop an estimate of the asymptotic per-epoch improvement of the objective $f(x^{tn})$, $t = 0, 1, 2, \ldots$. We provide an explanation in Section 3.5 for the large decrease in f that is often observed in the very first iteration of CD, a phenomenon that is not explained by the asymptotic analysis. Section 4 discusses RCD and CCD variants for the problem (1.3), (1.13), while Section 5 reports computational experience with the three variants.

1.4. **Notation.** In addition to the notation ρ_1 mentioned above, which denotes a scalar quantity of size not much greater than 1 and independent of n, ϵ, and δ, we make extensive use of vector quantities $\mathbf{r_1} \in \mathbb{R}^n$ and matrix quantities $\mathbf{R_1} \in \mathbb{R}^{n\times n}$ (symmetric in some contexts and nonsymmetric in others), which we assume are both bounded in norm by 1, that is,

$$\|\mathbf{r_1}\| \leq 1, \quad \|\mathbf{R_1}\| \leq 1. \tag{1.16}$$

In the case in which $\mathbf{R_1}$ is also symmetric, it follows from these assumptions that $-I \preceq \mathbf{R_1} \preceq I$. This notation is essential to capturing remainder terms that appear in our analysis. In particular, it allows us to keep explicit track of dependence of the remainder terms on n, ϵ, and δ. For example, a vector quantity whose size is bounded by a modest multiple of $\epsilon^2 n^{-1}$ can be represented by $\rho_1 \epsilon^2 n^{-1} \mathbf{r_1}$. The following estimate follows immediately from this notation:

$$\mathbf{R_1} v \mathbf{1}^T = \rho_1 \mathbf{r_1} \mathbf{1}^T \text{ provided } \|v\| \le \rho_1. \tag{1.17}$$

Matrix and vector norms $\|\cdot\|$ signify $\|\cdot\|_2$ throughout, unless some other subscript is specified.

2. Quadratic functions with Hessians of the form (1.13)

We discuss here the matrix of the form (1.13), explaining its relationship to (1.12) and to (1.11), and giving some preliminaries for the analysis of RPCD on the corresponding quadratic function.

2.1. Relating (1.13) to (1.12).

Given $\epsilon > 0$ and $\delta \in (0,1)$, suppose that $u \in \mathbb{R}^n$ satisfies

$$\min_{i=1,2,\ldots,n} |u_i| = \sqrt{\frac{\delta}{\delta+\epsilon}}, \quad \max_{i=1,2,\ldots,n} |u_i| = 1. \tag{2.1}$$

Consider the matrix B_u from (1.12). Defining $U := \mathrm{diag}(u)$, we have

$$A_\epsilon := U^{-1} B_u U^{-1} = \delta U^{-2} + (1-\delta)\mathbf{1}\mathbf{1}^T, \tag{2.2}$$

and note that the diagonal elements of U^{-2} are in the range $[1, \epsilon/\delta + 1]$. Thus we can write $\delta U^{-2} = \delta I + \epsilon D$, where D is diagonal with elements in $[0,1]$, so in fact A_ϵ in (2.2) has the form (1.13).

We verify in Appendix A that the iterates generated by Algorithm 1 for a given sequence of indices $i(\ell, j)$ to (1.3) with $A = B_u$ from (1.12), and with starting point $\tilde{x}^0$ and exact line search are isomorphic to the iterates generated by applying the same algorithm with the same index sequence to (1.3) with $A = A_\epsilon$ from (2.2), with starting point $x^0 = U\tilde{x}^0$. Specifically, we have $x^k = U\tilde{x}^k$ for all $k \ge 0$, where $\{\tilde{x}^k\}$ is the iterate sequence corresponding to (1.12) and $\{x^k\}$ is the sequence corresponding to (2.2). Note that the function values coincide at each iteration, that is,

$$\frac{1}{2}(\tilde{x}^k)^T B_u \tilde{x}^k = \frac{1}{2}(x^k)^T A_\epsilon x^k, \quad k = 0, 1, 2, \ldots. \tag{2.3}$$

Thus we expect to see similar asymptotic behavior for the quadratic objectives based on matrices (1.12) and (1.13), from starting points with the same distribution.

We note too that the matrix A_ϵ from (1.13) is "sandwiched" between scalar multiples of two matrices of the form (1.11). We have

$$\delta I + (1-\delta)\mathbf{1}\mathbf{1}^T \le A_\epsilon \le (1+\epsilon)\left(\delta' I + (1-\delta')\mathbf{1}\mathbf{1}^T\right), \tag{2.4}$$

where $\delta' = (\delta+\epsilon)/(1+\epsilon)$ and "$\le$" denotes element-wise inequality. This observation suggests similar behavior for RPCD to that proved for the matrices (1.11) in [3]. Indeed, we observe similar behavior empirically, but we could not find a way to exploit the relationship (2.4) in our convergence analysis. The distinctiveness of the components of D plays a key role; the effects of D in (1.13) persist through the epochs. The analysis techniques in [3] make strong use of the fact that the epoch-wise iteration matrix C_P defined in (1.5) is independent of P, a fact that no longer holds for matrices (1.13).

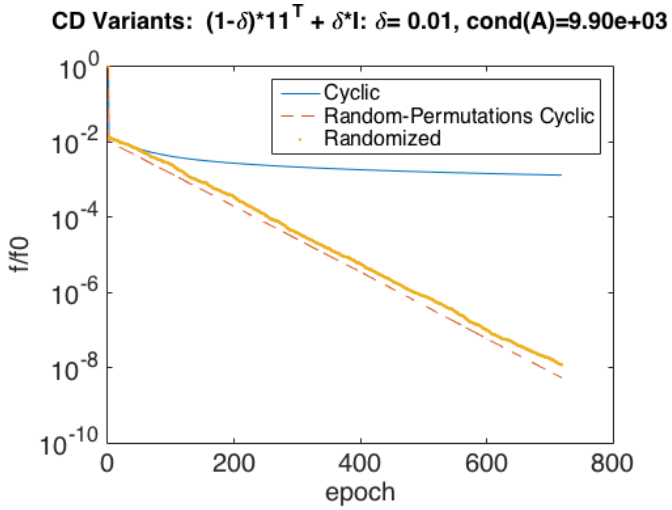
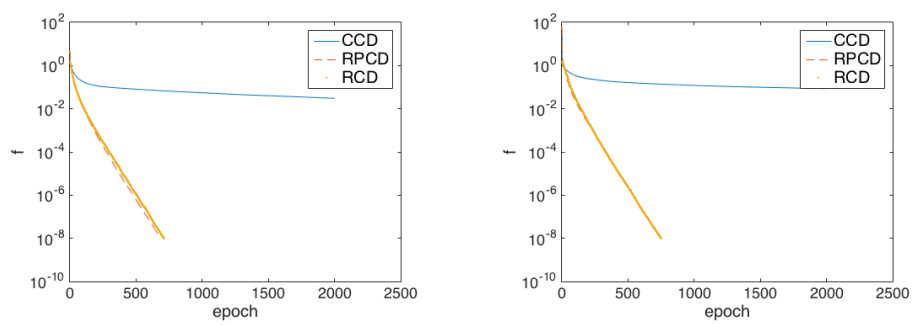


FIGURE 1. CCD, RPCD, and RCD on convex quadratic objective, where A is defined by (1.11) with $n = 100$ and $\delta = .01$.

Representative numerical results for the three versions of CD on quadratics with Hessians of the form (1.11) are shown in Figure 1. We note here the nearly identical linear rates of the RPCD and RCD variants, and the much slower rate of the CCD variant. The same pattern is observed in Figure 2, which considers matrices of the forms (1.12) and (1.13). Note in particular that the latter two matrices are indistinguishable in their empirical behavior, further justifying our focus on the form (1.13) in our analysis.



(A) Matrix (1.13). (B) Matrix (1.12) with u satisfying (2.1).

FIGURE 2. Comparison between CCD, RPCD, and RCD on different matrices with $n = 100$ and $(\delta, \epsilon) = (.01, .05)$.

2.2. **RPCD Preliminaries.** We now define some notation to be used in the remainder of the analysis: the matrix C_P that defines the change in iterate x over one epoch and the matrix $\bar{A}_\epsilon^{(\ell)}$ that defines the value $f(x^{\ell n})$ of the objective after ℓ epochs.

Applying to (1.13) the decomposition (1.4) into triangular and diagonal matrices, we obtain

$$\begin{aligned} P^T A_\epsilon P &= (1-\delta)E + P^T(\delta I + \epsilon D)P + (1-\delta)E^T \\ &= (1-\delta)E + (\delta I + \epsilon D_P) + (1-\delta)E^T, \end{aligned} \tag{2.5}$$

where

$$D_P := P^T D P, \quad E := \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 & 0 \end{bmatrix}. \tag{2.6}$$

Following (1.5), we have for A_ϵ that the epoch matrix is

$$C_P := -(1-\delta)\left[(1-\delta)E + (I + \epsilon D_P)\right]^{-1} E^T. \tag{2.7}$$

Our interest is in the quantity

$$\mathbb{E}_{P_1,P_2,\dots,P_\ell} f(x^{\ell n}), \quad \ell = 1, 2, \dots, \tag{2.8}$$

where $f(x^{\ell n})$ is defined by (1.8). Adapting notation from [3], we define the matrices $\bar{A}_\epsilon^{(t)}$, $t = 0, 1, 2, \dots, \ell$ as follows:

$$\begin{aligned} \bar{A}_\epsilon^{(0)} &= A_\epsilon, \\ \bar{A}_\epsilon^{(1)} &= \mathbb{E}_{P_\ell}\left((P_\ell C_{P_\ell}^T P_\ell^T) A_\epsilon (P_\ell C_{P_\ell} P_\ell^T)\right), \\ &\vdots \\ \bar{A}_\epsilon^{(\ell)} &= \mathbb{E}_{P_1,\dots,P_\ell}\left((P_1 C_{P_1}^T P_1^T)\dots(P_\ell C_{P_\ell}^T P_\ell^T) A_\epsilon (P_\ell C_{P_\ell} P_\ell^T)\dots(P_1 C_{P_1} P_1^T)\right). \end{aligned}$$

We have the following recursive relationship between successive terms in the sequence $\bar{A}_\epsilon^{(t)}$, $t = 0, 1, 2, \dots$:

$$\begin{aligned} \bar{A}_\epsilon^{(t)} &= \mathbb{E}_{P_{\ell-t+1}}(P_{\ell-t+1} C_{P_{\ell-t+1}}^T P_{\ell-t+1}^T \bar{A}_\epsilon^{(t-1)} P_{\ell-t+1} C_{P_{\ell-t+1}} P_{\ell-t+1}^T) \\ &= \mathbb{E}_P(P C_P^T P^T \bar{A}_\epsilon^{(t-1)} P C_P P^T), \end{aligned} \tag{2.9}$$

where we have dropped the subscript on $P_{\ell-t+1}$ in the second equality, since the permutation matrices for each epoch are i.i.d. Using this matrix, we can compute (2.8) by

$$\mathbb{E}_{P_1,P_2,\dots,P_\ell} f(x^{\ell n}) = \frac{1}{2}\left(x^0\right)^T \bar{A}_\epsilon^{(t)} x^0. \tag{2.10}$$

3. Epoch-Wise Convergence of Expected Function Value

In this section, we analyze the behavior of the sequence of matrices $\{\bar{A}_\epsilon^{(t)}\}$ that govern the expected value of the objective function f after t epochs of RPCD. By focusing on the operation (2.9) which tracks the change from one element of this sequence to the next, we show that this sequence is bounded in norm by a quantity that decreases to zero at an asymptotic rate similar to the known rate for the fully-random variant RCD.

We show that the matrix sequence $\{\bar{A}_\epsilon^{(t)}\}$ is dominated[2] by another sequence of positive definite matrices $\{\hat{A}_\epsilon^{(t)}\}$ that can be represented as a four-term recurrence

$$\hat{A}_\epsilon^{(t)} = \hat{\eta}_t I + \hat{\nu}_t \mathbf{1}\mathbf{1}^T + \hat{\epsilon}_t D + \hat{\tau}_t(\mathbf{1}\mathbf{r_1}^T + \mathbf{r_1}\mathbf{1}^T), \tag{3.1}$$

where $\mathbf{r_1}$ is a vector such that $\|\mathbf{r_1}\| \le 1$ (as defined in Section 1.4) and $(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)$ is a quadruplet of scalar coefficients for all $t = 0, 1, 2, \cdots$. (Note that the quantities $\mathbf{r_1}$ in the final term are generally different for each t.) We set $\hat{A}_\epsilon^{(0)} = \bar{A}_\epsilon^{(0)} = A_\epsilon$, with

$$\hat{\eta}_0 = \delta, \quad \hat{\nu}_0 = 1 - \delta, \quad \hat{\epsilon}_0 = \epsilon, \quad \hat{\tau}_0 = 0, \tag{3.2}$$

and define the sequence $\{\hat{A}_\epsilon^{(t)}\}$ so that successive elements satisfy the same relationship as shown in (2.9) for $\{\bar{A}_\epsilon^{(t)}\}$, namely

$$\hat{A}_\epsilon^{(t+1)} \succeq \mathbb{E}_P(PC_P^T P^T \hat{A}_\epsilon^{(t)} PC_P P^T).$$

Our analysis consists chiefly of analyzing the convergence to zero of the sequence of quadruplets $\{(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)\}_{t=0,1,2\ldots}$ corresponding to $\{\hat{A}_\epsilon^{(t)}\}_{1,2,\ldots}$.

After several definitions and technical results in Section 3.1, we derive in Section 3.2 a tractable representation of the matrix C_P from (1.5) that defines the transition between successive elements of the sequences $\{\bar{A}_\epsilon^{(t)}\}$ and $\{\hat{A}_\epsilon^{(t)}\}$. In Section 3.3, we examine the effect of the operation of C_P on each of the four terms in the bounding sequence (3.1). In Section 3.4, we define the recurrence that relates successive elements of the sequence $\{(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)\}_{t=0,1,2\ldots}$, and examine the rate at which this sequence converges to 0. We show that the per-epoch rate is bounded by a scalar sequence that converges at a nearly linear rate of $1 - 1.4\delta$. (Our analysis is conservative; the true rate, observed in experiments, is often closer to $1 - 2\delta$.)

In most of this section, we consider the regime for parameters n, ϵ, and δ defined by (1.14). The inequality $\delta \le \epsilon$ is made mostly for convenience; it implies that we can replace δ by ϵ in remainder terms, and it allows wide divergence in the diagonal elements of the matrix (1.13). (We expect that the main convergence results will continue to apply in a regime in which $0 \le \epsilon < \delta$, which indeed is closer to the matrix (1.11) studied in [3], which has constant diagonals, but the remainder terms in the analysis will need to be handled differently.) In the analysis of Section 3.4, we make additional assumptions on n, ϵ, and δ.

3.1. **Definitions and Technical Results.** We start by defining some useful quantities, drawing on [3], and proving several elementary results. While technical, these results give an idea of the effects of applying expectations over permutations to matrices that arise in the subsequent analysis.

From (1.13) and (3.3), we have

$$d = D\mathbf{1}, \quad d_{\text{av}} := \mathbf{1}^T d/n, \quad d_{\text{av},2} := \frac{1}{n}\mathbf{1}^T D^2 \mathbf{1}. \tag{3.3}$$

From the definition of D in (1.13), we have $d_{\text{av}} \in (0, 1)$ and $d_{\text{av},2} \in (0, 1)$. We use π to denote the permutation of $\{1, 2, \ldots, n\}$ associated with the permutation matrix

[2]Given two symmetric matrices F and G, we say that F dominates G if $F - G$ is positive semidefinite.

P, so that for any vector $u \in \mathbb{R}^n$, we have

$$P^T u = \begin{bmatrix} u_{\pi(1)} \\ u_{\pi(2)} \\ \vdots \\ u_{\pi(n)} \end{bmatrix}, \quad D_P = P^T D P = \operatorname{diag}(d_{\pi(1)}, d_{\pi(2)}, \ldots, d_{\pi(n)}). \tag{3.4}$$

We can see immediately that

$$P\mathbf{1} = \mathbf{1}, \tag{3.5a}$$

$$\mathbb{E}_P \, P e_j = \frac{1}{n}\mathbf{1}, \quad \text{for any } j = 1, 2, \ldots, n. \tag{3.5b}$$

A useful conditional probability is as follows:

$$\mathbb{E}_{P \mid P_{i1}=1} P e_2 = \frac{1}{n-1}(1 - e_i). \tag{3.6}$$

This claim follows because Pe_2 contains $n-1$ zeros and a single 1, and the 1 cannot appear in position i (because $Pe_1 = e_i$) but may appear in any other position with equal likelihood.

A quantity that appears frequently in the analysis is the matrix F defined by

$$F := \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 \end{bmatrix}, \tag{3.7}$$

that is, the $n \times n$ matrix of all zeros except for 1 on the diagonal immediately above the main diagonal. We see immediately that $\|F\| = 1$. Several identities follow:

$$F^T e_1 = e_2, \quad F e_1 = 0, \quad F\mathbf{1} = \mathbf{1} - e_n. \tag{3.8}$$

We also have

$$\mathbb{E}_P \, (PFP^T) = \frac{1}{n}(\mathbf{1}\mathbf{1}^T - I). \tag{3.9}$$

To verify this claim, note that the diagonals of PFP^T are zero for all permutation matrices P, while the off-diagonals are 1 with equal probability. Thus the expected value of the $n(n-1)$ off-diagonal elements is obtained by distributing the $n-1$ nonzeros in F with equal weight among all off-diagonal elements, giving an expected value of $1/n$ for each of these elements, as in (3.9).

We have the following results about quantities involving F.

Lemma 3.1.

$$PFP^T D P e_1 = 0, \tag{3.10a}$$

$$\mathbb{E}_P \, (PF^T P^T D P e_1) = \frac{1}{n-1}\left[d_{av}\mathbf{1} - \frac{1}{n} d \right]. \tag{3.10b}$$

Proof. For (3.10a), we see that $P^T D P e_1$ is a multiple of e_1, and that $F e_1 = 0$.

For (3.10b), we use $\mathbb{E}_i$ to denote the expectation with respect to index i uniformly distributed over $\{1, 2, \ldots, n\}$, and recall that π denotes the permutation

corresponding to P. We have

$$\begin{aligned}
\mathbb{E}_P\,(PF^TP^TDPe_1) &= \mathbb{E}_P\, d_{\pi(1)}PF^Te_1 && \text{from (3.4)}\\
&= \mathbb{E}_P\, d_{\pi(1)}Pe_2 && \text{from (3.8)}\\
&= \mathbb{E}_i\, d_i\mathbb{E}_{P\,|\,P_{i1}=1}\, Pe_2 \\
&= \mathbb{E}_i\, d_i\frac{1}{n-1}(\mathbf{1}-e_i) && \text{from (3.6)}\\
&= \frac{1}{n-1}\left[d_{\text{av}}\mathbf{1}-\frac{1}{n}d\right],
\end{aligned}$$

as required. □

Finally, we make frequent use of the following trivial result about the norm of rank-1 matrices: for any vectors $v, w \in \mathbb{R}^n$, we have

$$\|vw^T\| = \|v\|\|w\|. \tag{3.11}$$

In particular, we have from $\|\mathbf{1}\| = n^{1/2}$ that

$$\|\mathbf{1}v^T\| = n^{1/2}\|v\|, \tag{3.12}$$

and in particular, using the notation of Section 1.4, we have

$$\|\mathbf{1}\mathbf{r_1}^T\| \le n^{1/2}. \tag{3.13}$$

3.2. **Properties of the Epoch Matrix C_P.** As in [3], we define

$$\bar{L} := -(I + (1-\delta)E)^{-1}. \tag{3.14}$$

We noted in [3] that

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

so by using notation (3.7), we have

$$\bar{L} = -I + F^T + \delta\mathbf{R_1}. \tag{3.15}$$

We have further from a standard matrix-norm inequality together with the facts that $\|\bar{L}\|_1 \le 2$ and $\|\bar{L}\|_\infty \le 2$ that

$$\|\bar{L}\| \le \sqrt{\|\bar{L}\|_1\|\bar{L}\|_\infty} \le 2. \tag{3.16}$$

Moreover, from [3, Section 2.2], we have

$$(\bar{L}E^T)_{ij} = \begin{cases} -\delta^{i-1} & \text{for } i < j\\ \delta^{i-j} - \delta^{i-1} & \text{for } i \ge j, \end{cases}$$

so that

$$\bar{L}E^T = I - e_1\mathbf{1}^T + \delta F^T - \delta e_2\mathbf{1}^T + \rho_1\delta^2(\mathbf{R_1} + \mathbf{r_1}\mathbf{1}^T). \tag{3.17}$$

(The validity of the remainder term in this expression follows from the fact that the coefficients of $\delta^2, \delta^3, \ldots, \delta^{n-1}$ in $\bar{L}E^T$ all have the form $\mathbf{R_1} + \mathbf{r_1}\mathbf{1}^T$, so we can absorb them all into a single term of order δ^2 by summation.)

The following lemma provides a useful estimate of the epoch matrix C_P.

Lemma 3.2. *Suppose that* (1.14) *holds. Then for* C_P *defined by* (1.5) *and* (2.7), *we have*

$$(1-\delta)^{-1}C_P = I - e_1\mathbf{1}^T + \epsilon(-D_P + F^T D_P)(I - e_1\mathbf{1}^T) + \delta(F^T - e_2\mathbf{1}^T) + \epsilon^2(\rho_1\mathbf{r_1}\mathbf{1}^T + \rho_1\mathbf{R_1}). \tag{3.18}$$

Proof. Note first that for a matrix Y with $\|Y\| \le \rho_1$ and for ϵ satisfying (1.14), we have

$$(I-\epsilon Y)^{-1} = I + \epsilon Y + \epsilon^2(I-\epsilon Y)^{-1}Y^2 = I + \epsilon Y + \rho_1\epsilon^2\mathbf{R_1}. \tag{3.19}$$

From (2.7), using definition (3.14), we have

$$\begin{aligned}(1-\delta)^{-1}C_P &= -\left[(I+(1-\delta)E) + \epsilon D_P\right]^{-1} E^T \\ &= [\bar{L}^{-1} - \epsilon D_P]^{-1}E^T \\ &= [I - \epsilon\bar{L}D_P]^{-1}(\bar{L}E^T).\end{aligned}$$

By substituting from (3.17) and (3.19) (noting that $\|\bar{L}D_P\| \le \|\bar{L}\| \le 2$ from (3.16)), we have

$$\begin{aligned}(1-\delta)^{-1}C_P &= \left[I + \epsilon\bar{L}D_P + \rho_1\epsilon^2\mathbf{R_1}\right]\left[I - e_1\mathbf{1}^T + \delta F^T - \delta e_2\mathbf{1}^T + \delta^2(\rho_1\mathbf{R_1} + \rho_1\mathbf{r_1}\mathbf{1}^T)\right] \\ &= \left[I - e_1\mathbf{1}^T + \epsilon\bar{L}D_P(I - e_1\mathbf{1}^T) + \delta(F^T - e_2\mathbf{1}^T) + \epsilon^2(\rho_1\mathbf{R_1} + \rho_1\mathbf{r_1}\mathbf{1}^T)\right],\end{aligned}$$

where we used $\delta \le \epsilon$ from (1.14) to absorb the term $\delta^2(\rho_1\mathbf{R_1} + \rho_1\mathbf{r_1}\mathbf{1}^T)$. The result follows immediately when we use (3.15) to substitute for $\bar{L}$, and again use $\delta \le \epsilon$ together with $\|D_P\| \le 1$ and (1.17) to absorb the remainder terms. □

3.3. **Single-Epoch Analysis.** In this section we analyze the change in each term in the expression (3.1) over a single epoch. We examine in turn the following terms:

- the I term: Lemma 3.3,
- the D term: Lemma 3.4,
- the $\mathbf{1}\mathbf{1}^T$ and $(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T)$ terms: Lemma 3.5.

Proofs of these technical results appear in Appendix B.

Lemma 3.3. *Suppose that* (1.14) *holds. We have*

$$\begin{aligned}&(1-\delta)^{-2}\mathbb{E}_P\left(PC_P^TC_PP^T\right) \\ &= \left[I + \left(1 - \frac{2}{n}\right)\mathbf{1}\mathbf{1}^T\right] \\ &\quad + \epsilon\left[-2\left(1+\frac{1}{n}\right)D + \frac{3n-2}{n(n-1)}(d\mathbf{1}^T + \mathbf{1}d^T) - 2\frac{n}{n-1}d_{av}\mathbf{1}\mathbf{1}^T\right] \\ &\quad + \delta\left(\frac{-2}{n}\right)I + \epsilon^2(\rho_1\mathbf{1}\mathbf{1}^T + \rho_1\mathbf{r_1}\mathbf{1}^T + \rho_1\mathbf{1}\mathbf{r_1}^T + \rho_1\mathbf{R_1}) \\ &\preceq (1+\rho_1\epsilon^2)I + (1+\rho_1\epsilon^2)\mathbf{1}\mathbf{1}^T + (\rho_1\epsilon n^{-1/2} + \rho_1\epsilon^2)(\mathbf{1}\mathbf{r_1}^T + \mathbf{r_1}\mathbf{1}^T).\end{aligned}$$

Lemma 3.4. *Suppose that* (1.14) *holds. We have*

$$
\begin{aligned}
&(1-\delta)^{-2}\mathbb{E}\left(PC_P^TP^TDPC_PP^T\right)\\
&= \left[D + d_{av}\mathbf{1}\mathbf{1}^T - \frac{1}{n}(\mathbf{1}d^T + d\mathbf{1}^T)\right] + \delta\left[-\frac{2}{n}D\right]\\
&\quad + \epsilon\left[-2\left(1+\frac{1}{n}\right)D^2 - \frac{d_{av}}{n-1}(\mathbf{1}d^T + d\mathbf{1}^T) - 2d_{av,2}\mathbf{1}\mathbf{1}^T\right.\\
&\qquad\qquad \left. + \frac{2}{n}dd^T + \frac{2n-1}{n(n-1)}\left(\mathbf{1}\mathbf{1}^TD^2 + D^2\mathbf{1}\mathbf{1}^T\right)\right]\\
&\quad + \epsilon^2(\rho_1\mathbf{1}\mathbf{1}^T + \rho_1(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T) + \rho_1\mathbf{R_1})\\
&\preceq D + (2\epsilon + \rho_1\epsilon^2)I + (d_{av} + \rho_1\epsilon^2)\mathbf{1}\mathbf{1}^T + (\rho_1 n^{-1/2} + \rho_1\epsilon^2)(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T).
\end{aligned}
$$

Lemma 3.5. *Suppose that* (1.14) *holds. For any* $v \in \mathbb{R}^n$*, we have*

$$
\begin{aligned}
&(1-\delta)^{-2}\mathbb{E}_P\left(PC_P^TP^T(\mathbf{1}v^T + v\mathbf{1}^T)PC_PP^T\right)\\
&= -\epsilon\left[\frac{1}{n}(dv^T + vd^T) - \frac{\mathbf{1}^Tv}{n(n-1)}(d\mathbf{1}^T + \mathbf{1}d^T) + \frac{1}{n(n-1)}(Dv\mathbf{1}^T + \mathbf{1}v^TD)\right]\\
&\quad - \delta\left[\frac{1}{n-1}(\mathbf{1}v^T + v\mathbf{1}^T) - \frac{2\mathbf{1}^Tv}{n(n-1)}\mathbf{1}\mathbf{1}^T\right]\\
&\quad + \epsilon^2 n^{1/2}\|v\|(\rho_1\mathbf{R_1} + \rho_1(\mathbf{1}\mathbf{r_1}^T + \mathbf{r_1}\mathbf{1}^T) + \rho_1\mathbf{1}\mathbf{1}^T)
\end{aligned}
\tag{3.20}
$$

so that

$$
\begin{aligned}
&(1-\delta)^{-2}\mathbb{E}_P\left(PC_P^TP^T(\mathbf{1}v^T + v\mathbf{1}^T)PC_PP^T\right)\\
&\qquad \preceq \rho_1\|v\|(\epsilon n^{-1/2} + \epsilon^2 n)I + \rho_1\|v\|(\epsilon n^{-3/2} + \epsilon^2 n^{1/2})\mathbf{1}\mathbf{1}^T.
\end{aligned}
\tag{3.21}
$$

When $v = \mathbf{1}$*, we have*

$$
(1-\delta)^{-2}\mathbb{E}_P\left(PC_P^TP^T(\mathbf{1}\mathbf{1}^T)PC_PP^T\right) = \rho_1\epsilon^2\mathbf{R_1} \preceq \rho_1\epsilon^2 I. \tag{3.22}
$$

The following result summarizes Lemmas 3.3, 3.4, and 3.5, using the assumption $n\epsilon \le 1$ from (1.14) to simplify some terms.

Theorem 3.6. *Suppose that* (1.14) *holds. We have*

(3.23a)
$$(1-\delta)^{-2}\mathbb{E}_P\,(PC_P^TC_PP^T) \preceq (1+\rho_1\epsilon^2)I + (1+\rho_1\epsilon^2)\mathbf{1}\mathbf{1}^T + \rho_1\epsilon n^{-1/2}(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T),$$

(3.23b)
$$(1-\delta)^{-2}\mathbb{E}_P\,(PC_P^TP^T\mathbf{1}\mathbf{1}^TPC_PP^T) \preceq \rho_1\epsilon^2 I,$$

(3.23c)
$$(1-\delta)^{-2}\mathbb{E}\,(PC_P^TP^TDPC_PP^T) \preceq D + (2\epsilon+\rho_1\epsilon^2)I + (d_{av}+\rho_1\epsilon^2)\mathbf{1}\mathbf{1}^T + (\rho_1 n^{-1/2}+\rho_1\epsilon^2)(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T)$$

(3.23d)
$$(1-\delta)^{-2}\mathbb{E}_P\,(PC_P^TP^T(\mathbf{1}\mathbf{r_1}^T + \mathbf{r_1}\mathbf{1}^T)PC_PP^T) \preceq \rho_1\epsilon I + \rho_1\epsilon n^{-1/2}\mathbf{1}\mathbf{1}^T.$$

Proof. The first result (3.23a) follows immediately from Lemma 3.3 when we note that $\epsilon^2 = n^{-1/2}(\epsilon n^{-1/2})(\epsilon n) \le \epsilon n^{-1/2}$. The bound (3.23b) is immediate from (3.22) in Lemma 3.5. Lemma 3.4 immediately yields (3.23c). For (3.23d), we use $\epsilon n \le 1$ and $\epsilon^2 n^{1/2} = \epsilon(n\epsilon)n^{-1/2} \le \epsilon n^{-1/2}$ to simplify the coefficients of I and $\mathbf{1}\mathbf{1}^T$ in (3.21). □

3.4. **The Four-Term Recurrence and Convergence Bound for RPCD.** In this section we discuss the sequence of $n\times n$ symmetric matrices $\hat{A}_\epsilon^{(t)}$ that dominates the sequence $\bar{A}_\epsilon^{(t)}$ defined in Section 2.2. Using the results of the previous subsection, together with the four-term parametrization of $\hat{A}_\epsilon^{(t)}$ defined in (3.1), we derive a recurrence relationship for the sequence of quadruplets $\{(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)\}_{t=0,1,2,\ldots}$. By finding the rate at which this sequence decreases to zero, we derive a bound on the expected values of f after each epoch of RPCD.

We now show the main result for recurrence of the representation (3.1).

Theorem 3.7. *Suppose that* (1.14) *holds. Consider a nonnegative sequence of quadruplets* $\{(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)\}_{t=0,1,2,\ldots}$ *satisfying*

$$\hat{\eta}_0 = \delta, \quad \hat{\nu}_0 = 1-\delta, \quad \hat{\epsilon}_0 = \epsilon, \quad \hat{\tau}_0 = 0, \tag{3.24}$$

along with the recurrence

$$\begin{bmatrix}\tilde{\eta}_{t+1}\\ \tilde{\nu}_{t+1}\\ \tilde{\epsilon}_{t+1}\\ \tilde{\tau}_{t+1}\end{bmatrix} = (1-\delta)^2\hat{M}\begin{bmatrix}\hat{\eta}_t\\ \hat{\nu}_t\\ \hat{\epsilon}_t\\ \hat{\tau}_t\end{bmatrix}, \quad \begin{bmatrix}\hat{\eta}_{t+1}\\ \hat{\nu}_{t+1}\\ \hat{\epsilon}_{t+1}\\ \hat{\tau}_{t+1}\end{bmatrix} = \begin{bmatrix}\max(\tilde{\eta}_{t+1},0)\\ \max(\tilde{\nu}_{t+1},0)\\ \max(\tilde{\epsilon}_{t+1},0)\\ \max(\tilde{\tau}_{t+1},0)\end{bmatrix}, \tag{3.25}$$

where

$$\hat{M} = \begin{bmatrix}1+\rho_1\epsilon^2 & \rho_1\epsilon^2 & 2\epsilon+\rho_1\epsilon^2 & \rho_1\epsilon\\ 1+\rho_1\epsilon^2 & 0 & d_{av}+\rho_1\epsilon^2 & \rho_1\epsilon n^{-1/2}\\ 0 & 0 & 1 & 0\\ \rho_1\epsilon n^{-1/2} & 0 & \rho_1 n^{-1/2}+\rho_1\epsilon^2 & 0\end{bmatrix}, \tag{3.26}$$

where each ρ_1 represents a positive quantity not much greater than 1 *and independent of n, ϵ, and δ. Then we have for $\hat{A}_\epsilon^{(t)}$ defined by* (3.1) *that $\hat{A}_\epsilon^{(t)} \succeq \bar{A}_\epsilon^{(t)}$ for all t.*

Proof. By definition, we have $\hat{A}_\epsilon^{(0)} \succeq \bar{A}_\epsilon^{(0)}$. Supposing that $\hat{A}_\epsilon^{(t)} \succeq \bar{A}_\epsilon^{(t)}$ for some $t \geq 0$, we have from (2.9) that

$$\mathbb{E}_P(PC_P^T P^T \hat{A}_\epsilon^{(t)} PC_P P^T) \succeq \mathbb{E}_P(PC_P^T P^T \bar{A}_\epsilon^{(t)} PC_P P^T) = \bar{A}_\epsilon^{(t+1)}. \tag{3.27}$$

Analogous to (3.1), we define the following matrix, parametrized by the coefficients $(\tilde{\eta}_{t+1}, \tilde{\nu}_{t+1}, \tilde{\epsilon}_{t+1}, \tilde{\tau}_{t+1})$ defined in (3.25):

$$\tilde{A}_\epsilon^{(t+1)} = \tilde{\eta}_{t+1} I + \tilde{\nu}_{t+1} \mathbf{1}\mathbf{1}^T + \tilde{\epsilon}_{t+1} D + \tilde{\tau}_{t+1}(\mathbf{1}\mathbf{r_1}^T + \mathbf{r_1}\mathbf{1}^T). \tag{3.28}$$

Since $(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t) \geq 0$, we can use Theorem 3.6 to ensure that

$$\mathbb{E}_P(PC_P^T P^T \hat{A}_\epsilon^{(t)} PC_P P^T) \preceq \tilde{A}_\epsilon^{(t+1)}. \tag{3.29}$$

A little more explanation is needed here. Because the matrices I, $\mathbf{1}\mathbf{1}^T$, and D (the coefficients of $\hat{\eta}_t$, $\hat{\nu}_t$, and $\hat{\epsilon}_t$, respectively) are positive semidefinite, we can use the upper bounds in (3.23a), (3.23b), and (3.23c) to derive the $\preceq$ relationship. The coefficient of $\hat{\tau}_t$ may not be positive definite, but since $\hat{\tau}_t \geq 0$, we can still use the bound (3.23d) to establish the $\preceq$ relationship. Moreover, we can assume that $\tilde{\tau}_{t+1} \geq 0$, by replacing $\mathbf{r_1}$ by $-\mathbf{r_1}$ in the representation (3.28) if necessary. Thus, from (3.25), we have

$$\tilde{A}_\epsilon^{(t+1)} - \hat{A}_\epsilon^{(t+1)} = \min(\tilde{\eta}_{t+1}, 0) I + \min(\tilde{\nu}_{t+1}, 0)\mathbf{1}\mathbf{1}^T + \min(\tilde{\epsilon}_{t+1}, 0) D \preceq 0.$$

By combining this expression with (3.27) and (3.29), we obtain $\bar{A}_\epsilon^{(t+1)} \preceq \hat{A}_\epsilon^{(t+1)}$, as required. □

We now analyze the decay of the sequence of quadruplets $(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)$ generated by this recursion. For purposes of this analysis, we assume that all quantities ρ_1 that appear in the matrix $\hat{M}$ defined by (3.26) are bounded in magnitude by constant $\bar{\rho}$. From this constant, we define

$$\hat{\rho} := 3.05 + 2.1\bar{\rho} + .6\bar{\rho}^2 + .01\bar{\rho}^3. \tag{3.30}$$

We place further restrictions on the allowable regime for values of n, ϵ, and δ, in addition to those in (1.14). Specifically, we require

$$\hat{\rho}\epsilon^2 \leq \frac{1}{2}\delta, \quad n \geq 5. \tag{3.31}$$

As immediate consequences of these bounds, in combination with (1.14) and (3.30), we have

$$\hat{\rho}\epsilon \leq \frac{1}{2}\frac{\delta}{\epsilon} \leq \frac{1}{2}, \tag{3.32a}$$

$$\epsilon \leq \frac{1}{n} \leq .2 \Rightarrow \delta \leq \epsilon \leq .2 \tag{3.32b}$$

$$n^{-1/2} \leq .5, \quad n^{-3/2} \leq .1, \quad n^{-2} \leq .04, \tag{3.32c}$$

$$\bar{\rho}\epsilon^2 \leq \frac{1}{2}\hat{\rho}\epsilon^2 \leq \frac{1}{4}\delta \leq .05, \tag{3.32d}$$

$$\epsilon^2 = \frac{(n\epsilon)^2}{n^2} \leq \frac{1}{n^2} \leq .04. \tag{3.32e}$$

Other useful consequences of (1.14) and (3.31), used repeatedly below, are as follows

$$(1-\delta)^2(1+\hat{\rho}\epsilon^2) \le (1-\delta)^2(1+\tfrac{1}{2}\delta) \le (1-1.4\delta), \tag{3.33a}$$

$$(1-\delta)^2 = (1-2\delta+\delta^2) \le (1-2\delta+\delta/n) \le (1-1.8\delta). \tag{3.33b}$$

We now define two sequences that can be used to bound in norm the quadruplets $(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)$. These are

$$\bar{\eta}_t := 1.5\hat{\rho}(1-1.4\delta)^t t\delta, \tag{3.34a}$$

$$\bar{\epsilon}_t := (1-1.8\delta)^t \epsilon. \tag{3.34b}$$

We note immediately by combining with (3.33a) and (3.33b) that

$$\bar{\eta}_{t-1} \le \frac{\bar{\eta}_t}{(1-1.4\delta)} \Rightarrow (1-\delta)^2\bar{\eta}_{t-1} \le \bar{\eta}_t, \tag{3.35a}$$

$$\bar{\epsilon}_{t-1} = \frac{\bar{\epsilon}_t}{(1-1.8\delta)} \Rightarrow (1-\delta)^2\bar{\epsilon}_{t-1} \le \bar{\epsilon}_t. \tag{3.35b}$$

The following lemma details how the sequence of quadruplets $(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)$ is bounded in terms of the quantities in (3.35). Its proof appears in Appendix C.

Lemma 3.8. *Assume that the conditions* (1.14) *and* (3.31) *hold, and let* $(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)$ *be defined as in Theorem 3.7, and* $\bar{\eta}_t$ *and* $\bar{\epsilon}_t$ *be defined as in* (3.34)*. Then the following bounds hold for all* $t = 1, 2, \ldots$*:*

$$0 \le \hat{\eta}_t \le \bar{\eta}_t, \tag{3.36a}$$

$$0 \le \hat{\epsilon}_t \le \bar{\epsilon}_t, \tag{3.36b}$$

$$0 \le \hat{\tau}_t \le .5\epsilon\bar{\rho}\bar{\eta}_t + .54\bar{\rho}\bar{\epsilon}_t \tag{3.36c}$$

$$\le .1\bar{\rho}\bar{\eta}_t + .54\bar{\rho}\bar{\epsilon}_t, \tag{3.36d}$$

$$0 \le \hat{\nu}_t \le (1.1 + .01\bar{\rho}^2)\bar{\eta}_t + (1.1 + .1\bar{\rho}^2)\bar{\epsilon}_t. \tag{3.36e}$$

We are now ready to prove the main convergence result.

Theorem 3.9. *Suppose that the RPCD version of Algorithm 1 is applied to function* f *defined by* (1.3) *with coefficient matrix satisfying* (1.13)*. Suppose that the quantities* ρ_1 *in each recurrence matrix* $\hat{M}$ *in* (3.25) *are all bounded in magnitude by* $\bar{\rho}$*, and that conditions* (1.14) *and* (3.31) *hold. Then there is a constant* C *such that for all* $t = 1, 2, \ldots,$ *we have*

$$\mathbb{E}_{P_1,P_2,\ldots,P_t} f(x^{tn}) \le C(1-1.4\delta)^t t\epsilon\|x^0\|^2.$$

indicating an asymptotic per-epoch convergence rate approaching $1 - 1.4\delta$*.*

Proof. The proof follows from (2.10) and (3.34) when we use Lemma 3.8, the bound $\|\hat{A}_\epsilon^{(t)}\| \le \bar{C}\|(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)\|$ for some $\bar{C} > 0$, and the bound

$$\|(\hat{\eta}_t, \hat{\nu}_t, \hat{\epsilon}_t, \hat{\tau}_t)\| \le \hat{C}\max(\bar{\eta}_t, \bar{\epsilon}_t) \le \hat{C}(1-1.4\delta)^t t\epsilon,$$

for some $\hat{C} > 0$, where we used $\delta \le \epsilon$ in the last step. The final claim follows by taking the ratio of the bound after $t+1$ and t epochs, which approaches $(1-1.4\delta)$ as $t \to \infty$. □

3.5. **Decrease in the First Iteration.** A behavior of all CD variants that we observe in Figures 1-2 is that the objective value decreases dramatically in the very first iteration of the algorithm. The theorem below shows that this phenomenon can be explained for both RCD and RPCD by using an extension of the analysis in [3, Theorem 3.4]. (Similar reasoning also applies for CCD in most cases, but there is no guarantee, since adversarial examples consisting of particular choices of x^0 can be constructed.) Geometrically, the phenomenon is due to the function (1.3), (1.13) increasing rapidly along just one direction — the all-one direction $\mathbf{1}$ — and more gently in other directions. Thus an exact line search along *any* coordinate search direction will identify a point near the bottom of this multidimensional "trench."

Our result for first-iteration decrease is as follows.

Theorem 3.10. *Consider solving* (1.3) *with the matrix* $A = A_\epsilon$ *defined in* (1.13), *and* $\epsilon \in (0,1)$, *using CCD, RCD, or RPCD with exact line search. Then after a single iteration, we have*

$$f(x^1) \le \frac{1}{2}\sum_{j\neq i}(x_j^0)^2(\delta+\epsilon d_j) + \frac{(1-\delta)\,(\delta+\epsilon)}{2(1+\epsilon)}\left(\sum_{j\neq i} x_j^0\right)^2, \tag{3.37}$$

where $i = i(0,0)$ *is the coordinate chosen for updating in the first iteration. When RCD or RPCD is used, we further have that*

$$\mathbb{E}_i f(x^1) \le \frac{\delta+\epsilon}{2n}\left(n - \frac{\delta+\epsilon}{1+\epsilon}\right)\|x^0\|^2 + \frac{n-2}{n}\frac{(1-\delta)(\delta+\epsilon)}{(1+\epsilon)}\left(\mathbf{1}^T x^0\right)^2. \tag{3.38}$$

Proof. Suppose that $i \in \{1,2,\ldots,n\}$ is the component chosen for updating in the first iteration, which is chosen uniformly at random from $\{1,2,\ldots,n\}$ for RPCD and RCD. After a single step of CD, we have

$$x_i^1 = x_i^0 - \left(x_i^0 + (1-\delta)\sum_{j\neq i}\frac{x_j^0}{1+\epsilon d_i}\right) = -\frac{1-\delta}{1+\epsilon d_i}\left(\sum_{j\neq i} x_j^0\right);$$
$$x_j^1 = x_j^0, \quad \text{for } j \neq i.$$

Thus, from (1.13), we have

$$
\begin{aligned}
f(x^1) &= \frac{1}{2}\delta\|x^1\|^2 + \frac{1}{2}(1-\delta)\left(\sum_{j=1}^n x_j^1\right)^2 + \frac{1}{2}\epsilon\sum_{j=1}^n d_j(x_j^1)^2 \\
&= \frac{1}{2}\delta\left[\sum_{j\neq i}(x_j^0)^2 + \left(\frac{1-\delta}{1+\epsilon d_i}\right)^2\left(\sum_{j\neq i}x_j^0\right)^2\right] \\
&\qquad + \frac{1}{2}(1-\delta)\left[\sum_{j\neq i}x_j^0 - \frac{1-\delta}{1+\epsilon d_i}\sum_{j\neq i}x_j^0\right]^2 \\
&\qquad + \frac{1}{2}\epsilon\left[\sum_{j\neq i}(x_j^0)^2 d_j + \left(\frac{1-\delta}{1+\epsilon d_i}\right)^2 d_i\left(\sum_{j\neq i}x_j^0\right)^2\right] \\
&= \frac{1}{2}\delta\sum_{j\neq i}(x_j^0)^2 + \frac{1}{2}\epsilon\sum_{j\neq i}(x_j^0)^2 d_j \\
&\qquad + \frac{\left(\sum_{j\neq i}x_j^0\right)^2}{2(1+\epsilon d_i)^2}\left[\delta\left(1-\delta\right)^2 + (1-\delta)\left(\epsilon d_i+\delta\right)^2 + \epsilon d_i\left(1-\delta\right)^2\right].
\end{aligned}
\tag{3.39}
$$

Since $d_i \in [0,1]$, $\delta \in (0,1)$, and $\epsilon \in (0,1)$, it can be shown that

$$
\frac{1}{(1+\epsilon d_i)^2} \le \frac{1}{(1+\epsilon)^2}, \quad \frac{\delta+\epsilon d_i}{1+\epsilon d_i} \le \frac{\delta+\epsilon}{1+\epsilon}, \quad \frac{d_i}{(1+\epsilon d_i)^2} \le \frac{1}{(1+\epsilon)^2}.
$$

Thus by substitution into (3.39), we obtain

$$
\begin{aligned}
f(x^1) &\le \frac{1}{2}\delta\sum_{j\neq i}(x_j^0)^2 + \frac{1}{2}\epsilon\sum_{j\neq i}(x_j^0)^2 d_j \\
&\qquad + \frac{\left(\sum_{j\neq i}x_j^0\right)^2}{2(1+\epsilon)^2}\left[\delta(1-\delta)^2 + (1-\delta)(\epsilon+\delta)^2 + \epsilon\left(1-\delta\right)^2\right].
\end{aligned}
$$

Further, by noting that

$$
\begin{aligned}
\delta(1-\delta)^2 + (1-\delta)(\epsilon+\delta)^2 + \epsilon\left(1-\delta\right)^2 &= (1-\delta)\left[(\delta+\epsilon)(1-\delta) + (\epsilon+\delta)^2\right] \\
&= (1-\delta)\left(\delta+\epsilon\right)(1+\epsilon),
\end{aligned}
$$

the desired result (3.37) is obtained.

The result (3.38) is then obtained by noting that $d_i \le 1$ and that

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

whose derivation can be found in the proof of [3, Theorem 3.4]. □

We can compare $f(x^1)$ from this theorem with $f(x^0)$ obtained by substituting into (1.13), which is

$$f(x^0) = \frac{1}{2}\sum_{i=1}^{n}(x_i^0)^2(\delta + \epsilon d_i) + \frac{1}{2}(1-\delta)(\mathbf{1}^T x^0)^2.$$

Note that the second term, which involves $(\mathbf{1}^T x^0)^2$, decreases by a factor of approximately $(\delta + \epsilon)$ in the first iteration, whereas the first term, which involves $\|x^0\|^2$, does not change much from its original value, which is typically already small. For most starting points, the decrease is dramatic.

4. Analysis for CCD and RCD

The analysis for the RCD variant of coordinate descent for (1.3), (1.13) follows from the standard analysis [5]. The modulus of convexity μ is δ, while the maximum coordinate-wise Lipschitz constant for the gradient $L_{\max}$ is $1+\epsilon$. The per-epoch linear rate of expected improvement in f for RCD on A_ϵ is thus

$$\rho_{\mathrm{RCD}} \le \left(1 - \frac{\delta}{n(1+\epsilon)}\right)^n \approx 1 - \delta + \delta\epsilon + O\left(\delta\epsilon^2\right), \tag{4.1}$$

yielding a complexity of $O(|\log \hat{\epsilon}|/(\delta(1-\epsilon)))$ iterations for reaching an $\hat{\epsilon}$-accurate objective. The (slightly tighter) complexity of RCD from [5, Section 4] improves this epoch bound by approximately a factor of 2, to

$$O\left(|\log \hat{\epsilon}|\frac{1+\epsilon+\delta}{2\delta}\right) \quad \text{iterations.} \tag{4.2}$$

That is, the per-epoch convergence rate of a bound on ρ_{RCD} is approximately $1 - 2\delta/(1+\epsilon+\delta)$.

It is also shown in [5] that one can get an improved rate by non-uniform sampling of the coordinates when the coordinate-wise Lipschitz constants are not identical. In particular, for (1.3), (1.13), if the probability that the ith coordinate is sampled is proportional to the value of $(A_\epsilon)_{ii}$, the result in [5] improves ρ_{RCD} to

$$\rho_{\mathrm{RCD}} \le \left(1 - \frac{\delta}{n(1+d_{\mathrm{av}}\epsilon)}\right)^n \approx 1 - \delta + \delta d_{\mathrm{av}}\epsilon + O\left(\delta d_{\mathrm{av}}^2\epsilon^2\right).$$

Since $d_{\mathrm{av}} \in (0,1)$, this improvement is rather insignificant, and the rate is still worse than that of (4.2). (Whether nonuniform sampling can improve the complexity expression (4.2) is unknown.)

For CCD, we note that the iterates have the form

$$x^{\ell n} = C^\ell x^0,$$

where $C = -(L+\Delta)^{-1}L^T$, where $A = L + \Delta + L^T$ is the triangular-diagonal splitting of A (that is, $C = C_I$ from (1.4), (1.5)). Thus

$$f(x^{\ell n}) = \frac{1}{2}(x^0)^T(C^\ell)^T A C^\ell x^0,$$

and the asymptotic behavior of the sequence of function values is governed by $\|C^\ell\|^2$. By Gelfand's formula [2], the asymptotic per-epoch decrease factor is thus

approximately $\rho(C)^2$. Proposition 3.1 of [10] yields an upper bound on the per-epoch decrease factor. Noting that the largest eigenvalue of A_ϵ is bounded above by $n(1-\delta)+\delta+\epsilon$, their bound is as follows:

(4.3)
$$\rho_{\mathrm{CCD}} \le 1 - \max\left\{ \frac{\delta}{n(n(1-\delta)+\delta+\epsilon)}, \frac{\delta}{(n(1-\delta)+\delta+\epsilon)^2(2+\log n/\pi)^2}, \frac{\delta}{n^2} \right\},$$

which is approximately $1-\delta/n^2$ for the ranges of values of δ and ϵ of interest in this paper. The implied iteration complexity guarantee is about a factor of n^2 worse than that for RCD. In our computational experiments, we compare empirical observations of CCD convergence rate with $\rho(C)^2$ rather than with (4.3).

Note that the upper bounds for convergence rates of RCD and CCD are worst-case guarantees. On the problem class (1.13), we show that the convergence rate of RPCD is similar to the bound for RCD, and we see in the next section that both bounds are quite tight in practice. The worst-case bounds on CCD are looser, in the sense that the computational behavior is not quite as poor as these bounds suggest. Nevertheless, comparison of the worst-case bounds correctly foreshadows that relative behavior of the different variants on these problems, seen in Figure 2: CCD is much slower than RCD or RPCD on this class of problems.

5. Computational Results

We report here on some experiments with variants of CD on problems of the form (1.3), (1.13). Fixing $n=100$, we tried different settings of ϵ and δ, and ran the three variants CCD, RCD, and RPCD for many epochs. Results are reported in Tables 1 and 2. We obtain empirical estimates of the per-epoch asymptotic convergence rate by geometrically averaging the rate over the last 10 epochs, tabulating these observations as $\rho_{\mathrm{CCD}}(\delta, \text{observed})$, $\rho_{\mathrm{RCD}}(\delta, \text{observed})$, and $\rho_{\mathrm{RPCD}}(\delta, \text{observed})$. Since we report the difference between these quantities and 1 in the tables, larger numbers correspond to faster rates. (The numbers in the table are reported in scientific notation, with $a(b)$ representing $a \times 10^b$.) As noted in Section 4, we use $\rho(C)^2$ as the theoretical bound on the convergence rate for CCD, while we use ρ_{RCD} from [5, Section 4] (which corresponds to the complexity (4.2)) as the theoretical estimate of the convergence rate for RCD. For RPCD, we used 2δ as a "benchmark" value, corresponding to a per-epoch rate of $1-2\delta$, slightly faster than the $1-1.4\delta$ rate proved in Section 3.

Not all the settings of parameters n, ϵ, and δ in these tables satisfy the conditions (1.14), (3.31) that were assumed in our analysis. We mark with an asterisk those entries for which these conditions are not satisfied. We note that the benchmark rate of $1-2\delta$ continues to hold in regimes beyond the reach of our theory. This accords with the observation that the matrix $\hat{M}$ defined in (3.26) indeed has norm very close to 1, so that behavior of the sequence is governed chiefly by the $(1-\delta)^2$ factor in the recurrence (3.25).

These tables confirm that the empirical performance of RCD and RPCD is quite similar, across a wide range of parameter values, and markedly faster than CCD.

Acknowledgments

We are grateful for the careful reading and penetrating comments of two referees, which resulted in significant improvement of the results.

δ	1.0000 (-03)	3.0000 (-03)	1.0000 (-02)	3.0000 (-02)	1.0000 (-01)
$1-\rho_{\mathrm{CCD}}(\delta,\text{observed})$	3.4122 (-04)	3.3170 (-04)	3.3527 (-04)	6.1266 (-04)	8.1036 (-04)
$1-\rho(C)^2$	5.9018 (-06)	1.7170 (-05)	6.0912 (-05)	1.9453 (-04)	7.5546 (-04)
$1-\rho_{\mathrm{RCD}}(\delta,\text{observed})$	2.6814 (-03)	5.8265 (-03)	2.1983 (-02)	6.8824 (-02)	1.4427 (-01)
$1-\rho_{\mathrm{RCD}}(\delta,\text{predicted})$	1.9940 (-03)	5.9466 (-03)	1.9419 (-02)	5.5047 (-02)	1.5364 (-01)
$1-\rho_{\mathrm{RPCD}}(\delta,\text{observed})$	2.7048 (-03)	6.3637 (-03)	2.1723 (-02)	6.9230 (-02)	2.0842 (-01)
Benchmark 2δ	2.0000 (-03)	6.0000 (-03)	2.0000 (-02)	6.0000 (-02)*	2.0000 (-01)*

Table 1. Comparison of CCD, RPCD, and RCD on the matrix (1.13) with $n = 100$ and $\epsilon = \delta$.

δ	1.0000 (-03)	3.0000 (-03)	1.0000 (-02)	3.0000 (-02)	1.0000 (-01)
$1-\rho_{\mathrm{CCD}}(\delta,\text{observed})$	2.2372 (-04)	3.9800 (-04)	3.3538 (-04)	2.8511 (-04)	7.9319 (-04)
$1-\rho(C)^2$	2.7954 (-05)	5.0165 (-05)	9.7958 (-05)	2.3542 (-04)	7.8096 (-04)
$1-\rho_{\mathrm{RCD}}(\delta,\text{observed})$	2.6143 (-03)	8.6962 (-03)	1.7869 (-02)	5.8402 (-02)	1.4545 (-01)
$1-\rho_{\mathrm{RCD}}(\delta,\text{predicted})$	1.9763 (-03)	5.8634 (-03)	1.9019 (-02)	5.3824 (-02)	1.5364 (-01)
$1-\rho_{\mathrm{RPCD}}(\delta,\text{observed})$	2.8377 (-03)	7.1350 (-03)	2.1157 (-02)	6.6712 (-02)	2.0501 (-01)
Benchmark 2δ	2.0000 (-03)	6.0000 (-03)*	2.0000 (-02)*	6.0000 (-02)*	2.0000 (-01)*

Table 2. Comparison of CCD, RPCD, and RCD on the matrix (1.13) with $n = 100$ and $\epsilon = \sqrt{\delta/10}$.

References

1. A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM Journal on Optimization **23** (2013), no. 4, 2037–2060.
2. I. Gelfand, *Normierte ringe*, Rech. Math. [Mat. Sbornik] **9** (1941), 3–24.
3. C.-p. Lee and S. J. Wright, *Random permutations fix a worst case for cyclic coordinate descent*, IMA Journal on Numerical Analysis **39** (2019), no. 3, 1246–1275.
4. X. Li, T. Zhao, R. Arora, H. Liu, and M. Hong, *On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization*, Journal of Machine Learning Research **18** (2017), no. 1, 6741–6764.
5. Y. E. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization **22** (2012), no. 2, 341–362.
6. B. Recht and C. Ré, *Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences*, Proceedings of the 25th Annual Conference on Learning Theory, vol. 23, 2012, pp. 11.1–11.24.
7. S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, Journal of Machine Learning Research **14** (2013), no. Feb, 567–599.
8. R. Sun and M. Hong, *Improved iteration complexity bounds of cyclic block coordinate descent for convex problems*, Advances in Neural Information Processing Systems, 2015, pp. 1306–1314.
9. R. Sun, Z.-Q. Luo, and Y. Ye, *On the efficiency of random permutation for admm and coordinate descent*, Tech. Report arXiv:1503.06387, 2019, To appear in *Mathematics of Operations Research*.
10. R. Sun and Y. Ye, *Worst-case complexity of cyclic coordinate descent: $O(n^2)$ gap with randomized version*, Mathematical Programming (2019), 1–34, Online first.
11. S. J. Wright, *Computations with coordinate descent methods*, Presentation at Workshop on *Challenges in Optimization for Data Science*, `https://pcombet.math.ncsu.edu/data2015/`, July 2015.
12. ———, *Coordinate descent methods*, Colloquium, Courant Institute of Mathematical Sciences, December 2015.

Appendix A. Invariance of Coordinate Descent under Diagonal Scaling

Coordinate descent applied to quadratics (1.3) with exact line search at each iterate is invariant under symmetric diagonal scalings of A. For any symmetric

positive definite A and nonzero diagonal F, define

$$\tilde{A} = F^{-1}AF^{-1}. \tag{A.1}$$

Note that $\tilde{A}$ is symmetric positive definite. Consider the objective functions (1.3) defined with Hessians A and $\tilde{A}$. For a given x^0, define $\tilde{x}^0 = Fx^0$. The function values match at these points, that is,

$$(\tilde{x}^0)^T\tilde{A}\tilde{x}^0 = (Fx^0)^T\tilde{A}(Fx^0) = (x^0)^TAx^0. \tag{A.2}$$

Considering the iterates generated by Algorithm 1 for the two functions, with α_k defined by exact line searches, and the same choices of coordinates $i(\ell, j)$ at each iteration. Assume that $Fx^t = \tilde{x}^t$ for $t = 1, 2, \ldots, k$. Suppose that coordinate i is chosen at iteration k, the updates are

$$x^{k+1} = x^k - \frac{(Ax^t)_i}{A_{ii}}e_i, \quad \tilde{x}^{k+1} = \tilde{x}^k - \frac{(\tilde{A}\tilde{x}^t)_i}{\tilde{A}_{ii}}e_i.$$

By noting that

$$(\tilde{A}\tilde{x}^t)_i = F_{ii}^{-1}(Ax^t)_i, \quad \tilde{A}_{ii} = F_{ii}^{-2}A_{ii},$$

and using the inductive hypothesis, it is easy to verify that $\tilde{x}^{k+1} = Fx^{k+1}$, as required.

Appendix B. Proofs of Lemmas from Section 3.3

B.1. **Proof of Lemma 3.3.**

Proof. From Lemma 3.2, we have

$$\begin{aligned}
&(1-\delta)^{-2}PC_P^TC_PP^T \\
&\text{(B.1a)} \\
&= P\left[(I - \mathbf{1}e_1^T) + \epsilon(I - \mathbf{1}e_1^T)(-D_P + D_PF) + \delta(F - \mathbf{1}e_2^T) + \epsilon^2(\rho_1\mathbf{1}\mathbf{r_1}^T + \rho_1\mathbf{R_1})\right] \\
&\qquad \left[(I - e_1\mathbf{1}^T) + \epsilon(-D_P + F^TD_P)(I - e_1\mathbf{1}^T) + \delta(F^T - e_2\mathbf{1}^T) + \epsilon^2(\rho_1\mathbf{r_1}\mathbf{1}^T + \rho_1\mathbf{R_1})\right]P^T \\
&\text{(B.1b)} \\
&= \Big\{P(I - \mathbf{1}e_1^T)(I - e_1\mathbf{1}^T)P^T \\
&\qquad + \delta\big[P(F - \mathbf{1}e_2^T)(I - e_1\mathbf{1}^T)P^T + P(I - \mathbf{1}e_1^T)(F^T - e_2\mathbf{1}^T)P^T\big] \\
&\qquad + \epsilon P(I - \mathbf{1}e_1^T)(-2D_P + D_PF + F^TD_P)(I - e_1\mathbf{1}^T)P^T\Big\} \\
&\qquad + \epsilon^2(\rho_1(\mathbf{1}\mathbf{r_1}^T + \mathbf{r_1}\mathbf{1}^T) + \rho_1\mathbf{1}\mathbf{1}^T + \rho_1\mathbf{R_1}).
\end{aligned}$$

(We give further details on the ϵ^2 term below.) For the $O(1)$ term in (B.1b), we have from (3.5) and $e_1^Te_1 = 1$ that

$$\mathbb{E}_P\left(P(I - \mathbf{1}e_1^T)(I - e_1\mathbf{1}^T)P^T\right) = I - \frac{2}{n}\mathbf{1}\mathbf{1}^T + \mathbf{1}\mathbf{1}^T = I + \left(1 - \frac{2}{n}\right)\mathbf{1}\mathbf{1}^T. \tag{B.2}$$

For the first part of the $O(\delta)$ term, we have from (3.5), (3.8), (3.9), and $e_1^T e_2 = 0$ that

$$\begin{aligned}\mathbb{E}_P\left(P(F-\mathbf{1}e_2^T)(I-e_1\mathbf{1}^T)P^T\right) &= \mathbb{E}_P\left(PFP^T - P\mathbf{1}e_2^T P^T\right)\\ &= \mathbb{E}_P\left(PFP^T - \mathbf{1}\left(\frac{1}{n}\mathbf{1}\right)^T\right)\\ &= \frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}I - \frac{1}{n}\mathbf{1}\mathbf{1}^T = -\frac{1}{n}I. \end{aligned} \tag{B.3}$$

(By symmetry, the second part of the $O(\delta)$ term will also have expectation $-\frac{1}{n}I$.)

For the $O(\epsilon)$ term in (B.1b), we have from (2.6), (3.8), (3.9), Lemma 3.1, and the fact that $\mathbb{E}_P e_1^T D_P e_1 = d_{\text{av}}$ that

$$\begin{aligned}&\mathbb{E}_P\left\{P(I-\mathbf{1}e_1^T)(-2D_P + D_P F + F^T D_P)(I-e_1\mathbf{1}^T)P^T\right\}\\ &= \mathbb{E}_P\left\{(P-\mathbf{1}e_1^T)(-2P^T DP + P^T DPF + F^T P^T DP)(P^T - e_1\mathbf{1}^T)\right\}\\ &= -2D + \mathbb{E}_P(DPFP^T + PF^T P^T D)\\ &\qquad - \mathbb{E}_P\left\{\mathbf{1}e_1^T(-2P^T D + P^T DPFP^T) + (-2DP + PF^T P^T DP)e_1\mathbf{1}^T\right\}\\ &\qquad + \mathbb{E}_P(-2e_1^T(P^T DP)e_1)\mathbf{1}\mathbf{1}^T\\ &= -2D + \frac{1}{n}(D(\mathbf{1}\mathbf{1}^T - I) + (\mathbf{1}\mathbf{1}^T - I)D) + \frac{2}{n}(\mathbf{1}\mathbf{1}^T D + D\mathbf{1}\mathbf{1}^T)\\ &\qquad - \frac{1}{n-1}\left[2d_{\text{av}}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T D - \frac{1}{n}D\mathbf{1}\mathbf{1}^T\right] - 2d_{\text{av}}\mathbf{1}\mathbf{1}^T\\ &= -2\left(1+\frac{1}{n}\right)D + \left(\frac{3}{n} + \frac{1}{n(n-1)}\right)\left(d\mathbf{1}^T + \mathbf{1}d^T\right) - 2\left(1+\frac{1}{n-1}\right)d_{\text{av}}\mathbf{1}\mathbf{1}^T.\end{aligned}$$

The lower-order terms in the main result follows by substituting this estimate along with (B.2) and (B.3) into (B.1b).

We now address the ϵ^2 term in (B.1b). Gathering together all terms with coefficients ϵ^2, $\epsilon\delta$, and δ^2 from (B.1a), we have

$$\begin{aligned}&\epsilon^2 P(\rho_1\mathbf{1}\mathbf{r_1}^T + \rho_1\mathbf{R_1}_(I - e_1\mathbf{1}^T)P^T + (\text{transpose})\\ &+ \epsilon^2 P(I-\mathbf{1}e_1^T)(-D_P)(I-F)(I-F^T)(-D_P)(I-e_1\mathbf{1}^T)P^T\\ &+ \epsilon\delta P(I-\mathbf{1}e_1^T)(-D_P)(I-F)(F^T - e_2\mathbf{1}^T)P^T + (\text{transpose})\\ &+ \delta^2 P(F-\mathbf{1}e_2^T)(F^T - e_2\mathbf{1}^T)P^T.\end{aligned}$$

The first term in this expression (and its transpose) is clearly accounted for by the ϵ^2 term in (B.1b). For the other ϵ^2 term, and also the $\epsilon\delta$ terms, we use the facts that $\|F\| = 1$ and $\|D_P(I-F)\| \le \|D_P\|(\|I\| + \|F\|) = \rho_1$, along with $\delta \le \epsilon$, to deduce that these terms too are accounted for by the ϵ^2 term in (B.1b). From $\|F\| = 1$ and $e_2^T e_2 = 1$, we can say the same too for the coefficient of δ^2.

For the final "$\preceq$" claim in the lemma, we use the facts (1.16), $d\mathbf{1}^T = \rho_1 n^{1/2}\mathbf{r_1}\mathbf{1}^T$ (from (3.11)), $d_{\text{av}} \in (0,1]$, and $D \succeq 0$, □

B.2. **Proof of Lemma 3.4.**

Proof. From Lemma 3.2, we have

$$
\begin{aligned}
&(1-\delta)^{-2} P C_P^T D_P C_P P^T \\
&\text{(B.4a)} \\
&= P\left[(I - \mathbf{1}e_1^T) + \epsilon(I - \mathbf{1}e_1^T)(-D_P + D_P F) + \delta(F - \mathbf{1}e_2^T) + \epsilon^2(\rho_1 \mathbf{1}\mathbf{r_1}^T + \rho_1 \mathbf{R_1})\right] D_P \\
&\qquad \left[(I - e_1\mathbf{1}^T) + \epsilon(-D_P + F^T D_P)(I - e_1\mathbf{1}^T) + \delta(F^T - e_2\mathbf{1}^T) + \epsilon^2(\rho_1 \mathbf{r_1}\mathbf{1}^T + \rho_1 \mathbf{R_1})\right] P^T \\
&= P(I - \mathbf{1}e_1^T) D_P (I - e_1\mathbf{1}^T) P^T \\
&\quad + \epsilon P(I - \mathbf{1}e_1^T) D_P(-I + F) D_P (I - e_1\mathbf{1}^T) P^T \\
&\quad + \epsilon P(I - \mathbf{1}e_1^T) D_P(-I + F^T) D_P (I - e_1\mathbf{1}^T) P^T \\
&\quad + \delta P(F - \mathbf{1}e_2^T) D_P (I - e_1\mathbf{1}^T) P^T \\
&\quad + \delta P(I - \mathbf{1}e_1^T) D_P (F^T - e_2\mathbf{1}^T) P^T \\
&\quad + \epsilon^2(\rho_1 \mathbf{1}\mathbf{1}^T + \rho_1(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T) + \rho_1 \mathbf{R_1}) \\
&\text{(B.4b)} \\
&= P(I - \mathbf{1}e_1^T) D_P (I - e_1\mathbf{1}^T) P^T \\
&\quad + \epsilon P(I - \mathbf{1}e_1^T) D_P(-2I + F + F^T) D_P (I - e_1\mathbf{1}^T) P^T \\
&\quad + \delta P(F D_P - \mathbf{1}e_2^T D_P + D_P F^T - D_P e_2 \mathbf{1}^T) P^T \\
&\quad + \epsilon^2(\rho_1 \mathbf{1}\mathbf{1}^T + \rho_1(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T) + \rho_1 \mathbf{R_1}),
\end{aligned}
$$

where we used (3.10a) from Lemma 3.1 along with $e_2^T D_P e_1 = 0$ to simplify the coefficient of δ. (Further justification for the form of the ϵ^2 term appears below.)

For the $O(1)$ term, we have that

$$
\begin{aligned}
P(I - \mathbf{1}e_1^T) D_P (I - e_1\mathbf{1}^T) P^T &= (P - \mathbf{1}e_1^T) D_P (P^T - e_1\mathbf{1}^T) \\
&= D - \mathbf{1}e_1^T P^T D - D P e_1 \mathbf{1}^T + (e_1^T D_P e_1)\mathbf{1}\mathbf{1}^T.
\end{aligned}
$$

Thus from (3.5), we have by taking expectations over P that

$$
\begin{aligned}
\mathbb{E}_P(P(I - \mathbf{1}e_1^T) D_P (I - e_1\mathbf{1}^T) P^T) &= D - \frac{1}{n}\mathbf{1}\mathbf{1}^T D - \frac{1}{n} D \mathbf{1}\mathbf{1}^T + d_{\text{av}}\mathbf{1}\mathbf{1}^T \\
&= D - \frac{1}{n}(\mathbf{1}d^T + d\mathbf{1}^T) + d_{\text{av}}\mathbf{1}\mathbf{1}^T,
\end{aligned}
$$

as required.

For the coefficient of δ, we have

$$
\begin{aligned}
&P(F D_P - \mathbf{1}e_2^T D_P + D_P F^T - D_P e_2 \mathbf{1}^T) P^T \\
&\quad = P F P^T D - \mathbf{1}e_2^T P^T D + D P F^T P^T - D P e_2 \mathbf{1}^T.
\end{aligned}
$$

Taking expectations with respect to D, we have from (3.5) and (3.9) that

$$
\begin{aligned}
&\mathbb{E}_P(P F P^T) D - \mathbf{1}\mathbb{E}_P(e_2^T P^T) D + D\mathbb{E}_P(P F^T P^T) - D\mathbb{E}_P(P e_2)\mathbf{1}^T \\
&\quad = \frac{1}{n}(\mathbf{1}\mathbf{1}^T - I) D - \frac{1}{n}\mathbf{1}\mathbf{1}^T D + \frac{1}{n} D(\mathbf{1}\mathbf{1}^T - I) - \frac{1}{n} D \mathbf{1}\mathbf{1}^T \\
&\quad = -\frac{2}{n} D.
\end{aligned}
$$

For the coefficient of ϵ, we have

$$\begin{aligned}
&P(I-\mathbf{1}e_1^T)P^TDP(-2I+F+F^T)P^TDP(I-e_1\mathbf{1}^T)P^T\\
&\quad= DP(-2I+F+F^T)P^TD\\
&\qquad-\mathbf{1}e_1^TP^TDP(-2I+F+F^T)P^TD-DP(-2I+F+F^T)P^TDPe_1\mathbf{1}^T\\
&\qquad+\left[e_1^TP^TDP(-2I+F+F^T)P^TDPe_1\right]\mathbf{1}\mathbf{1}^T\\
&\quad= DP(-2I+F+F^T)P^TD\\
&\qquad-\mathbf{1}e_1^TP^TDP(-2I+F)P^TD-DP(-2I+F^T)P^TDPe_1\mathbf{1}^T\\
&\qquad-2\left[e_1^TP^TD^2Pe_1\right]\mathbf{1}\mathbf{1}^T, \qquad \text{(B.5)}
\end{aligned}$$

where we used (3.10a) in Lemma 3.1 to eliminate terms that are multiples of $Fe_1 = 0$.

For the first term in (B.5), we have from (3.9) that

$$\begin{aligned}
&\mathbb{E}_P(DP(-2I+F+F^T)P^TD)\\
&\quad= -2D^2+D\mathbb{E}_P(PFP^T)D+D\mathbb{E}_P(PF^TP^T)D\\
&\quad= -2D^2+\frac{2}{n}D(\mathbf{1}\mathbf{1}^T-I)D\\
&\quad= -2\left(1+\frac{1}{n}\right)D^2+\frac{2}{n}dd^T. \qquad \text{(B.6)}
\end{aligned}$$

For the second term in (B.5), we have

$$\begin{aligned}
&-\mathbf{1}\mathbb{E}_P(e_1^TP^TDP(-2I+F)P^TD)\\
&\quad= 2\mathbf{1}\mathbb{E}_P(e_1^TP^T)D^2-\mathbf{1}\mathbb{E}_P(e_1^TP^TDPFP^T)D\\
&\quad= \frac{2}{n}\mathbf{1}\mathbf{1}^TD^2-\frac{1}{n-1}\mathbf{1}\left(d_{\text{av}}\mathbf{1}^T-\frac{1}{n}\mathbf{1}^TD\right)D\\
&\quad= \frac{2}{n}\mathbf{1}\mathbf{1}^TD^2-\frac{d_{\text{av}}}{n-1}\mathbf{1}d^T+\frac{1}{n(n-1)}\mathbf{1}\mathbf{1}^TD^2\\
&\quad= \frac{1}{n}\frac{2n-1}{n-1}\mathbf{1}\mathbf{1}^TD^2-\frac{d_{\text{av}}}{n-1}\mathbf{1}d^T, \qquad \text{(B.7)}
\end{aligned}$$

where we used (3.10b) from Lemma 3.1 and the definition of d_{av} in (3.3). The third term in (B.5) is the transpose of this second term. For the final term in (B.5), we have

$$-2\mathbb{E}_P(e_1^TP^TD^2Pe_1)\mathbf{1}\mathbf{1}^T=-2d_{\text{av},2}\mathbf{1}\mathbf{1}^T. \qquad \text{(B.8)}$$

By substituting (B.6), (B.7), and (B.8) into (B.5), we obtain the required coefficient of ϵ.

We return to verifying the form of the ϵ^2 term in (B.4b). The coefficients of ϵ^2, $\epsilon\delta$, and δ^2 terms from (B.4a) are are follows:

$$\begin{aligned}
&\epsilon^2P(I-\mathbf{1}e_1^T)D_P(\rho_1\mathbf{r_1}\mathbf{1}^T+\rho_1\mathbf{R_1})P^T+\text{(transpose)}\\
&+\epsilon^2P(I-\mathbf{1}e_1^T)(-D_P)(I-F)D_P(I-F^T)(-D_P)(I-e_1\mathbf{1}^T)P^T\\
&+\epsilon\delta P(I-\mathbf{1}e_1^T)(-D_P)(I-F)D_P(F^T-e_2\mathbf{1}^T)P^T+\text{(transpose)}\\
&+\delta^2P(F-\mathbf{1}e_2^T)D_P(F^T-e_2\mathbf{1}^T).
\end{aligned}$$

By making use of the bounds $\|I\| = \|F\| = 1$, $\|D_P\| \le 1$, $\|e_1\| = \|e_2\| = 1$, and $\delta \le \epsilon$, we see that this expression is accounted for by the coefficient of ϵ^2 in (B.4b).

For the final "$\preceq$" relationship, we use $dd^T \preceq nI$ to obtain $(2/n)\epsilon dd^T \preceq 2\epsilon I$, $\mathbf{1}\mathbf{1}^T D^2 = n^{1/2}\mathbf{1}\mathbf{r_1}^T$ to bound the terms with $\mathbf{1}\mathbf{1}^T D^2$ (and similarly for $D^2\mathbf{1}\mathbf{1}^T$), $D^2 \succeq 0$, $\mathbf{1}d^T = n^{1/2}\mathbf{1}\mathbf{r_1}^T$ to obtain $-(1/n)\mathbf{1}d^T \preceq n^{-1/2}\mathbf{1}\mathbf{r_1}^T$, and $\mathbf{R_1} \preceq I$. □

B.3. **Proof of Lemma 3.5.**

Proof. We have

$$PC_P^T P^T(\mathbf{1}v^T + v\mathbf{1}^T)PC_P P^T = PC_P^T P^T \mathbf{1}v^T PC_P P^T + \text{(transpose)},$$

where we use (transpose) to denote the transpose of the explicitly stated terms. From Lemma 3.2 and $P^T\mathbf{1} = \mathbf{1}$, we have

$$(1-\delta)^{-2} PC_P^T P^T \mathbf{1}v^T PC_P P^T$$

(B.9a)

$$= P\left[(I - \mathbf{1}e_1^T) + \epsilon(I - \mathbf{1}e_1^T)(-D_P + D_P F) + \delta(F - \mathbf{1}e_2^T) + \epsilon^2(\rho_1 \mathbf{1}\mathbf{r_1}^T + \rho_1\mathbf{R_1})\right]\mathbf{1}v^T P$$
$$\left[(I - e_1\mathbf{1}^T) + \epsilon(-D_P + F^T D_P)(I - e_1\mathbf{1}^T) + \delta(F^T - e_2\mathbf{1}^T) + \epsilon^2(\rho_1\mathbf{r_1}\mathbf{1}^T + \rho_1\mathbf{R_1})\right]P^T$$

(B.9b)

$$\begin{aligned}
&= P(I - \mathbf{1}e_1^T)\mathbf{1}v^T P(I - e_1\mathbf{1}^T)P^T \\
&\quad - \epsilon\big\{P(I - \mathbf{1}e_1^T)D_P(I - F)\mathbf{1}v^T P(I - e_1\mathbf{1}^T)P^T \\
&\qquad + P(I - \mathbf{1}e_1^T)\mathbf{1}v^T P(I - F^T)D_P(I - e_1\mathbf{1}^T)P^T\big\} \\
&\quad + \delta\big\{P(F - \mathbf{1}e_2^T)\mathbf{1}v^T P(I - e_1\mathbf{1}^T)P^T \\
&\qquad + P(I - \mathbf{1}e_1^T)\mathbf{1}v^T P(F^T - e_2\mathbf{1}^T)P^T\big\} \\
&\quad + \epsilon^2 n^{1/2}\|v\|(\rho_1\mathbf{R_1} + \rho_1(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T) + \rho_1\mathbf{1}\mathbf{1}^T).
\end{aligned}$$

To derive the remainder term (the coefficient of ϵ^2 in (B.9b)), we need to consider the coefficients of ϵ^2, $\delta\epsilon$, and δ^2 from (B.9a). The coefficient of the ϵ^2 term is

$$\begin{aligned}
&\epsilon^2 P(I - \mathbf{1}e_1^T)\mathbf{1}v^T P(\rho_1\mathbf{r_1}\mathbf{1}^T + \rho_1\mathbf{R_1})P^T \\
&\quad + \epsilon^2 P(\rho_1\mathbf{1}\mathbf{r_1}^T + \rho_1\mathbf{R_1})\mathbf{1}v^T P(I - e_1\mathbf{1}^T)P^T \\
&\quad + \epsilon^2 P(I - \mathbf{1}e_1^T)(-D_P + D_P F)\mathbf{1}v^T(-D_P + F^T D_P)(I - e_1\mathbf{1}^T)P^T.
\end{aligned} \tag{B.10}$$

Using $(I - \mathbf{1}e_1^T)\mathbf{1} = \mathbf{1} - \mathbf{1} = 0$, we see that the first term in this expression vanishes. From (3.7) and (3.8), we have several other identities:

$$(I - F)\mathbf{1} = e_n, \quad (F - \mathbf{1}e_2^T)\mathbf{1} = F\mathbf{1} - \mathbf{1} = -e_n. \tag{B.11}$$

In the third term, we thus have that $(-D_P + D_P F)\mathbf{1} = -D_P(I - F)\mathbf{1} = -D_P e_n = \rho_1\mathbf{r_1}$. We also have that $e_1^T D_P e_n = 0$ and $v^T(-D_P + F^T D_P) = \rho_1\|v\|\mathbf{r_1}^T$. Thus (B.10) becomes

$$\begin{aligned}
&\epsilon^2 P(\rho_1(\mathbf{r_1}^T\mathbf{1})\mathbf{1} + \rho_1\mathbf{R_1}\mathbf{1})v^T(I - Pe_1\mathbf{1}^T) \\
&\quad + \rho_1\epsilon^2\|v\|P(I - \mathbf{1}e_1^T)(-D_P e_n)\mathbf{r_1}^T(I - e_1\mathbf{1}^T)P^T \\
&= \epsilon^2(\rho_1 n^{1/2}\mathbf{1} + \rho_1 n^{1/2}\mathbf{r_1})(v^T - \|v\|\rho_1\mathbf{1}^T) \\
&\quad + \rho_1\epsilon^2\|v\|\mathbf{r_1}\mathbf{r_1}^T(I - e_1\mathbf{1}^T)P^T \\
&= \epsilon^2 n^{1/2}\|v\|(\rho_1\mathbf{R_1} + \rho_1(\mathbf{1}\mathbf{r_1}^T + \mathbf{r_1}\mathbf{1}^T) + \rho_1\mathbf{1}\mathbf{1}^T),
\end{aligned}$$

which is accounted for by the ϵ^2 term in (B.9b). We turn next to the coefficient of $\epsilon\delta$ in (B.9a). This term consists of the following expression plus its transpose:

$$\begin{aligned}
&\delta\epsilon P(I-\mathbf{1}e_1^T)(-D_P)(I-F)\mathbf{1}v^TP(F^T-e_2\mathbf{1}^T)P^T\\
&=\delta\epsilon(P-\mathbf{1}e_1^T)(-D_P)e_nv^TP(F^T-e_2\mathbf{1}^T)P^T \quad \text{from (B.11)}\\
&=\delta\epsilon(PD_Pe_n)v^TP(F^T-e_2\mathbf{1}^T)P^T \quad \text{since } e_1^TD_Pe_n=0\\
&=\delta\epsilon\mathbf{r_1}(\rho_1\|v\|\mathbf{r_1}^T-\rho_1\|v\|\mathbf{1}^T).
\end{aligned}$$

Because $\delta\le\epsilon$, this term (plus its transpose) can also be accounted for by the ϵ^2 term in (B.9b). For the coefficient of δ^2 in (B.9a), we have, using (B.11) again,

$$\begin{aligned}
&\delta^2P(F-\mathbf{1}e_2^T)\mathbf{1}v^TP(F^T-e_2\mathbf{1}^T)P^T\\
&=-\delta^2Pe_nv^TP(F^T-e_2\mathbf{1}^T)P^T=\delta^2\mathbf{r_1}\|v\|(\rho_1\mathbf{r_1}^T+\rho_1\mathbf{1}^T)=\delta^2\|v\|(\rho_1\mathbf{R_1}+\rho_1\mathbf{r_1}\mathbf{1}^T),
\end{aligned}$$

which can also be absorbed into the ϵ^2 term in (B.9b).

Returning to the lower-order terms in (B.9b), we use again the fact that $(I-\mathbf{1}e_1^T)\mathbf{1}=0$ to eliminate the $O(1)$ term, and also one of the two terms in the coefficients of both ϵ and δ. We thus obtain

$$\begin{aligned}
&(1-\delta)^{-2}PC_P^TP^T\mathbf{1}v^TPC_PP^T\\
&=-\epsilon\big\{P(I-\mathbf{1}e_1^T)D_P(I-F)\mathbf{1}v^TP(I-e_1\mathbf{1}^T)P^T\big\}\\
&\qquad+\delta\big\{P(F-\mathbf{1}e_2^T)\mathbf{1}v^TP(I-e_1\mathbf{1}^T)P^T\big\}\\
&\qquad+\epsilon^2n^{1/2}\|v\|(\rho_1\mathbf{R_1}+\rho_1(\mathbf{r_1}\mathbf{1}^T+\mathbf{1}\mathbf{r_1}^T)+\rho_1\mathbf{1}\mathbf{1}^T).
\end{aligned}\tag{B.12}$$

Additionally, we have from (2.6) that

$$\begin{aligned}
P(I-e_1\mathbf{1}^T)P^T&=I-(Pe_1)\mathbf{1}^T\\
P(I-\mathbf{1}e_1^T)D_P&=P(I-\mathbf{1}e_1^T)P^TDP=(I-\mathbf{1}(Pe_1)^T)DP.
\end{aligned}$$

By substituting these identities into (B.12), we obtain

$$\begin{aligned}
&(1-\delta)^{-2}PC_P^TP^T\mathbf{1}v^TPC_PP^T\\
&=-\epsilon\big\{(I-\mathbf{1}(Pe_1)^T)DPe_nv^T(I-(Pe_1)\mathbf{1}^T)\big\}\\
&\qquad-\delta\big\{(Pe_n)v^T(I-(Pe_1)\mathbf{1}^T)\big\}\\
&\qquad+\epsilon^2n^{1/2}\|v\|(\rho_1\mathbf{R_1}+\rho_1(\mathbf{1}\mathbf{r_1}^T+\mathbf{r_1}\mathbf{1}^T)+\rho_1\mathbf{1}\mathbf{1}^T)\\
&=-\epsilon\big\{D(Pe_n)v^T(I-(Pe_1)\mathbf{1}^T)\big\}-\delta\big\{(Pe_n)v^T(I-(Pe_1)\mathbf{1}^T)\big\}\\
&\qquad+\epsilon^2n^{1/2}\|v\|(\rho_1\mathbf{R_1}+\rho_1(\mathbf{1}\mathbf{r_1}^T+\mathbf{r_1}\mathbf{1}^T)+\rho_1\mathbf{1}\mathbf{1}^T)
\end{aligned}\tag{B.13}$$

where the second equality follows from

$$(Pe_1)^TD(Pe_n)=e_1^TD_Pe_n=0,$$

since D_P is diagonal for all P.

Taking expectations, we have for the coefficient of $(-\delta)$ in (B.13) that

$$\begin{aligned}
\mathbb{E}_P\left((Pe_n)v^T(I-(Pe_1)\mathbf{1}^T)\right) &= \mathbb{E}_P\,(Pe_n)v^T - \left[\mathbb{E}_P\,(Pe_n)(v^TPe_1)\right]\mathbf{1}^T \\
&= \frac{1}{n}\mathbf{1}v^T - \frac{1}{n(n-1)}\left(\sum_{\substack{i,j=1\\ i\neq j}}^{n} v_ie_j\right)\mathbf{1}^T \\
&= \frac{1}{n}\mathbf{1}v^T - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} v_i\sum_{\substack{j=1\\ i\neq j}}^{n} e_j\right)\mathbf{1}^T \\
&= \frac{1}{n}\mathbf{1}v^T - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} v_i(\mathbf{1}-e_i)\right)\mathbf{1}^T \\
&= \frac{1}{n}\mathbf{1}v^T - \frac{1}{n(n-1)}\left((\mathbf{1}^Tv)\mathbf{1}\mathbf{1}^T - v\mathbf{1}^T)\right) \\
&= \frac{1}{n}\mathbf{1}v^T - \frac{(\mathbf{1}^Tv)}{n(n-1)}\mathbf{1}\mathbf{1}^T + \frac{1}{n(n-1)}v\mathbf{1}^T, \qquad \text{(B.14)}
\end{aligned}$$

where the second equality is from a conditional expectation over permutation matrices P such that $Pe_1 = j$ and $Pe_n = i$, for all $i, j = 1, 2, \ldots, n$ with $i \neq j$. By combining (B.14) with its transpose, we obtain the full coefficient of $(-\delta)$ in (3.20), which is

$$\left(\frac{1}{n} + \frac{1}{n(n-1)}\right)(\mathbf{1}v^T + v\mathbf{1}^T) - \frac{2(\mathbf{1}^Tv)}{n(n-1)}\mathbf{1}\mathbf{1}^T = \frac{1}{n-1}(\mathbf{1}v^T + v\mathbf{1}^T) - \frac{2(\mathbf{1}^Tv)}{n(n-1)}\mathbf{1}\mathbf{1}^T.$$

This verifies the $O(\delta)$ term in (3.20).

We note that the coefficient of $(-\epsilon)$ in (B.13) is the same as the coefficient of $(-\delta)$, except for being multiplied from the left by D, which is independent of P. Thus the expectation of this term is simply (B.14), multiplied from the left by D, that is,

$$\frac{1}{n}D\mathbf{1}v^T - \frac{\mathbf{1}^Tv}{n(n-1)}(D\mathbf{1})\mathbf{1}^T + \frac{1}{n(n-1)}Dv\mathbf{1}^T = \frac{1}{n}dv^T - \frac{\mathbf{1}^Tv}{n(n-1)}d\mathbf{1}^T + \frac{1}{n(n-1)}Dv\mathbf{1}^T.$$

We obtain the full coefficient of $(-\epsilon)$ in (3.20) by adding this quantity to its transpose, to obtain

$$\frac{1}{n}(dv^T + vd^T) - \frac{\mathbf{1}^Tv}{n(n-1)}(d\mathbf{1}^T + \mathbf{1}d^T) + \frac{1}{n(n-1)}(Dv\mathbf{1}^T + \mathbf{1}v^TD),$$

as required.

For the "$\preceq$" result (3.21), we use

$$\|dv^T\| = n^{1/2}\rho_1\|v\|, \quad \|d\| \leq n^{1/2}, \quad |\mathbf{1}^Tv| \leq n^{1/2}\|v\|, \quad \|Dv\| \leq \|d\|,$$

together with $\delta \leq \epsilon$ from (1.14). We also use $\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T = n^{1/2}\mathbf{R_1}$ and $-I \preceq \mathbf{R_1} \preceq I$ for symmetric $\mathbf{R_1}$ to absorb the $\epsilon^2(\mathbf{r_1}\mathbf{1}^T + \mathbf{1}\mathbf{r_1}^T)$ term in (3.20) into the I term in (3.21).

For (3.22), we have

$$
\begin{aligned}
&(1-\delta)^{-1}PC_P^TP^T\mathbf{1}\\
&=(1-\delta)^{-1}PC_P^T\mathbf{1}\\
&=P\left[(I-\mathbf{1}e_1^T)+\epsilon(I-\mathbf{1}e_1^T)D_P(-I+F)+\delta(F-\mathbf{1}e_2^T)+\epsilon^2(\rho_1\mathbf{1}\mathbf{r_1}^T+\rho_1\mathbf{R_1})\right]\mathbf{1}\\
&=P\left[-\epsilon(I-\mathbf{1}e_1^T)D_Pe_n-\delta e_n+\epsilon^2n^{1/2}(\rho_1\mathbf{1}+\rho_1\mathbf{r_1})\right]\\
&=P\left[-\epsilon D_Pe_n-\delta e_n+\epsilon^2n^{1/2}(\rho_1\mathbf{1}+\rho_1\mathbf{r_1})\right],
\end{aligned}
\tag{B.15}
$$

where we used the following identities for the third equality:

$$
\begin{aligned}
&(I-\mathbf{1}e_1^T)\mathbf{1}=0,\quad (-I+F)\mathbf{1}=-e_n,\quad (F-\mathbf{1}e_2^T)\mathbf{1}=-e_n,\\
&\mathbf{r_1}^T\mathbf{1}\le n^{1/2},\quad \mathbf{R_1}\mathbf{1}=\rho_1n^{1/2}\mathbf{r_1},
\end{aligned}
$$

and $e_1^TD_Pe_n=0$ for the fourth equality. By substituting $D_P=P^TDP$ into (B.15), we obtain

$$
(1-\delta)^{-1}PC_PP^T\mathbf{1}=-\epsilon DPe_n-\delta Pe_n+\epsilon^2n^{1/2}(\rho_1\mathbf{1}+\rho_1\mathbf{r_1}).
$$

By taking the outer product of this vector with itself, and using $\delta\le\epsilon$, we obtain

$$
\begin{aligned}
&(1-\delta)^{-2}PC_P^TP^T\mathbf{1}\mathbf{1}^TPC_PP^T\\
&=\epsilon^2D(Pe_n)(Pe_n)^TD+\delta\epsilon[D(Pe_n)(Pe_n)^T+(Pe_n)(Pe_n)^TD]+\delta^2(Pe_n)(Pe_n)^T\\
&\qquad+\epsilon^3n^{1/2}(\rho_1(\mathbf{1}\mathbf{r_1}^T+\mathbf{r_1}\mathbf{1}^T)+\rho_1\mathbf{R_1}),
\end{aligned}
$$

where in the remainder term we used $\|Pe_n\|=1$ and $\|D\|\le 1$. By taking expectations over P, and using $\mathbb{E}_P(Pe_n)(Pe_n)^T=n^{-1}I$, we obtain

$$
\begin{aligned}
&(1-\delta)^{-2}\mathbb{E}_P(PC_P^TP^T\mathbf{1}\mathbf{1}^TPC_PP^T)\\
&=n^{-1}\epsilon^2D^2+2n^{-1}\epsilon\delta D+n^{-1}\delta^2I+\epsilon^3n^{1/2}(\rho_1(\mathbf{1}\mathbf{r_1}^T+\mathbf{r_1}\mathbf{1}^T)+\rho_1\mathbf{R_1})\\
&=\rho_1\epsilon^2n^{-1}\mathbf{R_1}+\rho_1\epsilon^3n\mathbf{R_1}=\rho_1\epsilon^2\mathbf{R_1},
\end{aligned}
$$

where we used (1.14) in the last expression to deduce that $\epsilon^3n\le\epsilon^2$.

□

Appendix C. Proof of Lemma 3.8

Proof. Note first that $\hat{\eta}_t$, $\hat{\nu}_t$, $\hat{\epsilon}_t$, and $\hat{\tau}_t$ are all nonnegative by definition. In this proof, we use repeatedly that they can be bounded by $|\tilde{\eta}_t|$, $|\tilde{\nu}_t|$, $|\tilde{\epsilon}_t|$, and $|\tilde{\tau}_t|$, respectively, though the $|\cdot|$ are unnecessary in the case of $\tilde{\epsilon}_t$ (since its exact value can be determined trivially from (3.25)) and in the case of $\tilde{\tau}_t$ (which can be assumed without loss of generality to be nonnegative, as mentioned in the proof of Theorem 3.7).

The proof is by induction on t. We show first that the bounds (3.36) hold for $t=1$.

We have from (3.33b) and the obvious property $\hat{\epsilon}_t=(1-\delta)^{2t}\epsilon$ from (3.25) that

$$
\hat{\epsilon}_1=(1-\delta)^2\epsilon\le(1-1.8\delta)\epsilon=\bar{\epsilon}_1,
$$

verifying (3.36b) for $t = 1$. For (3.36a), we have from (3.25) with $t = 0$, using the initial values (3.24) and the bounds in (3.32) that that

$$\begin{aligned}(1-\delta)^{-2}\hat{\eta}_1 \le (1-\delta)^{-2}|\tilde{\eta}_1| &\le (1+\bar{\rho}\epsilon^2)\delta + \bar{\rho}\epsilon^2(1-\delta) + (2\epsilon + \bar{\rho}\epsilon^2)\epsilon \\ &= \delta + (2 + \bar{\rho} + \bar{\rho}\epsilon)\,\epsilon^2 \\ &\le \delta + \hat{\rho}\epsilon^2 \le 3\delta.\end{aligned}$$

It follows from $\hat{\rho} \ge 3$ and $\delta \le .2$ (see (3.32b)) that

$$\hat{\eta}_1 \le 3(1-\delta)^2\delta \le 3(1-1.4\delta)\delta \le 1.5\hat{\rho}(1-1.4\delta)\delta = \bar{\eta}_1,$$

verifying (3.36a) for $t = 1$. For (3.36c) with $t = 1$, we have

$$\begin{aligned}\hat{\tau}_1 = \tilde{\tau}_1 &\le (1-\delta)^2\bar{\rho}\epsilon n^{-1/2}\delta + (1-\delta)^2(\bar{\rho}n^{-1/2} + \bar{\rho}\epsilon^2)\epsilon \\ &\le (1-1.4\delta)(.5)\bar{\rho}\epsilon\delta + (1-1.8\delta)(.5\bar{\rho} + .04\bar{\rho})\epsilon,\end{aligned}$$

where for the second inequality we used $n^{-1/2} \le .5$ and $\epsilon^2 \le .04$. Continuing, we use $\bar{\eta}_1 \ge 4(1-1.4\delta)\delta$ and $\bar{\epsilon}_1 = (1-1.8\delta)\epsilon$ to write

$$\hat{\tau}_1 \le \frac{1}{8}\bar{\rho}\epsilon\bar{\eta}_1 + .54\bar{\rho}\bar{\epsilon}_1, \tag{C.1}$$

which suffices to prove (3.36c) for $t = 1$. For (3.36d), we simply use $\epsilon \le .2$ from (3.32).

For (3.36e) with $t = 1$, we have from (3.25) and (3.24), using again $\epsilon^2 \le .04$ from (3.32), as well as $d_{\text{av}} \le 1$ from (3.3) that

$$\begin{aligned}\hat{\nu}_t \le |\tilde{\nu}_t| &\le (1-\delta)^2(1+\bar{\rho}\epsilon^2)\delta + (1-\delta)^2(d_{\text{av}} + \bar{\rho}\epsilon^2)\epsilon \\ &\le (1-\delta)^2(1+\bar{\rho}\epsilon^2)\delta + (1-\delta)^2(1+.04\bar{\rho})\epsilon \\ &\le (1-1.4\delta)\hat{\rho}\delta + (1-1.8\delta)(1+.04\bar{\rho})\epsilon \\ &\le \bar{\eta}_1 + (1+.04\bar{\rho})\bar{\epsilon}_1,\end{aligned}$$

which suffices to demonstrate (3.36e) for $t = 1$.

Assuming now that (3.36) holds for some $t \ge 1$, we prove that the bounds holds for $t+1$ as well. We start with (3.36c). It follows from (3.25) that

$$\begin{aligned}(1-\delta)^{-2}\hat{\tau}_{t+1} \le (1-\delta)^{-2}|\tilde{\tau}_{t+1}| &\le \bar{\rho}\epsilon n^{-1/2}|\hat{\eta}_t| + (\bar{\rho}n^{-1/2} + \bar{\rho}\epsilon^2)|\hat{\epsilon}_t| \\ &\le .5\bar{\rho}\epsilon\bar{\eta}_t + (.5\bar{\rho} + .04\bar{\rho})\bar{\epsilon}_t \\ &\le .5\bar{\rho}\epsilon\bar{\eta}_t + .54\bar{\rho}\bar{\epsilon}_t.\end{aligned}$$

It then follows from (3.35) that

$$\hat{\tau}_{t+1} \le .5\bar{\rho}\epsilon\bar{\eta}_{t+1} + .54\bar{\rho}\bar{\epsilon}_{t+1},$$

as required. As earlier, (3.36d) follows immediately when we note that $\epsilon \le .2$, from (3.32).

For (3.36e), we have

$$\begin{aligned}(1-\delta)^{-2}\hat{\nu}_{t+1} &\le (1+\bar{\rho}\epsilon^2)\bar{\eta}_t + (d_{\text{av}} + \bar{\rho}\epsilon^2)\bar{\epsilon}_t + \bar{\rho}\epsilon n^{-1/2}|\hat{\tau}_t| \\ &\le (1+\bar{\rho}\epsilon^2 + (\bar{\rho}\epsilon n^{-1/2})(.1)\bar{\rho})\bar{\eta}_t + (d_{\text{av}} + \bar{\rho}\epsilon^2 + (\bar{\rho}\epsilon n^{-1/2})(.54)\bar{\rho})\bar{\epsilon}_t,\end{aligned}$$

where we used (3.36d) for the second inequality. Using now the bounds $\bar{\rho}\epsilon^2 \le .05$ and $n^{-1/2} \le .5$ (from (3.32)), $d_{\text{av}} \le 1$, and $\epsilon n^{-1/2} = (\epsilon n)n^{-3/2} \le n^{-3/2} \le .1$, we have

$$(1-\delta)^{-2}\hat{\nu}_{t+1} \le (1.1 + .01\bar{\rho}^2)\bar{\eta}_t + (1.1 + .1\bar{\rho}^2)\bar{\epsilon}_t,$$

so that

$$\hat{\nu}_{t+1} \le (1.1 + .01\bar{\rho}^2)\bar{\eta}_{t+1} + (1.1 + .1\bar{\rho}^2)\bar{\epsilon}_{t+1},$$

as required.

The proof for (3.36b) is trivial, since

$$\hat{\epsilon}_{t+1} = (1-\delta)^{2(t+1)}\epsilon = (1-\delta)^2\hat{\epsilon}_t \le (1-1.8\delta)\hat{\epsilon}_t \le (1-1.8\delta)\bar{\epsilon}_t = \bar{\epsilon}_{t+1}.$$

We now prove (3.36a) for t replaced by $t+1$. We have, substituting from the other formulas in (3.36), and using the bounds in (3.32), that

$$\begin{aligned}
(1-\delta)^{-2}\hat{\eta}_{t+1} &\le (1+\bar{\rho}\epsilon^2)\bar{\eta}_t + \bar{\rho}\epsilon^2\bar{\nu}_t + (2\epsilon + \bar{\rho}\epsilon^2)\bar{\epsilon}_t + \bar{\rho}\epsilon|\hat{\tau}_t| \\
&\le (1+\bar{\rho}\epsilon^2)\bar{\eta}_t + \bar{\rho}\epsilon^2[(1.1+.01\bar{\rho}^2)\bar{\eta}_t + (1.1+.1\bar{\rho}^2)\bar{\epsilon}_t] \\
&\qquad + (2\epsilon + \bar{\rho}\epsilon^2)\bar{\epsilon}_t + \bar{\rho}\epsilon[.5\epsilon\bar{\rho}\bar{\eta}_t + .54\bar{\rho}\bar{\epsilon}_t] \\
&\le [1 + \bar{\rho}\epsilon^2 + \bar{\rho}\epsilon^2(1.1+.01\bar{\rho}^2) + .5\bar{\rho}^2\epsilon^2]\bar{\eta}_t \\
&\qquad + [\bar{\rho}\epsilon^2(1.1+.1\bar{\rho}^2) + 2\epsilon + \bar{\rho}\epsilon^2 + .54\bar{\rho}^2\epsilon]\bar{\epsilon}_t \\
&\le [1 + \epsilon^2(\bar{\rho} + \bar{\rho}(1.1+.01\bar{\rho}^2) + .5\bar{\rho}^2)]\bar{\eta}_t \\
&\qquad + [.5\epsilon(1.1+.1\bar{\rho}^2) + 2\epsilon + .5\epsilon + .54\bar{\rho}^2\epsilon]\bar{\epsilon}_t \\
&\le [1 + \epsilon^2(2.1\bar{\rho} + .5\bar{\rho}^2 + .01\bar{\rho}^3)]\bar{\eta}_t + \epsilon[.55 + .05\bar{\rho}^2 + 2.5 + .54\bar{\rho}^2]\bar{\epsilon}_t \\
&\le (1 + \hat{\rho}\epsilon^2)\bar{\eta}_t + \hat{\rho}\epsilon\bar{\epsilon}_t,
\end{aligned}$$

where we used the definition (3.30) of $\hat{\rho}$ for the final inequality. Thus from (3.33), substituting from (3.34), and using $\epsilon^2 < \delta$ from (1.14), we have

$$\begin{aligned}
\hat{\eta}_{t+1} &\le (1-\delta)^2(1+\hat{\rho}\epsilon^2)\bar{\eta}_t + (1-\delta)^2\hat{\rho}\epsilon\bar{\epsilon}_t \\
&\le (1-1.4\delta)\bar{\eta}_t + (1-1.8\delta)\hat{\rho}\epsilon\bar{\epsilon}_t \\
&\le 1.5\hat{\rho}(1-1.4\delta)^{t+1}t\delta + (1-1.8\delta)^{t+1}\hat{\rho}\epsilon^2 \\
&\le 1.5\hat{\rho}(1-1.4\delta)^{t+1}t\delta + (1-1.4\delta)^{t+1}\hat{\rho}\epsilon^2 \\
&\le (1-1.4\delta)^{t+1}(1.5\hat{\rho}t\delta + \epsilon^2) \\
&\le (1-1.4\delta)^{t+1}(1.5\hat{\rho}t\delta + \delta) \\
&\le (1-1.4\delta)^{t+1}(1.5)\hat{\rho}(t+1)\delta = \bar{\eta}_{t+1},
\end{aligned}$$

as required. This completes the inductive step and hence the proof. □

Computer Sciences Department, University of Wisconsin-Madison, Madison, WI.
E-mail address: swright@cs.wisc.edu

Department of Mathematics, National University of Singapore.
E-mail address: leechingpei@gmail.com