On the Expected Convergence of Randomly Permuted ADMM

Ruoyu Sun * **Zhi-Quan Luo** † **Yinyu Ye** ‡

Abstract

The alternating direction method of multipliers (ADMM) is now widely used in many fields, and its convergence was proved when two blocks of variables are alternately updated. It is computationally beneficial to extend the ADMM directly to the case of a multi-block (multiple variable blocks) convex minimization problem. Unfortunately, such an extension fails to converge even when solving a simple square system of linear equations. In this paper, however, we prove that, if in each step one randomly and independently permutes the updating order of any given number of blocks followed by the regular multiplier update, the method will converge in expectation for solving the square system of linear equations. Our analysis of random permutation will also be of independent interest.

1 Introduction

Consider a convex minimization problem with a separable objective function and linear constraints:

$$\begin{array}{rl} \min & f_1(x_1) + \cdots + f_n(x_n), \\ \text{s.t.} & A_1 x_1 + \cdots + A_n x_n = b, \\ & x_i \in \mathcal{X}_i,\ i = 1, \ldots, n, \end{array} \tag{1}$$

where $A_i \in \mathbb{R}^{N \times d_i}, b \in \mathbb{R}^{N \times 1}, \mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ is a closed convex set, and $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ is a closed convex function, $i = 1, \ldots, n$. Many machine learning and engineering problems can be cast into linearly-constrained optimization problems with two blocks (see [3] for many examples) or more than two blocks (e.g. linear programming, robust principal component analysis, composite regularizers for structured sparsity; see [5, 24] for more examples).

ADMM (Alternating Direction Method of Multipliers) was first proposed in [10] (see also [4, 8]) to solve problem (1) when there are only two blocks (i.e. $n = 2$). In this 2-block case, the augmented Lagrangian function of (1) is

$$\mathcal{L}(x_1, x_2; \mu) = f_1(x_1) + f_2(x_2) - \mu^T (A_1 x_1 + A_2 x_2 - b) + \frac{\beta}{2} \|A_1 x_1 + A_2 x_2 - b\|^2, \tag{2}$$

where μ is the Lagrangian multiplier and $\beta > 0$ is the penalty parameter. Each iteration of ADMM consists of a cyclic update (i.e. Gauss-Seidal type update) of primal variables x_1, x_2 and a dual ascent type update of μ:

$$\begin{cases} x_1^{k+1} = \arg\min_{x_1 \in \mathcal{X}_1} \mathcal{L}(x_1, x_2^k; \mu^k), \\ x_2^{k+1} = \arg\min_{x_2 \in \mathcal{X}_2} \mathcal{L}(x_1^{k+1}, x_2; \mu^k), \\ \mu^{k+1} = \mu^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases} \tag{3}$$

*Ruoyu Sun is with the Department of Management Science and Engineering, School of Engineering, Stanford University, USA. Email: ruoyus@stanford.edu. Most of the work was done while this author was visiting Stanford University as a PhD student of the department of ECE, University of Minnesota.

†Zhi-Quan Luo is with the Chinese University of Hong Kong, Shenzhen, China. He is also affiliated with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: luozq@cuhk.edu.cn.

‡Yinyu Ye is with the Department of Management Science and Engineering, School of Engineering, Stanford University, USA. Email: yyye@stanford.edu.

Due to the separable structure of the objective function, each subproblem only involves one $f_i, i \in \{1,2\}$, thus may be easier to solve. This feature enables the wide application of ADMM in signal processing and statistical learning where the objective function of the problem can usually be decomposed as the sum of the loss function and the regularizer; see [3] for a review. The convergence of 2-block ADMM has been well studied; see [7, 9] for some recent reviews.

It is natural and computationally beneficial to extend the original 2-block ADMM directly to solve the general n-block problem (1):

$$\begin{cases} x_1^{k+1} = \arg\min_{x_1 \in \mathcal{X}_1} \mathcal{L}(x_1, x_2^k, \ldots, x_n^k; \mu^k), \\ \quad\quad \vdots \\ x_n^{k+1} = \arg\min_{x_n \in \mathcal{X}_n} \mathcal{L}(x_1^{k+1}, \ldots, x_{n-1}^{k+1}, x_n; \mu^k), \\ \mu^{k+1} = \mu^k - \beta(A_1 x_1^{k+1} + \cdots + A_n x_n^{k+1} - b), \end{cases} \tag{4}$$

where the augmented Lagrangian function

$$\mathcal{L}(x_1, \ldots, x_n; \mu) = \sum_{i=1}^{n} f_i(x_i) - \mu^T (\sum_i A_i x_i - b) + \frac{\beta}{2} \| \sum_i A_i x_i - b \|^2. \tag{5}$$

The convergence of the direct extension of ADMM to multi-block case had been a long standing open question, until a counter-example was recently given in [5]. More specifically, [5] showed that even for the simplest scenario where the objective function is 0 and the number of blocks is 3, ADMM can be divergent for a certain choice of $A = [A_1, A_2, A_3]$ (in fact, there is a positive measure of A such that ADMM can be divergent). There are several proposals to overcome the drawback (see, e.g., [6, 11–15, 17, 22]), but they either need to restrict the range of original problems being solved, add additional cost in each step of computation, or limit the stepsize in updating the Lagrangian multipliers. These solutions typically slow down the performance of ADMM for solving most practical problems. One may ask whether a "minimal" modification of cyclic multi-block ADMM (4) can lead to convergence.

One of the simplest modifications of (4) is to add randomness to the update order. Randomness in the update order has been very useful in the analysis of block coordinate gradient descent (BCGD) and stochastic gradient descent (SGD). In particular, the known iteration complexity bounds of randomized BCGD [18] and SAG (Stochastic Average Gradient, a variant of SGD) [20] are much better than the known iteration complexity bounds of their cyclic counterparts BCGD [1] and IAG (Incremental Aggregated Gradient) [2], respectively. [1] The iteration complexity bounds for randomized algorithms are usually established for independent randomization (sampling with replacement), while in practice, random permutation (sampling without replacement) has been reported to exhibit faster convergence (e.g. [19, 21, 23]). However, the theoretical analysis for random permutation seems to be very difficult since the picked blocks/components are not independent across iterations. We have tested both randomly permuted and independently randomized versions of ADMM. Interestingly, independently randomized versions can still be divergent, even for solving linear system of equations, while random permutation can make ADMM converge in all experiments we have conducted.

The main result of this paper is to support the above observation: when the objective function is zero and the constraint is a non-singular square linear system of equations, the expected output of randomly permuted ADMM (RP-ADMM) converges to the unique primal-dual optimal solution. Our contributions are two-fold. First, our result shows that RP-ADMM may serve as a simple solution to resolve the divergence issue of cyclic multi-block ADMM. Since multi-block ADMM is one promising candidate of fast algorithms for large-scale linearly constrained problems, we expect RP-ADMM to be one of the major solvers in big data optimization. Second, our result is one of the first direct analysis of random permutation (sampling without replacement) in optimization algorithms. Our proof framework and techniques will be of independent interest and can be used to analyze random permutation in other optimization algorithms.

We restrict to the simple category of solving linear system of equations, instead of the general convex optimization problems, since the counter-example in [5] belongs to this category and this category seems already difficult to handle for ADMM. The difficulty lies in how to proving the spectral

[1] Rigourously speaking, these two bounds are not directly comparable since the result for the randomized version only holds with high probability, while the result for the cyclic version always holds.

radius of the expected update matrix M is less than one. There are two issues: first, there are few mathematical tools to deal with the spectral radius of non-symmetric matrices; second, the entries of M are complicated functions of the entries of $A^T A$ (in fact, n-th order polynomials). To resolve the first issue, we build a relation between the eigenvalues of $M \in \mathbb{R}^{2N\times 2N}$ and the eigenvalues of a symmetric matrix $AQA^T \in \mathbb{R}^{N\times N}$ (see Lemma 1), where Q is the expectation of the inverse of a random matrix. To resolve the second issue, we use mathematical induction to implicitly utilize the relation of the entries of AQA^T and A. The induction analysis requires several techniques, including a three-level symmetrization technique to construct an induction formula that relates Q to its lower dimensional analogs.

Organization. In Section 2, we present RP-ADMM. Two other versions of randomized ADMM are presented in Section 3. In Section 4, we present our main results Theorem 1, Theorem 2 and their proofs. The proofs of the two technical results Lemma 1 and Lemma 2, which are used in the proof of Theorem 2, are provided in the supplement.

Notations. For a matrix X, we denote $X(i,j)$ as the (i,j)-th entry of X, $\text{eig}(X)$ as the set of eigenvalues of X, $\rho(X)$ as the spectral radius of X (i.e. the maximum modulus of the eigenvalues of X), $\|X\|$ as the spectral norm of X, and X^T as the transpose of X. When X is block partitioned, we use $X[i,j]$ to denote the (i,j)-th block of X. When X is a real symmetric matrix, let $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ denote the maximum and minimum eigenvalue of X respectively.

2 Randomly Permuted ADMM

In this section, we first present RP-ADMM (Randomly Permuted ADMM) for solving the optimization problem (1), then we specialize RP-ADMM for solving a linear system of equations.

Define Γ as

$$\Gamma \triangleq \{\sigma \mid \sigma \text{ is a permutation of } \{1,\dots,n\}\}. \tag{6}$$

At each round, we draw a permutation σ of $\{1,\dots,n\}$ uniformly at random from Γ, and update the primal variables in the order of the permutation, followed by updating the dual variables in a usual way. Obviously, all primal and dual variables are updated exactly once at each round. See Algorithm 1 for the details of RP-ADMM. Note that with a little abuse of notation, the function $\mathcal{L}(x_{\sigma(1)}, x_{\sigma(2)},\dots,x_{\sigma(n)};\mu)$ in this algorithm should be understood as $\mathcal{L}(x_1,x_2,\dots,x_n;\mu)$. For example, when $n=3$ and $\sigma=(231)$, $\mathcal{L}(x_{\sigma(1)},x_{\sigma(2)},x_{\sigma(3)};\mu)=\mathcal{L}(x_2,x_3,x_1;\mu)$ should be understood as $\mathcal{L}(x_1,x_2,x_3;\mu)$.

Algorithm 1 n-block Randomly Permuted ADMM (RP-ADMM)

Initialization: $x_i^0 \in \mathbb{R}^{d_i\times 1}, i=1,\dots,n;\ \mu^0 \in \mathbb{R}^{N\times 1}$.
Round k ($k=0,1,2,\dots$):
1) Primal update.
Pick a permutation σ of $\{1,\dots,n\}$ uniformly at random.
For $i=1,\dots,n$, compute $x_{\sigma(i)}^{k+1}$ by

$$x_{\sigma(i)}^{k+1} = \arg\min_{x_{\sigma(i)}\in\mathcal{X}_{\sigma(i)}} \mathcal{L}(x_{\sigma(1)}^{k+1},\dots,x_{\sigma(i-1)}^{k+1},x_{\sigma(i)},x_{\sigma(i+1)}^{k},\dots,x_{\sigma(n)}^{k};\mu^k) \tag{7}$$

2) Dual update. Update the dual variable by

$$\mu^{k+1} = \mu^k - \beta(\sum_{i=1}^n A_i x_i^{k+1} - b). \tag{8}$$

In this paper, we will only consider using Algorithm 1 to solve a square linear system of equations. Consider a special case of (1) where $f_i = 0$, $\mathcal{X}_i = \mathbb{R}^{d_i}, \forall i$ and $N=\sum_i d_i$ (i.e. the constraint is a square system of equations). Then problem (1) becomes

$$\begin{aligned} \min_{x\in\mathbb{R}^N} \quad & 0, \\ \text{s.t.} \quad & A_1x_1+\cdots+A_nx_n = b, \end{aligned} \tag{9}$$

where $A_i \in \mathbb{R}^{N\times d_i}, x_i \in \mathbb{R}^{d_i\times 1}, b \in \mathbb{R}^{N\times 1}$. Solving this feasibility problem (with 0 being the objective function) is equivalent to solving a linear system of equations

$$Ax = b, \tag{10}$$

where $A = [A_1, \ldots, A_n] \in \mathbb{R}^{N\times N}, x = [x_1^T, \ldots, x_n^T]^T \in \mathbb{R}^{N\times 1}, b \in \mathbb{R}^{N\times 1}$. Throughout this paper, we assume A is non-singular. Then the unique solution to (10) is $x = A^{-1}b$, and problem (9) has a unique primal-dual optimal solution $(x, \mu) = (A^{-1}b, 0)$. The augmented Lagrangian function (5) for the optimization problem (9) becomes

$$\mathcal{L}(x, \mu) = -\mu^T(Ax - b) + \frac{\beta}{2}\|Ax - b\|^2.$$

Throughout this paper, we assume $\beta = 1$; note that our algorithms and results can be extended to any $\beta > 0$ by simply scaling μ.

2.1 Example: 3-block ADMM

Before presenting the update equation for general RP-ADMM, we consider a simple case $N = n = 3, d_i = 1, \forall i$ and $\sigma = (123)$, and let $a_i = A_i \in \mathbb{R}^{3\times 1}$. The update equations (7) and (8) can be rewritten as

$$\begin{aligned}
-a_1^T\lambda^k + a_1^T(a_1x_1^{k+1} + a_2x_2^k + a_3x_3^k - b) = 0,\\
-a_2^T\lambda^k + a_2^T(a_1x_1^{k+1} + a_2x_2^{k+1} + a_3x_3^k - b) = 0,\\
-a_3^T\lambda^k + a_3^T(a_1x_1^{k+1} + a_2x_2^{k+1} + a_3x_3^{k+1} - b) = 0,\\
(a_1x_1^{k+1} + a_2x_2^{k+1} + a_3x_3^{k+1} - b) + \lambda^{k+1} - \lambda^k = 0.
\end{aligned}$$

Denote $y^k = [x_1^k; x_2^k; x_3^k; (\lambda^k)^T] \in \mathbb{R}^{6\times 1}$, then the above update equation becomes

$$\begin{bmatrix} a_1^Ta_1 & 0 & 0 & 0\\ a_2^Ta_1 & a_2^Ta_2 & 0 & 0\\ a_3^Ta_1 & a_3^Ta_2 & a_3^Ta_3 & 0\\ a_1 & a_2 & a_3 & I_{3\times 3}\end{bmatrix} y^{k+1} = \begin{bmatrix} 0 & -a_1^Ta_2 & -a_1^Ta_3 & a_1^T\\ 0 & 0 & -a_2^Ta_3 & a_2^T\\ 0 & 0 & 0 & a_3^T\\ 0 & 0 & 0 & I_{3\times 3}\end{bmatrix} y^k + \begin{bmatrix} A^Tb\\ b\end{bmatrix}. \tag{11}$$

Define

$$L \triangleq \begin{bmatrix} a_1^Ta_1 & 0 & 0\\ a_2^Ta_1 & a_2^Ta_2 & 0\\ a_3^Ta_1 & a_3^Ta_2 & a_3^Ta_3\end{bmatrix}, \quad R \triangleq \begin{bmatrix} 0 & -a_1^Ta_2 & -a_1^Ta_3\\ 0 & 0 & -a_2^Ta_3\\ 0 & 0 & 0\end{bmatrix}. \tag{12}$$

The relation between L and R is

$$L - R = A^TA.$$

Define

$$\bar{L} \triangleq \begin{bmatrix} L & 0\\ A & I_{3\times 3}\end{bmatrix}, \quad \bar{R} \triangleq \begin{bmatrix} R & A^T\\ 0 & I_{3\times 3}\end{bmatrix}, \quad \bar{b} = \begin{bmatrix} A^Tb\\ b\end{bmatrix} \tag{13}$$

then the update equation (11) becomes $\bar{L}y^{k+1} = \bar{R}y^k$, i.e.

$$y^{k+1} = (\bar{L})^{-1}\bar{R}y^k + \bar{L}^{-1}\bar{b}. \tag{14}$$

As a side remark, reference [5] provides a specific example of $A \in \mathbb{R}^{3\times 3}$ so that $\rho((\bar{L})^{-1}\bar{R}) > 1$, which implies the divergence of the above iteration if the update order $\sigma = (123)$ is used all the time. This counterexample disproves the convergence of cyclic 3-block ADMM.

2.2 General Update Equation of RP-ADMM

In this case, the optimization problem is (9), and the primal update (7) becomes

$$-A_{\sigma(i)}^T\mu^k + A_{\sigma(i)}^T(\sum_{j=1}^{i} A_{\sigma(j)}x_{\sigma(j)}^{k+1} + \sum_{l=i+1}^{n} A_{\sigma(l)}x_{\sigma(l)}^k - b) = 0,\ i = 1, \ldots, n. \tag{15}$$

Denote the output of Algorithm 1 after round $(k-1)$ as

$$y^k \triangleq [x_1^k; \ldots; x_n^k; \mu^k] \in \mathbb{R}^{2N\times 1}.$$

Similar to the previous subsection, the update equations of Algorithm 1 for solving (9), i.e. (15) and (8), can be written in the matrix form as (when the permutation is σ and $\beta = 1$)

$$y^{k+1} = \bar{L}_\sigma^{-1} \bar{R}_\sigma y^k + \bar{L}_\sigma^{-1} \bar{b}, \tag{16}$$

where $\bar{L}_\sigma, \bar{R}_\sigma, L_\sigma, R_\sigma, \bar{b}$ are defined by

$$\bar{L}_\sigma \triangleq \begin{bmatrix} L_\sigma & 0 \\ A & I_{N\times N} \end{bmatrix}, \quad \bar{R}_\sigma \triangleq \begin{bmatrix} R_\sigma & A^T \\ 0 & I_{N\times N} \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} A^T b \\ b \end{bmatrix}, \tag{17}$$

in which $L_\sigma \in \mathbb{R}^{N\times N}$ has $n \times n$ blocks and the (i, j)-th block is defined as

$$L_\sigma[\sigma(i), \sigma(j)] \triangleq \begin{cases} A_{\sigma(i)}^T A_{\sigma(j)} & j \le i, \\ 0 & j > i, \end{cases} \tag{18}$$

and R_σ is defined as

$$R_\sigma \triangleq L_\sigma - A^T A.$$

When $n = 3, d_i = 1, \forall i$ and $\sigma = (123)$, L_σ defined above is the same as L defined in (12).

3 Other Randomized ADMM

In this section, we present two other versions of randomized ADMM which can be divergent according to simulations. The failure of these versions makes us focus on analyzing RP-ADMM in this paper.

In the first algorithm, called primal-dual randomized ADMM (PD-RADMM), the whole dual variable is viewed as the $(n+1)$-th block. In particular, at each iteration, the algorithm draws one index i from $\{1, \dots, n, n+1\}$, then performs the following update: if $i \le n$, update the i-th block of the primal variable; if $i = n + 1$, update the whole dual variable. The details are given in Algorithm 2. We have tested PD-RADMM for the counter-example given in [5], and found that PD-RADMM always diverges (for random initial points).

Algorithm 2 Primal-Dual Randomized ADMM (PD-RADMM)

Iteration t ($t = 0, 1, 2, \dots$):
  Pick $i \in \{1, \dots, n, n+1\}$ uniformly at random;
    If $1 \le i \le n$:
      $x_i^{t+1} = \arg\min_{x_i \in \mathcal{X}_i} \mathcal{L}(x_1^t, \dots, x_{i-1}^t, x_i, x_{i+1}^t, \dots, x_n^t; \mu^t)$,
      $x_j^{t+1} = x_j^t, \ \forall\, j \in \{1, \dots, n\} \backslash \{i\}$,
      $\mu^{t+1} = \mu^t$.
    Else If $i = n + 1$:
      $\mu^{t+1} = \mu^t - \beta(\sum_{i=1}^n A_i x_i^{t+1} - b)$,
      $x_j^{t+1} = x_j^t, \ \forall\, j \in \{1, \dots, n\}$.
    End

In the second algorithm, called primal randomized ADMM (P-RADMM), we only perform randomization for the primal variables. In particular, at each round, we first draw n independent random variables $j_1, \dots, j_n$ from the uniform distribution of $\{1, \dots, n\}$ and update $x_{j_1}, \dots, x_{j_n}$ sequentially, then update the dual variable in the usual way. The details are given in Algorithm 3. This algorithm looks quite similar to RP-ADMM as they both update n primal blocks at each round; the difference is that RP-ADMM samples *without replacement* while this algorithm P-RADMM samples *with replacement*. In other words, RP-ADMM updates each block exactly once at each round, while P-RADMM may update one block more than one times or does not update one block at each round. We have tested P-RADMM in various settings. For the counter-example given in [5], we found that P-RADMM does converge. However, if $n \ge 30$ and A is a Gaussian random matrix (each entry is drawn i.i.d. from $\mathcal{N}(0, 1)$), then P-RADMM diverges in almost all cases we have tested. This phenomenon is rather strange since for random Gaussian matrices A the cyclic ADMM actually converges (according to simulations). An implication is that randomized versions do not always outperform their deterministic counterparts in terms of convergence.

Since both Algorithm 2 and Algorithm 3 can diverge in certain cases, we will not further study them in this paper. In the rest of the paper, we will focus on RP-ADMM (i.e. Algorithm 1).

Algorithm 3 Primal Randomized ADMM (P-RADMM)

Round k ($k = 0, 1, 2, \dots$):

1) Primal update.

Pick $l_1, \dots, l_n$ independently from the uniform distribution of $\{1, \dots, n\}$.

For $i = 1, \dots, n$:

$t = kn + i - 1$,

$x_{l_i}^{t+1} = \arg\min_{x_{l_i} \in \mathcal{X}_{l_i}} \mathcal{L}(x_1^t, \dots, x_{l_i-1}^t, x_{l_i}, x_{l_i+1}^t, \dots, x_n^t; \mu^t)$,

$x_j^{t+1} = x_j^t, \ \forall\, j \in \{1, \dots, n\} \backslash \{l_i\}$,

$\mu^{t+1} = \mu^t$.

End.

2) Dual update.

$\mu^{(k+1)n} = \mu^{kn} - \beta(\sum_{i=1}^n A_i x_i^{(k+1)n} - b)$.

4 Main Results

Let σ_i denote the permutation used in round i of Algorithm 1, which is a uniform random variable drawn from the set of permutations Γ. After round k, Algorithm 1 generates a random output y^{k+1}, which depends on the observed draw of the random variable

$$\xi_k = (\sigma_0, \sigma_1, \dots, \sigma_k). \tag{19}$$

We will show that the expected output

$$\phi^k = E_{\xi_{k-1}}(y^k) \tag{20}$$

converges to the primal-dual solution of the problem (9). Note that the expected iterate convergence does not necessarily implies that the iterates converge. However, it strongly indicates that random permutation make a dramatic difference in multi-block ADMM (i.e. ADMM with more than two blocks).

Theorem 1 *Assume the coefficient matrix $A = [A_1, \dots, A_n]$ of the constraint in* (9) *is a non-singular square matrix. Suppose Algorithm 1 is used to solve problem* (9)*, then the expected output converges to the unique primal-dual optimal solution to* (9)*, i.e.*

$$\{\phi^k\}_{k\to\infty} \longrightarrow \begin{bmatrix} A^{-1}b \\ 0 \end{bmatrix}. \tag{21}$$

Since the update matrix does not depend on previous iterates, we claim (and prove in Section 4.1) that Theorem 1 holds if the expected update matrix has a spectral radius less than 1, i.e. if the following Theorem 2 holds.

Theorem 2 *Suppose $A = [A_1, \dots, A_n] \in \mathbb{R}^{N\times N}$ is non-singular, and $\bar{L}_\sigma^{-1}, \bar{R}_\sigma$ are defined by* (17) *for any permutation σ. Define*

$$M \triangleq E_\sigma(\bar{L}_\sigma^{-1}\bar{R}_\sigma) = \frac{1}{n!}\sum_{\sigma\in\Gamma}(\bar{L}_\sigma^{-1}\bar{R}_\sigma), \tag{22}$$

where the expectation is taken over the uniform random distribution over Γ, the set of permutations of $\{1, 2, \dots, n\}$. Then the spectral radius of M is smaller than 1*, i.e.*

$$\rho(M) < 1. \tag{23}$$

Remark 4.1 *For the counterexample in [5] where $A = [1, 1, 1; 1, 1, 2; 1, 2, 2]$, we have $\rho(M_\sigma) > 1.02$ for any permutation of $(1, 2, 3)$. Theorem 2 shows that even if each M_σ is "bad", the average of them is always "good".*

Theorem 2 is just a linear algebra result, and can be understood even without knowing the details of the algorithm. However, the proof of Theorem 2 is rather non-trivial and forms the main body of the paper. This proof will be provided in Section 4.2, and the technical results used in this proof will be provided in the supplement.

4.1 Proof of Theorem 1

Denote σ_k as the permutation used in round k, and define ξ_k as in (19). Rewrite the update equation (16) below (replacing σ by σ_k):

$$y^{k+1} = \bar{L}_{\sigma_k}^{-1}\bar{R}_{\sigma_k} y^k + \bar{L}_{\sigma_k}^{-1}\bar{b}. \tag{24}$$

We first prove (21) for the case $b = 0$. By (17) we have $\bar{b} = 0$, then (24) is simplified to $y^{k+1} = \bar{L}_{\sigma_k}^{-1}\bar{R}_{\sigma_k} y^k$. Taking the expectation of both sides of this equation in ξ_k (see its definition in (19)), and note that y^k is independent of σ_k, we get

$$\phi^{k+1} = E_{\xi_k}(\bar{L}_{\sigma_k}^{-1}\bar{R}_{\sigma_k} y^k) = E_{\sigma_k}\left(E_{\xi_{k-1}}(\bar{L}_{\sigma_k}^{-1}\bar{R}_{\sigma_k} y^k)\right) = E_{\sigma_k}(\bar{L}_{\sigma_k}^{-1}\bar{R}_{\sigma_k}\phi^k) = M\phi^k.$$

Since the spectral radius of M is less than 1 by Theorem 2, we have that $\{\phi^k\} \to 0$, i.e. (21).

We then prove (21) for general b. Let $y^* = [A^{-1}b; 0]$ denote the optimal solution. Then it is easy to verify that

$$y^* = \bar{L}_{\sigma_k}^{-1}\bar{R}_{\sigma_k} y^* + \bar{L}_{\sigma_k}^{-1}\bar{b}$$

for all $\sigma_k \in \Gamma$ (i.e. the optimal solution is the fixed point of the update equation for any order). Compute the difference between this equation and (24) and letting $\hat{y}^k = y^k - y^*$, we get $\hat{y}^{k+1} = \bar{L}_{\sigma_k}^{-1}\bar{R}_{\sigma_k}\hat{y}^k$. According to the proof for the case $b = 0$, we have $E(\hat{y}^k) \longrightarrow 0$, which implies $E(y^k) \longrightarrow y^*$. $\square$

4.2 Proof of Theorem 2

The difficulty of proving Theorem 2 (bounding the spectral radius of M) is two-fold. First, M is a non-symmetric matrix, and there are very few tools to bound the spectral radius of a non-symmetric matrix. In fact, spectral radius is neither subadditive nor submultiplicative (see, e.g. [16]). Note that the spectral norm of M can be much larger than 1 (there are examples that $\|M\| > 2$), thus we cannot bound the spectral radius simply by the spectral norm. Second, although it is possible to explicitly write each entry of M as a function of the entries of A^TA, these functions are very complicated (n-th order polynomials).

The proof outline of Theorem 2 and the main techniques are described below. In Step 0, we provide an expression of the expected update matrix M. In Step 1, we establish the relationship between the eigenvalues of M and the eigenvalues of a simple symmetric matrix AQA^T. As a consequence, the spectral radius of M is smaller than one iff the eigenvalues of AQA^T lie in the region $(0, 4/3)$. This step partially resolves the first difficulty, i.e. how to deal with the spectral radius of a non-symmetric matrix. In Step 2, we show that the eigenvalues of AQA^T do lie in $(0, 4/3)$ using mathematical induction. The induction analysis circumvents the second difficulty, i.e. how to utilize the relation between M and A. Note that we will perform induction analysis for QA^TA (with the same eigenvalues as AQA^T) which is non-symmetric, and we will use several techniques in Step 1 again to transform non-symmetric matrices to symmetric matrices.

Step 0: compute the expression of the expected update matrix M. Define

$$Q \triangleq E_\sigma(L_\sigma^{-1}) = \frac{1}{n!}\sum_{\sigma\in\Gamma} L_\sigma^{-1}. \tag{25}$$

It is easy to prove that Q defined by (25) is symmetric. In fact, note that $L_\sigma^T = L_{\bar{\sigma}}, \forall \sigma \in \Gamma$, where $\bar{\sigma}$ is a reverse permutation of σ satisfying $\bar{\sigma}(i) = \sigma(n+1-i), \forall\, i$, thus $Q = \frac{1}{n!}\sum_\sigma Q_\sigma = (\frac{1}{n!}\sum_\sigma Q_{\bar{\sigma}})^T = Q^T$, where the last step is because the sum of all $Q_{\bar{\sigma}}$ is the same as the sum of all Q_σ. Denote

$$M_\sigma \triangleq \bar{L}_\sigma^{-1}\bar{R}_\sigma = \bar{L}_\sigma^{-1}\begin{bmatrix} R_\sigma & A^T \\ 0 & I \end{bmatrix}. \tag{26}$$

Substituting the expression of $\bar{L}_\sigma^{-1}$ into the above relation, and replacing R_σ by $L_\sigma - A^TA$, we obtain

$$M_\sigma = \begin{bmatrix} L_\sigma^{-1} & 0 \\ -AL_\sigma^{-1} & I \end{bmatrix}\begin{bmatrix} L_\sigma - A^TA & A^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} I - L_\sigma^{-1}A^TA & L_\sigma^{-1}A^T \\ -A + AL_\sigma^{-1}A^TA & I - AL_\sigma^{-1}A^T \end{bmatrix}. \tag{27}$$

Since M_σ is linear in L_σ^{-1}, we have

$$\begin{aligned} M = E_\sigma(M_\sigma) &= \begin{bmatrix} I - E_\sigma(L_\sigma^{-1})A^TA & E_\sigma(L_\sigma^{-1})A^T \\ -A + AE_\sigma(L_\sigma^{-1})A^TA & I - AE_\sigma(L_\sigma^{-1})A^T \end{bmatrix} \\ &= \begin{bmatrix} I - QA^TA & QA^T \\ -A + AQA^TA & I - AQA^T \end{bmatrix}. \end{aligned} \tag{28}$$

Step 1: relate M to a simple symmetric matrix. The main result of Step 1 is given below, and the proof of this result is given in the supplement.

Lemma 1 *Suppose $A \in \mathbb{R}^{N\times N}$ is non-singular and $Q \in \mathbb{R}^{N\times N}$ is an arbitrary matrix. Define $M \in \mathbb{R}^{2N\times 2N}$ as*

$$M = \begin{bmatrix} I - QA^TA & QA^T \\ -A + AQA^TA & I - AQA^T \end{bmatrix}. \tag{29}$$

Then

$$\lambda \in eig(M) \Longleftrightarrow \frac{(1-\lambda)^2}{1-2\lambda} \in eig(QA^TA). \tag{30}$$

Furthermore, when Q is symmetric, we have

$$\rho(M) < 1 \Longleftrightarrow eig(QA^TA) \subseteq (0, \frac{4}{3}). \tag{31}$$

Remark: For our problem, the matrix Q as defined by (25) is symmetric (see the argument after equation (25)). Lemma 1 implies (31) holds. Note that the first conclusion (30) holds even if Q is non-symmetric.

Step 2: Bound the eigenvalues of QA^TA. The main result of Step 2 is summarized in the following Lemma 2. The proof of Lemma 2 is based on an induction formula that relates Q to its lower dimensional analogs, and several techniques used in the proof of Lemma 1; see the supplement for the details of the proof.

Lemma 2 *Suppose $A = [A_1, \dots, A_n] \in \mathbb{R}^{N\times N}$ is non-singular. Define Q as*

$$Q \triangleq E_\sigma(L_\sigma^{-1}) = \frac{1}{n!}\sum_{\sigma\in\Gamma} L_\sigma^{-1}, \tag{32}$$

in which L_σ is defined by (18) and Γ is defined by (6). Then all eigenvalues of QA^TA lie in $(0, 4/3)$, i.e.

$$eig(QA^TA) \subseteq (0, \frac{4}{3}). \tag{33}$$

Theorem 2 follows immediately from Lemma 1 and Lemma 2.

5 Conclusion

In this paper, we propose randomly permuted ADMM (RP-ADMM) and prove the expected convergence of RP-ADMM for solving a non-singular square system of equations. Multi-block ADMM is one promising candidate for solving large-scale linearly constrained problems in big data applications, but its cyclic version is known to be possibly divergent. Our result shows that RP-ADMM may serve as a simple solution to resolve the divergence issue of cyclic multi-block ADMM. One interesting aspect is that while it is possible that every single permutation leads to a "bad" update matrix, averaging these permutations always leads to a "good" update matrix. Our result is also one of the first direct analysis of random permutation (sampling without replacement) in optimization algorithms, though independent randomization (sampling with replacement) has been extensively studied for BCD and SGD. It is not hard to extend our result to non-square (including tall and wide) system of equations. Future directions include extending our result to general convex problems and proving the convergence of RP-ADMM with high probability.

References

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

[2] D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[4] T. F. Chan and R. Glowinski. *Finite element approximation and iterative solution of a class of mildly non-linear elliptic equations*. Computer Science Department, Stanford University Stanford, 1978.

[5] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, pages 1–23, 2014.

[6] W. Deng, M.-J. Lai, Z. Peng, and W. Yin. Parallel multi-block ADMM with o (1/k) convergence. *arXiv preprint arXiv:1312.3040*, 2013.

[7] J. Eckstein and W. Yao. Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. *RUTCOR Research Reports*, 32, 2012.

[8] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[9] R. Glowinski. On alternating direction methods of multipliers: a historical perspective. In *Modeling, Simulation and Optimization for Science and Technology*, pages 59–82. Springer, 2014.

[10] R. Glowinski and A. Marroco. Approximation par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

[11] D. Han and X. Yuan. A note on the alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 155(1):227–238, 2012.

[12] B. He, M. Tao, and X. Yuan. Alternating direction method with Gaussian back substitution for separable convex programming. *SIAM Journal on Optimization*, 22(2):313–340, 2012.

[13] B. He, M. Tao, and X. Yuan. Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming. *Math. Oper. Res., under revision*, 2, 2012.

[14] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo. A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. *arXiv preprint arXiv:1401.7079*, 2014.

[15] M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

[16] F. Kittaneh. Spectral radius inequalities for Hilbert space operators. *Proceedings of the American Mathematical Society*, pages 385–390, 2006.

[17] T. Lin, S. Ma, and S. Zhang. On the convergence rate of multi-block ADMM. *arXiv preprint arXiv:1408.4265*, 2014.

[18] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[19] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[20] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.

[21] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

[22] D. Sun, K.-C. Toh, and L. Yang. A convergent proximal alternating direction method of multipliers for conic programming with 4-block constraints. *arXiv preprint arXiv:1404.5378*, 2014.

[23] R. Sun. *Matrix Completion via Nonconvex Factorization: Algorithms and Theory*. PhD thesis, University of Minnesota, 2015.

[24] H. Wang, A. Banerjee, and Z.-Q. Luo. Parallel direction method of multipliers. In *Advances in Neural Information Processing Systems*, pages 181–189, 2014.

Supplemental Materials

6 Proof of Lemma 1

The proof of Lemma 1 relies on two simple techniques. The first technique, as elaborated in the Step 1 below, is to factorize M and rearrange the factors. The second technique, as elaborated in the Step 2 below, is to reduce the dimension by eliminating a variable from the eigenvalue equation.

Step 1: Factorizing M and rearranging the order of multiplication. The following observation is crucial: the matrix M defined by (29) can be factorized as

$$M = \begin{bmatrix} I & 0 \\ -A & I \end{bmatrix} \begin{bmatrix} QA^T & I \\ I & A \end{bmatrix} \begin{bmatrix} -A & I \\ I & 0 \end{bmatrix}.$$

Switching the order of the products by moving the first component to the last, we get a new matrix

$$M' \triangleq \begin{bmatrix} QA^T & I \\ I & A \end{bmatrix} \begin{bmatrix} -A & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ -A & I \end{bmatrix} = \begin{bmatrix} QA^T & I \\ I & A \end{bmatrix} \begin{bmatrix} -2A & I \\ I & 0 \end{bmatrix} = \begin{bmatrix} I - 2QA^TA & QA^T \\ -A & I \end{bmatrix}. \tag{34}$$

Note that $\text{eig}(XY) = \text{eig}(YX)$ for any two square matrices, thus

$$\text{eig}(M) = \text{eig}(M').$$

To prove (30), we only need to prove

$$\lambda \in \text{eig}(M') \Longleftrightarrow \frac{(1-\lambda)^2}{1-2\lambda} \in \text{eig}(QA^TA). \tag{35}$$

Step 2: Relate the eigenvalues of M' to the eigenvalues of QA^TA, i.e. prove (35). This step is simple as we only use the definition of eigenvalues. However, note that, without Step 1, just applying the definition of eigenvalues of the original matrix M may not lead to a simple relationship as (35).

We first prove one direction of (30):

$$\lambda \in \text{eig}(M') \Longrightarrow \frac{(1-\lambda)^2}{1-2\lambda} \in \text{eig}(QA^TA). \tag{36}$$

Suppose $v \in \mathbb{C}^{2N\times 1}\backslash\{0\}$ is an eigenvector of M' corresponding to the eigenvalue λ, i.e.

$$M'v = \lambda v.$$

Partition v as $v = \begin{bmatrix} v_1 \\ v_0 \end{bmatrix}$, where $v_1, v_0 \in \mathbb{C}^{N\times 1}$. Using the expression of M' in (34), we can write the above equation as

$$\begin{bmatrix} I - 2QA^TA & QA^T \\ -A & I \end{bmatrix} \begin{bmatrix} v_1 \\ v_0 \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_0 \end{bmatrix},$$

which implies

$$(I - 2QA^TA)v_1 + QA^Tv_0 = \lambda v_1, \tag{37a}$$

$$-Av_1 + v_0 = \lambda v_0. \tag{37b}$$

We claim that (36) holds when $v_1 = 0$. In fact, in this case we must have $v_0 \neq 0$ (otherwise $v = 0$ cannot be an eigenvector). By (37b) we have $\lambda v_0 = v_0$, thus $\lambda = 1$. By (37a) we have $0 = QA^Tv_0 = QA^TA(A^{-1}v_0)$, which implies $\frac{(1-\lambda)^2}{1-2\lambda} = 0 \in \text{eig}(QA^TA)$, therefore (36) holds in this case.

We then prove (36) for the case

$$v_1 \neq 0. \tag{38}$$

The equation (37b) implies $(1-\lambda)v_0 = Av_1$. Multiplying both sides of (37a) by $(1-\lambda)$ and invoking this equation, we get

$$(1-\lambda)(I - 2QA^TA)v_1 + QA^TAv_1 = (1-\lambda)\lambda v_1.$$

This relation can be simplified to

$$(1-2\lambda)QA^TAv_1 = (1-\lambda)^2 v_1. \tag{39}$$

We must have $\lambda \neq \frac{1}{2}$; otherwise, the above relation implies $v_1 = 0$, which contradicts (38). Then (39) becomes

$$QA^TAv_1 = \frac{(1-\lambda)^2}{1-2\lambda}v_1. \tag{40}$$

Therefore, $\frac{(1-\lambda)^2}{1-2\lambda}$ is an eigenvalue of QA^TA, with the corresponding eigenvector $v_1 \neq 0$, which finishes the proof of (36).

The other direction [2]

$$\lambda \in \text{eig}(M) \Longleftarrow \frac{(1-\lambda)^2}{1-2\lambda} \in \text{eig}(QA^TA) \tag{41}$$

is easy to prove. Suppose $\frac{(1-\lambda)^2}{1-2\lambda} \in \text{eig}(QA^TA)$. We consider two cases.

Case 1: $\frac{(1-\lambda)^2}{1-2\lambda} = 0$. In this case $\lambda = 1$. Since $0 = \frac{(1-\lambda)^2}{1-2\lambda} \in \text{eig}(QA^TA)$, there exists $v_0 \in \mathbb{C}^N \backslash \{0\}$ such that $QA^TAv_0 = 0$ and Let $v_1 = (0, \ldots, 0)^T \in \mathbb{C}^{N\times 1}$, then v_0, v_1 and $\lambda = 1$ satisfy (37). Thus $v = \begin{bmatrix} v_1 \\ v_0 \end{bmatrix} \in \mathbb{C}^{2N} \backslash \{0\}$ satisfies $Mv = \lambda v$, which implies $\lambda = 1 \in \text{eig}(M)$.

Case 2: $\frac{(1-\lambda)^2}{1-2\lambda} \neq 0$, then $\lambda \neq 1$. Let v_1 be the eigenvector corresponding to $\frac{(1-\lambda)^2}{1-2\lambda}$ (i.e. pick v_1 that satisfies (40)), and define $v_0 = v_1/(1-\lambda)$. It is easy to verify that $v = \begin{bmatrix} v_1 \\ v_0 \end{bmatrix}$ satisfies $Mv = \lambda v$, which implies $\lambda \in \text{eig}(M)$.

Step 3: When Q is symmetric, prove (31) by simple algebraic computation.

Since Q is symmetric, we know that $\text{eig}(QA^TA) = \text{eig}(AQA^T) \subseteq \mathbb{R}$. Suppose $\tau \in \mathbb{R}$ is an eigenvalue of QA^TA, then any λ satisfying $\frac{(1-\lambda)^2}{1-2\lambda} = \tau$ is an eigenvalue of M. This relation can be rewritten as $\lambda^2 + 2(\tau - 1)\lambda + (1-\tau) = 0$, which, as a real-coefficient quadratic equation in λ, has two roots

$$\lambda_1 = 1 - \tau + \sqrt{\tau(\tau-1)}, \quad \lambda_2 = 1 - \tau - \sqrt{\tau(\tau-1)}. \tag{42}$$

Note that when $\tau(\tau-1) < 0$, the expression $\sqrt{\tau(\tau-1)}$ denotes a complex number $i\sqrt{\tau(1-\tau)}$, where i is the imaginary unit. To prove (31), we only need to prove

$$\max\{|\lambda_1|, |\lambda_2|\} < 1 \Longleftrightarrow 0 < \tau < \frac{4}{3}. \tag{43}$$

Consider three cases.

Case 1: $\tau < 0$. Then $\tau(\tau-1) = |\tau|(|\tau|+1) > 0$. In this case, $\lambda_1 = 1 + |\tau| + \sqrt{|\tau|(|\tau|+1)} > 1$.

Case 2: $0 < \tau < 1$. Then $\tau(\tau-1) < 0$, and (42) can be rewritten as

$$\lambda_{1,2} = 1 - \tau \pm i\sqrt{\tau(1-\tau)},$$

which implies $|\lambda_1| = |\lambda_2| = \sqrt{(1-\tau)^2 + \tau(1-\tau)} = \sqrt{1-\tau} < 1$.

Case 3: $\tau > 1$. Then $\tau(\tau-1) > 0$. According to (42), it is easy to verify $\lambda_1 > 0 > \lambda_2$ and

$$|\lambda_2| = \tau - 1 + \sqrt{\tau(\tau-1)} > 1 - \tau + \sqrt{\tau(\tau-1)} = |\lambda_1|.$$

Then we have

$$\max\{|\lambda_1|, |\lambda_2|\} < 1 \Longleftrightarrow |\lambda_2| = \tau - 1 + \sqrt{\tau(\tau-1)} < 1 \Longleftrightarrow 1 < \tau < \frac{4}{3}.$$

Combining the conclusions of the three cases immediately leads to (43).

7 Proof of Lemma 2 for the case $d_i = 1, \forall i$

In this section, we prove Lemma 2 for the case $d_i = 1, \forall i$. The proof for general d_i's is quite similar (but not exactly the same), and will be given in Section 8.

7.1 Proof Overview

We will use mathematical induction to prove Lemma 2, and the reason of doing so is the following. A major difficulty of proving Lemma 2 is that each entry of Q is a complicated function (in fact, n-th order polynomial) of the entries of A^TA. To circumvent this difficulty, we will implicitly exploit the property of Q by an induction analysis on n, the number of blocks.

[2]For the purpose of proving Theorem 2, we do not need to prove this direction. Here we present the proof since it is quite straightforward and makes the result more comprehensive.

The difficulty of using induction to prove Lemma 2 is two-fold. First, it is not obvious how Q is related to an analogous matrix in a lower dimension. Second, the simulations show that $\|QA^TA\| < \frac{4}{3} \ll \|Q\|\|A^TA\|$, thus we have to bound the eigenvalues of the product QA^TA, instead of the eigenvalues of Q. Even if we know the relationship between Q and a lower-dimensional matrix $\hat{Q}$, it is not obvious how $\text{eig}(QA^TA)$ and $\text{eig}(\hat{Q}\hat{A}^T\hat{A})$ are related, where $\hat{A}$ is a lower-dimensional analog of A.

The proof outline of Lemma 2 and the main techniques are described below. In Step 1, we prove an induction formula in Proposition 1, which states that Q can be decomposed as the sum of n symmetric matrices, where each symmetric matrix contains an $(n-1)\times(n-1)$ sub-matrix $\hat{Q}_k$ that is analogous to Q. In other words, we relate Q to n analogous matrices $\hat{Q}_k, k = 1, \ldots, n$ in a lower dimension. To prove the induction formula, we use a three-level symmetrization technique. This induction formula resolves the first difficulty. In Step 2, we prove the induction step, i.e. under the induction hypothesis that $\text{eig}(\hat{Q}_k\hat{A}_k^T\hat{A}_k) \subseteq (0, \frac{4}{3}), k = 1, \ldots, n$, where $\hat{A}_k$ is a certain sub-matrix of A, the desired result $\text{eig}(QA^TA) \subseteq (0, \frac{4}{3})$ holds. To build the relation between $\text{eig}(QA^TA)$ and $\text{eig}(\hat{Q}_k\hat{A}_k^T\hat{A}_k)$, we will apply the two simple techniques used in the proof of Lemma 1: factorize and rearrange, and reduce the dimension by eliminating a variable from the eigenvalue equation. Nevertheless, the subsequent analysis is more complicated than the proof of Lemma 1.

7.2 Proof of Lemma 2 for $d_i = 1, \forall i$ and Two Propositions

Without loss of generality, we can assume $\|a_i\|^2 = 1, \forall i$ (see the first paragraph of Section 8 for an explanation).

We use mathematical induction to prove Lemma 2 for the n-coordinate case. For the basis of the induction ($n = 1$), Lemma 2 holds since $QA^TA = 1$. Assume Lemma 2 holds for $n-1$, we will prove Lemma 2 for n.

7.2.1 Step 1: Induction formula for general n

Since $d_i = 1, \forall\, i$, we have $N = \sum_i d_i = n$. Denote $a_i \triangleq A_i \in \mathbb{R}^{n\times 1}(i = 1, \ldots, n)$. Denote $[n] \triangleq \{1, \ldots, n\}$. For any $k \in [n]$, define

$$\Gamma_k \triangleq \{\sigma' \mid \sigma' \text{ is a permutation of } [n]\backslash\{k\}\}. \tag{44}$$

For any $\sigma' \in \Gamma_k$, similar to (**??**), we can define $L_{\sigma'}, Q_k \in \mathbb{R}^{(n-1)\times(n-1)}$ as

$$L_{\sigma'}(\sigma'(i), \sigma'(j)) \triangleq \begin{cases} a_{\sigma'(i)}^T a_{\sigma'(j)} & i \geq j, \\ 0 & i < j, \end{cases} \tag{45}$$

$$\hat{Q}_k \triangleq \frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} L_{\sigma'}^{-1},\ k = 1, \ldots, n. \tag{46}$$

Note that $L_{\sigma'}$ and $\hat{Q}_k$ are lower-dimensional analogs of L_σ and Q respectively.

Define w_k as the k-th column of A^TA excluding the entry $a_k^Ta_k$, i.e.

$$w_k \triangleq [a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n]^T a_k \in \mathbb{R}^{(n-1)\times 1}. \tag{47}$$

Define permutation matrices $S_1, \ldots, S_n \in \mathbb{R}^{n\times n}$ as follows:

$$S_k(i,i) = 1, i = 1, \ldots, k-1; \quad S_k(k+1, k) = \cdots = S_k(n, n-1) = 1; \quad S_k(k, n) = 1, \tag{48}$$

and all other entries of S_k are zero. S_k is called a permutation matrix since it corresponds to a permutation $(1, \ldots, k-1, k+1, \ldots, n, k)$; in fact, $(1, 2, \ldots, n)S_k = (1, \ldots, k-1, k+1, \ldots, n, k)$. Replacing $1, 2, \ldots, n$ by column vectors $b_1, \ldots, b_n$, we get $(b_1, b_2, \ldots, b_n)S_k = (b_1, \ldots, b_{k-1}, b_{k+1}, \ldots, b_n, b_k)$. This relation can be interpreted as the following column-moving property of S_k: right-multiply a matrix by S_k will move the k-th column to the end (i.e. in the new matrix it becomes the last column). Similarly, S_k^T has the following row-moving property: left-multiply a matrix by S_k^T will move the k-th row to the end. Note that S_n is the identity matrix. Another property is

$$S_k^T = S_k^{-1}. \tag{49}$$

We give an example to illustrate the expressions of S_k. When $n = 3$, $S_1, S_2, S_3 \in \mathbb{R}^{3\times 3}$ defined in (48) can be explicitly written as

$$S_1 \triangleq \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},\ S_2 \triangleq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},\ S_3 = I_{3\times 3}.$$

The column-moving property means $[b_1, b_2, b_3]S_1 = [b_2, b_3, b_1]$, and $[b_1, b_2, b_3]S_2 = [b_1, b_3, b_2]$. Similarly, the row-moving property means $S_1^T \begin{bmatrix} b_1^T \\ b_2^T \\ b_3^T \end{bmatrix} = \begin{bmatrix} b_2^T \\ b_3^T \\ b_1^T \end{bmatrix}$ and $S_2^T \begin{bmatrix} b_1^T \\ b_2^T \\ b_3^T \end{bmatrix} = \begin{bmatrix} b_1^T \\ b_3^T \\ b_2^T \end{bmatrix}$.

With these definitions, we are ready to present the induction formula, which builds a relation between Q and its lower-dimensional analogs $\hat{Q}_k, k = 1, \dots, n$.

Proposition 1 *The matrix $Q = \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} L_\sigma^{-1}$, where L_σ, Γ are defined by* (**??**) *and* (6) *respectively, can be decomposed as follows:*

$$Q = \frac{1}{n} \sum_{k=1}^{n} S_k Q_k S_k^T, \tag{50}$$

where

$$Q_k \triangleq \begin{bmatrix} \hat{Q}_k & -\frac{1}{2}\hat{Q}_k w_k \\ -\frac{1}{2} w_k^T \hat{Q}_k & 1 \end{bmatrix}, \tag{51}$$

in which $\hat{Q}_k$ is defined by (46)*.*

The proof of Proposition 1 for the case $n = 3$ will be given in Section 7.3. We relegate the proof of Proposition 1 for general n to Appendix 9.

7.2.2 Step 2: bounding eigenvalues of each Q_k

According to (50), we have

$$AQA^T = \frac{1}{n} \sum_{k=1}^{n} AS_k Q_k S_k^T A^T.$$

Note that $\hat{Q}_k$ defined by (46) is symmetric, thus Q_k defined by (51) is symmetric, which implies that each $AS_k Q_k S_k^T A^T$ is a symmetric matrix. From the above relation, we have

$$\frac{1}{n} \sum_{k=1}^{n} \lambda_{\min}(AS_k Q_k S_k^T A^T) \le \lambda_{\min}(AQA^T) \le \lambda_{\max}(AQA^T) \le \frac{1}{n} \sum_{k=1}^{n} \lambda_{\max}(AS_k Q_k S_k^T A^T). \tag{52}$$

Therefore, to prove eig$(AQA^T) \subseteq (0, 4/3)$, we only need to prove for any $k = 1, \dots, n$,

$$\text{eig}(AS_k Q_k S_k^T A^T) = \text{eig}(Q_k S_k^T A^T A S_k) \subseteq (0, 4/3). \tag{53}$$

By the column moving property of S_k, we have

$$\bar{A}_k \triangleq AS_k = [\hat{A}_k, a_k], \tag{54}$$

where $\hat{A}_k \triangleq [a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_n]$. Note that $\hat{Q}_k$ only depends on the entries of $\hat{A}_k^T \hat{A}_k \in \mathbb{R}^{(n-1)\times(n-1)}$, thus by the induction hypothesis, we have

$$\text{eig}(\hat{Q}_k \hat{A}_k^T \hat{A}_k) \subseteq (0, 4/3). \tag{55}$$

We claim that (53) follows from the induction hypothesis (55) and the expressions (54) and (51). In fact, the following Proposition 2 directly proves (53) for $k = n$. If we replace $A, \hat{A}_n, a_n, \hat{Q}_n, Q_n$ by $\bar{A}_k, \hat{A}_k, a_k, \hat{Q}_k, Q_k$ respectively in Proposition 2, we will obtain (53) for any k. As mentioned earlier, the desired result eig$(AQA^T) \subseteq (0, 4/3)$ follows immediately from (53) and (52).

Proposition 2 *Suppose $A = [\hat{A}_n, a_n] \in \mathbb{R}^{n\times n}$ is non-singular, where $\hat{A}_n \in \mathbb{R}^{n\times(n-1)}$ and $a_n \in \mathbb{R}^{(n-1)\times 1}$ satisfies $\|a_n\| = 1$. Suppose $\hat{Q}_n \in \mathbb{R}^{(n-1)\times(n-1)}$ is a symmetric matrix which satisfies eig$(\hat{Q}_n \hat{A}_n^T \hat{A}_n) \subseteq (0, \frac{4}{3})$. Define*

$$w_n \triangleq \hat{A}_n^T a_n, \quad Q_n \triangleq \begin{bmatrix} \hat{Q}_n & -\frac{1}{2}\hat{Q}_n w_n \\ -\frac{1}{2} w_n^T \hat{Q}_n & 1 \end{bmatrix}. \tag{56}$$

Then eig$(Q_n A^T A) \subseteq (0, \frac{4}{3})$.

The proof of Proposition 2 is given in Section 7.4.

7.3 Proof of Proposition 1 for $n = 3$

In this subsection, we prove the induction formula for $n = 3$. Before the formal proof, we briefly describe the ideas of contructing this induction formula. The key idea is symmetrization: we start from an obvious relation between L_σ^{-1} and its lower-dimensional analog, and by three levels of symmetrization we can obtain a relation between Q and its lower-dimensional analogs.

Define

$$\begin{aligned} w_{ij} =& a_i^T a_j,\ \forall i, j, \\ w_1 = [w_{12}, w_{13}]^T,\ w_2 =& [w_{21}, w_{23}]^T,\ w_3 = [w_{31}, w_{32}]^T. \end{aligned} \tag{57}$$

For $\sigma = (123)$, the expressions of L_σ and L_σ^{-1} are

$$L_{(123)} = \begin{bmatrix} 1 & 0 & 0 \\ w_{21} & 1 & 0 \\ w_{31} & w_{32} & 1 \end{bmatrix}, \quad L_{(123)}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -w_{21} & 1 & 0 \\ -w_{31} + w_{21}w_{32} & -w_{32} & 1 \end{bmatrix}. \tag{58}$$

Note, however, that the following expressions of L_σ and L_σ^{-1} are more useful:

$$L_{(123)} = \begin{bmatrix} L_{(12)} & 0 \\ w_3^T & 1 \end{bmatrix},$$

and

$$L_{(123)}^{-1} = \begin{bmatrix} L_{(12)}^{-1} & 0 \\ -w_3^T L_{(12)}^{-1} & 1 \end{bmatrix}. \tag{59}$$

The above equation provides a relation between $L_{(123)}^{-1}$ and an analogous matrix L_{12}^{-1} in a lower dimension. Such a kind of relation also exists between any L_σ^{-1} and $L_{\sigma'}^{-1}$ where σ' is a sub-permutation of σ. Here, we say $\sigma' \in \Gamma_k$ is a sub-permutation of $\sigma \in \Gamma$ if $\sigma'(j) = \sigma(j), \forall j \in \Gamma_k$. For example, (134) and (123) are both sub-permutations of (1234).

A natural question is: given the relation between L_σ and its lower dimensional analogs, how to build a relation between Q and its lower dimensional counterparts? To answer this question, the following intuition is crucial: since $Q = E_\sigma(L_\sigma^{-1})$ is a symmetrization of L_σ^{-1}, we should symmetrize RHS (Right-Hand-Side) of (59). There are three levels of "asymmetry" in the RHS of (59): i) $L_{(12)}^{-1}$ is non-symmetric; ii) the off-diagonal blocks are not transpose to each other; iii) in this block partition of L_σ the 1st and 2nd row/columns are grouped together, so this expression is not symmetric with respect to the permutation of $\{1, 2, 3\}$. Let us briefly explain below how to build the three levels of symmetry.

The first level of symmetry is built by the matrix $\hat{Q}_k$. For example,

$$\hat{Q}_3 = \frac{1}{2}(L_{(12)}^{-1} + L_{(21)}^{-1}) = \begin{bmatrix} 1 & -\frac{1}{2}w_{12} \\ -\frac{1}{2}w_{12} & 1 \end{bmatrix} \tag{60}$$

is a symmetrization of $L_{(12)}^{-1}$ and forms the first-level symmetrization of the RHS of (59) (more details are given later). The second level of symmetry is built by the matrix Q_k. For example, $Q_3 = \begin{bmatrix} \hat{Q}_3 & -\frac{1}{2}\hat{Q}_3 w_3 \\ -\frac{1}{2}w_3^T\hat{Q}_3 & 1 \end{bmatrix}$ is the symmetrization of $\begin{bmatrix} \hat{Q}_3 & 0 \\ -w_3^T\hat{Q}_3 & 1 \end{bmatrix}$, thus forming the second level of symmetrization for the RHS of (59). The third level of symmetry is built by averaging the three matrices Q_1, Q_2, Q_3 (up to permutation of rows/columns), as shown by the induction formula (50)

$$Q = \frac{1}{3}(S_1 Q_1 S_1^T + S_2 Q_2 S_2^T + Q_3). \tag{61}$$

Below, we prove the induction formula (61) in a rigous way.

Proof of (61)*:* As the first level symmetrization, we prove

$$\frac{1}{2}(L_{(123)}^{-1} + L_{(213)}^{-1}) = \begin{bmatrix} \hat{Q}_3 & 0 \\ -w_3^T\hat{Q}_3 & 1 \end{bmatrix}. \tag{62}$$

Recall that $L_{(123)} = \begin{bmatrix} L_{(12)} & 0 \\ w_3^T & 1 \end{bmatrix}$ implies $L_{(123)}^{-1} = \begin{bmatrix} L_{(12)}^{-1} & 0 \\ -w_3^T L_{(12)}^{-1} & 1 \end{bmatrix}$. Similarly, $L_{(213)} = \begin{bmatrix} L_{(21)} & 0 \\ w_3^T & 1 \end{bmatrix}$ implies $L_{(213)}^{-1} = \begin{bmatrix} L_{(21)}^{-1} & 0 \\ -w_3^T L_{(21)}^{-1} & 1 \end{bmatrix}$. Summing up these two relations and invoking the definition of $\hat{Q}_3$ in (60), we obtain (62).

As the second level symmetrization, we prove

$$Q_3 = \frac{1}{4}\left(L_{(123)}^{-1} + L_{(213)}^{-1} + L_{(321)}^{-1} + L_{(312)}^{-1}\right). \tag{63}$$

Note that the common feature of the four permutations $(123), (213), (321), (312)$ is: 1 and 2 are adjacent in these permutations. By the definition of L_σ in (**??**), we have $L_{(321)} = L_{(123)}^T, L_{(312)} = L_{(213)}^T$, thus $L_{(321)}^{-1} = L_{(123)}^{-T}, L_{(312)}^{-1} = L_{(213)}^{-T}$. Taking the transpose over both sides of (62), we obtain

$$\frac{1}{2}(L_{(321)}^{-1} + L_{(312)}^{-1}) = \begin{bmatrix} \hat{Q}_3 & -\hat{Q}_3 w_3 \\ 0 & 1 \end{bmatrix}. \tag{64}$$

Combining (62) and (64), and using the definition of Q_3 in (51), we obtain (63).

Using a similar argument, we can prove

$$S_1 Q_1 S_1^T = \frac{1}{4}\left(L_{(123)}^{-1} + L_{(132)}^{-1} + L_{(231)}^{-1} + L_{(321)}^{-1}\right). \tag{65}$$

Again, the common feature of the four permutations $(123), (132), (231), (321)$ is: 2 and 3 are adjacent in these permutations. The proof of (65) is almost the same as the proof of (63), except the extra step to move rows and columns. Similarly, we can prove

$$S_2 Q_2 S_2^T = \frac{1}{4}\left(L_{(132)}^{-1} + L_{(312)}^{-1} + L_{(231)}^{-1} + L_{(213)}^{-1}\right). \tag{66}$$

As the third level symmetrization, combining (63),(65) and (66), and invoking the definition of Q in (25), we obtain (61).

7.4 Proof of Proposition 2

For simplicity, throughout this proof, we denote

$$w \triangleq w_n,\ \hat{Q} \triangleq \hat{Q}_n,\ \hat{A} \triangleq \hat{A}_n.$$

We claim that

$$0 \leq \theta \triangleq w^T \hat{Q} w < \frac{4}{3}. \tag{67}$$

In fact, by the definition $w = \hat{A}^T a_n$ we have $\theta = a_n^T \hat{A}\hat{Q}\hat{A}^T a_n \leq \rho(\hat{A}\hat{Q}\hat{A}^T)\|a_n\|^2 = \rho(\hat{A}\hat{Q}\hat{A}^T) < \frac{4}{3}$, which proves the last inequality of (67). According to the assumption, $\text{eig}(\hat{Q}\hat{A}^T\hat{A}) \subseteq (0, 4/3) \subseteq (0, \infty)$ and $\hat{A}$ is non-singular, thus $\hat{Q} \succ 0$. Then we have $\theta = w^T\hat{Q}w \geq 0$, which proves the first inequality of (67).

We apply a trick that we have previously used: factorize Q_n and change the order of multiplication. To be specific, Q_n defined in (56) can be factorized as

$$Q_n = \begin{bmatrix} I & 0 \\ -\frac{1}{2}w^T & 1 \end{bmatrix}\begin{bmatrix} \hat{Q} & 0 \\ 0 & 1 - \frac{1}{4}w^T\hat{Q}w \end{bmatrix}\begin{bmatrix} I & -\frac{1}{2}w \\ 0 & 1 \end{bmatrix} = J\begin{bmatrix} \hat{Q} & 0 \\ 0 & c \end{bmatrix}J^T, \tag{68}$$

where $J \triangleq \begin{bmatrix} I & 0 \\ -\frac{1}{2}w^T & 1 \end{bmatrix}$, I denotes the $(n-1)$-dim identity matrix and

$$c \triangleq 1 - \frac{1}{4}w^T\hat{Q}w. \tag{69}$$

It is easy to prove

$$\text{eig}(AQ_nA^T) \subseteq (0, \infty). \tag{70}$$

In fact, since A is non-singular, we only need to prove $Q_n \succ 0$. According to (68), we only need to prove $\begin{bmatrix} \hat{Q} & 0 \\ 0 & c \end{bmatrix} \succ 0$. This follows from $\hat{Q} \succ 0$, and the fact $c = 1 - \frac{1}{4}w^T\hat{Q}w \overset{(67)}{>} 1 - \frac{1}{3} > 0$. Thus (70) is proved.

It remains to prove

$$\rho(AQ_nA^T) < \frac{4}{3}. \tag{71}$$

Denote $\hat{B} \triangleq \hat{A}^T\hat{A}$, then we can write A^TA as

$$A^TA = \begin{bmatrix} \hat{B} & w \\ w^T & 1 \end{bmatrix}.$$

We simplify the expression of $\rho(AQ_nA^T) = \rho(Q_nA^TA)$ as follows:

$$\rho(AQ_nA^T) \overset{(68)}{=} \rho\left(J\begin{bmatrix}\hat{Q} & 0\\ 0 & c\end{bmatrix}J^TA^TA\right) = \rho\left(\begin{bmatrix}\hat{Q} & 0\\ 0 & c\end{bmatrix}J^TA^TAJ\right). \tag{72}$$

By algebraic computation, we have

$$J^TA^TAJ = \begin{bmatrix}I & -\frac{1}{2}w\\ 0 & 1\end{bmatrix}\begin{bmatrix}\hat{B} & w\\ w^T & 1\end{bmatrix}\begin{bmatrix}I & 0\\ -\frac{1}{2}w^T & 1\end{bmatrix} = \begin{bmatrix}I & -\frac{1}{2}w\\ 0 & 1\end{bmatrix}\begin{bmatrix}\hat{B}-\frac{1}{2}ww^T & w\\ \frac{1}{2}w^T & 1\end{bmatrix} = \begin{bmatrix}\hat{B}-\frac{3}{4}ww^T & \frac{1}{2}w\\ \frac{1}{2}w^T & 1\end{bmatrix},$$

thus

$$Z \triangleq \begin{bmatrix}\hat{Q} & 0\\ 0 & c\end{bmatrix}J^TA^TAJ = \begin{bmatrix}\hat{Q} & 0\\ 0 & c\end{bmatrix}\begin{bmatrix}\hat{B}-\frac{3}{4}ww^T & \frac{1}{2}w\\ \frac{1}{2}w^T & 1\end{bmatrix} = \begin{bmatrix}\hat{Q}\hat{B}-\frac{3}{4}\hat{Q}ww^T & \frac{1}{2}\hat{Q}w\\ \frac{1}{2}cw^T & c\end{bmatrix}. \tag{73}$$

According to (72), $\mathrm{eig}(AQ_nA^T) = \mathrm{eig}(Z)$, thus to prove (71), we only need to prove $\rho(Z) < \frac{4}{3}$. Since we have proved that $\mathrm{eig}(Z) = \mathrm{eig}(AQ_nA^T) \subseteq (0,\infty)$, we only need to prove $\lambda_{\max}(Z) < 4/3$. In the rest, we will prove that for any eigenvalue of Z, denoted as λ, we have

$$\lambda < \frac{4}{3}. \tag{74}$$

Suppose $v \in \mathbb{R}^n\backslash\{0\}$ is the eigenvector corresponding to λ, i.e. $Zv = \lambda v$. Partition v into $v = \begin{bmatrix}v_1\\ v_0\end{bmatrix}$, where $v_1 \in \mathbb{R}^{n-1}, v_0 \in \mathbb{R}$. According to the expression of Z in (73), $Zv = \lambda v$ implies

$$(\hat{Q}\hat{B} - \frac{3}{4}\hat{Q}ww^T)v_1 + \frac{1}{2}\hat{Q}wv_0 = \lambda v_1, \tag{75a}$$

$$\frac{1}{2}cw^Tv_1 + cv_0 = \lambda v_0. \tag{75b}$$

If $\lambda = c$, then (74) holds since $c = 1 - \frac{1}{4}\theta \le 1$. In the following, we assume $\lambda \neq c$. An immediate consequence is

$$v_1 \neq 0.$$

Otherwise, assume $v_1 = 0$; then (75b) implies $cv_0 = \lambda v_0$, which leads to $v_0 = 0$ and thus $v = 0$, a contradiction.

By (75b) we get

$$v_0 = \frac{c}{2(\lambda - c)}w^Tv_1.$$

Plugging into (75a), we obtain

$$\lambda v_1 = (\hat{Q}\hat{B} - \frac{3}{4}\hat{Q}ww^T)v_1 + \frac{1}{2}\hat{Q}w\frac{c}{2(\lambda-c)}w^Tv_1 = (\hat{Q}\hat{B} + \phi\hat{Q}ww^T)v_1, \tag{76}$$

where

$$\phi = -\frac{3}{4} + \frac{c}{4(\lambda - c)} = \frac{\lambda}{4(\lambda-c)} - 1 = \frac{\lambda}{4\lambda - 4 + \theta} - 1. \tag{77}$$

Here we have used the definition $c = 1 - \frac{1}{4}w^T\hat{Q}w = 1 - \frac{1}{4}\theta$.

Denote $\hat{\lambda} \triangleq \rho(\hat{Q}\hat{B})$, then by the assumption $\hat{\lambda} = \rho(\hat{Q}\hat{A}^T\hat{A}) \in (0, 4/3)$. We prove that

$$\lambda \le \begin{cases}\hat{\lambda} + \phi\theta, & \phi > 0,\\ \hat{\lambda}, & \phi \le 0.\end{cases} \tag{78}$$

Since $\hat{Q} \in \mathbb{R}^{(n-1)\times(n-1)}$ is a (symmetric) positive definite matrix, there exists $U \in \mathbb{R}^{(n-1)\times(n-1)}$ such that

$$\hat{Q} = U^TU.$$

Pick a positive number $g = |\phi|\theta$. By (76) we have $(g + \lambda)v_1 = (\hat{Q}\hat{B} + \phi\hat{Q}ww^T + gI)v_1$, here I denotes the identity matrix with dimension $n-1$. Consequently,

$$g + \lambda \in \mathrm{eig}(\hat{Q}\hat{B} + \phi\hat{Q}ww^T + gI) = \mathrm{eig}(U\hat{B}U^T + \phi Uww^TU^T + gI). \tag{79}$$

Note that ϕUww^TU^T is a rank-one symmetric matrix with a (possibly) non-zero eigenvalue $\phi w^TU^TUw = \phi w^T\hat{Q}w = \phi\theta$. By our definition $g = |\phi|\theta \ge \phi\theta$, which implies that $gI + \phi w^TU^TUw \succeq 0$ and

$$\rho(gI + \phi w^TU^TUw) = \begin{cases}g + \phi\theta, & \phi > 0,\\ g, & \phi \le 0.\end{cases} \tag{80}$$

Since both $U\hat{B}U^T = U\hat{A}^T\hat{A}U^T$ and $\phi Uww^TU^T + gI$ are symmetric PSD (Positive Semi-Definite) matrices, (79) implies

$$\begin{aligned} g + \lambda &\le \rho(U\hat{B}U^T + \phi Uww^TU^T + gI) \\ &\le \rho(U\hat{B}U^T) + \rho(\phi Uww^TU^T + gI) \\ &= \begin{cases} \hat{\lambda} + g + \phi\theta, & \phi > 0, \\ \hat{\lambda} + g, & \phi \le 0, \end{cases} \end{aligned} \tag{81}$$

which immediately leads to (78).

We claim that (74) follows from (78). In fact, if $\phi \le 0$, then by (78) we have $\lambda \le \hat{\lambda} < \frac{4}{3}$, which proves (74). Next we assume $\phi > 0$, which, by the definition of ϕ in (77), means

$$1 < \frac{\lambda}{4\lambda - 4 + \theta}. \tag{82}$$

If $\lambda \le 1$, then (74) already holds; thus we can assume $\lambda > 1$, which implies $4\lambda - 4 > 0$. Combining with the fact $\theta \ge 0$, we have

$$1 < \frac{\lambda}{4\lambda - 4 + \theta} < \frac{\lambda}{4\lambda - 4},$$

which leads to $\lambda < \frac{4}{3}$. This finishes the proof of (74).

8 Proof of Lemma 2 for the general case

Without loss of generality, we can assume

$$A_i^TA_i = I_{d_i \times d_i},\ i = 1, \ldots, n.$$

To show this, let us write M_σ, M as $M_\sigma(A_1, \ldots, A_n)$ and $M(A_1, \ldots, A_n)$ respectively, i.e. functions of the coefficient matrix $(A_1, \ldots, A_n)$. Define $\tilde{A}_i = A_i(A_i^TA_i)^{-\frac{1}{2}}$ and

$$D \triangleq \text{Diag}((A_1^TA_1)^{-\frac{1}{2}}, \ldots, (A_n^TA_n)^{-\frac{1}{2}}, I_{N\times N}).$$

It is easy to verify that $M_\sigma(A_1, \ldots, A_n) = D^{-1}M_\sigma(\tilde{A}_1, \ldots, \tilde{A}_n)D$, which implies

$$M(A_1, \ldots, A_n) = D^{-1}M(\tilde{A}_1, \ldots, \tilde{A}_n)D.$$

Thus $\rho(M(A_1, \ldots, A_n)) = \rho(M(\tilde{A}_1, \ldots, \tilde{A}_n))$. In other words, normalizing A_i to $\tilde{A}_i$, which satisfies $\tilde{A}_i^T\tilde{A}_i = I_{d_i\times d_i}$, does not change the spectral radius of M.

8.1 Proof Outline of Lemma 2 and Two Propositions

We use mathematical induction to prove Lemma 2 for the n-block case. For the basis of the induction ($n = 1$), Lemma 2 holds since $QA^TA = I_{d_1\times d_1}$. Assume Lemma 2 holds for $n-1$, we will prove Lemma 2 for n.

Similar to the n-coordinate case, we will first derive the induction formula, and then use this formula to prove the induction step.

8.1.1 Step 2.1: Induction formula for the n-block case

For any matrix $Z \in \mathbb{R}^{N\times N}$ with $n \times n$ blocks, denote $Z[i,j]$ as the (i,j)-th block of Z, $1 \le i, j \le n$. We use the term "the i-th block-row" to describe the collection of blocks $Z[i,1], \ldots, Z[i,n]$, and "the i-th block-column" to describe the collection of blocks $Z[1,i], \ldots, Z[n,i]$. We say the row pattern of Z is $(r_1, \ldots, r_n)$ and the column pattern of Z is $(c_1, \ldots, c_n)$ if $Z[i,j] \in \mathbb{R}^{r_i\times c_j}$, $\forall\ 1 \le i, j \le n$. The multiplication of two block partitioned matrices $Z_1, Z_2 \in \mathbb{R}^{N\times N}$ can be expressed using only the blocks if the column pattern of Z_1 is the same as the row pattern of Z_2.

For $k = 1, \ldots, n$, we define block-permutation matrix $S_k \in \mathbb{R}^{N\times N}$ with $n \times n$ blocks as follows:

$$S_k[i,i] \triangleq I_{d_i\times d_i}, i = 1, \ldots, k-1; \quad S_k[j,j-1] \triangleq I_{d_j\times d_j}, j = k+1, \ldots, n; \quad S_k[k,n] \triangleq I_{d_k\times d_k}, \tag{83}$$

and all other entries of S_k are set to zero. Note that the row pattern of S_k is $(d_1, \ldots, d_n)$ and the colum pattern of S_k is $(d_1, \ldots, d_{k-1}, d_{k+1}, \ldots, d_n, d_k)$. When $d_i = 1, \forall i$, this definition reduces to the definition (48) in the n-coordinate case. Similar to the n-coordinate case, this matrix S_k has the following block-column moving property: right multiplying a matrix with n block-columns and column pattern

$(d_1, \ldots, d_n)$ by S_k will move the k-th block-column to the end, resulting in a new matrix with column pattern $(d_1, \ldots, d_{k-1}, d_{k+1}, \ldots, d_n, d_k)$. Consequently, S_k^T has the following block-row moving property: left multiplying a matrix with n block-rows by S_k^T will move the k-th block-row to the end. Note that S_n is the identity matrix. Another property is

$$S_k^T = S_k^{-1}. \tag{84}$$

For any $k \in [n]$, define Γ_k as in (44). For any $\sigma' \in \Gamma_k$, $L_{\sigma'} \in \mathbb{R}^{(N-d_k)\times(N-d_k)}$ is partitioned into $(n-1) \times (n-1)$ blocks and the $(\sigma'(i), \sigma'(j))$-th block is defined by

$$L_{\sigma'}[\sigma'(i), \sigma'(j)] \triangleq \begin{cases} A_{\sigma'(i)}^T A_{\sigma'(j)} & i \geq j, \\ 0 & i < j, \end{cases} \tag{85}$$

We then define $\hat{Q}_k \in \mathbb{R}^{(N-d_k)\times(N-d_k)}$ by

$$\hat{Q}_k \triangleq \frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} L_{\sigma'}^{-1}, \; k = 1, \ldots, n. \tag{86}$$

Define W_k as the k-th block-column of $A^T A$ excluding the block $A_k^T A_k$, i.e.

$$W_k = [A_k^T A_1, \ldots, A_k^T A_{k-1}, A_k^T A_{k+1}, \ldots, A_k^T A_n]^T, \; \forall k \in [n]. \tag{87}$$

With these definitions, we are ready to present the induction formula.

Proposition 3 *The matrix $Q = \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} L_\sigma^{-1}$, where L_σ and Γ are defined by* (18) *and* (6) *respectively, can be decomposed as follows:*

$$Q = \frac{1}{n} \sum_{k=1}^{n} S_k Q_k S_k^T, \tag{88}$$

where

$$Q_k \triangleq \begin{bmatrix} \hat{Q}_k & -\frac{1}{2}\hat{Q}_k W_k \\ -\frac{1}{2} W_k^T \hat{Q}_k & I_{d_k \times d_k} \end{bmatrix}, \tag{89}$$

in which $\hat{Q}_k$ is defined by (86).

Proposition 3 is a generalization of Proposition 1 from the n-coordinate case to the n-block case, and its proof is similar to the proof of Proposition 1 (with a slight difference due to the block partition). We relegate this proof to Appendix 10.

8.1.2 Step 2.2: Bounding eigenvalues of each Q_k

According to (88), we have

$$AQA^T = \frac{1}{n} \sum_{k=1}^{n} AS_k Q_k S_k^T A^T.$$

Consequently,

$$\frac{1}{n} \sum_{k=1}^{n} \lambda_{\min}(AS_k Q_k S_k^T A^T) \leq \lambda_{\min}(AQA^T) \leq \lambda_{\max}(AQA^T) \leq \frac{1}{n} \sum_{k=1}^{n} \lambda_{\max}(AS_k Q_k S_k^T A^T). \tag{90}$$

To prove $\text{eig}(AQA^T) \subseteq (0, 4/3)$, we only need to prove for any $k = 1, \ldots, n$,

$$\text{eig}(AS_k Q_k S_k^T A^T) = \text{eig}(Q_k S_k^T A^T A S_k) \subseteq (0, 4/3). \tag{91}$$

By the block-column moving property of S_k, we have

$$\bar{A}_k \triangleq AS_k = [\hat{A}_k, A_k], \tag{92}$$

where $\hat{A}_k \triangleq [A_1, \ldots, A_{k-1}, A_{k+1}, \ldots, A_n]$. Note that $\hat{Q}_k$ only depends on the entries of $\hat{A}_k^T \hat{A}_k \in \mathbb{R}^{(N-d_k)\times(N-d_k)}$ which has $(n-1) \times (n-1)$ blocks, thus by the induction hypothesis, we have

$$\text{eig}(\hat{Q}_k \hat{A}_k^T \hat{A}_k) \subseteq (0, 4/3). \tag{93}$$

We claim that (91) follows from the induction hypothesis (93) and the expressions (92) and (89). In fact, the following proposition directly proves (91) for $k = n$. If we replace $A, \hat{A}_n, A_n, \hat{Q}_n, Q_n$ by $\bar{A}_k, \hat{A}_k, A_k, \hat{Q}_k, Q_k$ respectively in the following proposition, we will obtain (91) for any k. As mentioned earlier, the desired result $\text{eig}(AQA^T) \subseteq (0, 4/3)$ follows immediately from (91) and (90).

Proposition 4 *Suppose $A = [\hat{A}_n, A_n] \in \mathbb{R}^{N\times N}$ is a non-singular matrix, where $\hat{A}_n \in \mathbb{R}^{N\times(N-d_n)}$, and $A_n \in \mathbb{R}^{N\times d_n}$ satisfies $A_n^T A_n = I_{d_n\times d_n}$. Suppose $\hat{Q}_n \in \mathbb{R}^{(N-d_n)\times(N-d_n)}$ is symmetric and*

$$\mathit{eig}(\hat{Q}_n \hat{A}_n^T \hat{A}_n) \subseteq (0, 4/3). \tag{94}$$

Define

$$W_n \triangleq \hat{A}_n^T A_n \in \mathbb{R}^{(N-d_n)\times d_n}, \quad Q_n \triangleq \begin{bmatrix} \hat{Q}_n & -\frac{1}{2}\hat{Q}_n W_n \\ -\frac{1}{2}W_n^T \hat{Q}_n & I_{d_n\times d_n} \end{bmatrix}. \tag{95}$$

Then $\mathit{eig}(AQ_nA^T) \subseteq (0, \frac{4}{3})$.

Proposition 4 is a generalization of Proposition 2 from the n-coordinate case to the n-block case, and its proof is similar to the proof of Proposition 2 (with a few minor differences). The proof of Proposition 4 is given in Section 8.2.

8.2 Proof of Proposition 4

This proof is similar to the proof of Proposition 2 for the n-coordinate case, with a few minor differences due to the fact $d_n > 1$.

For simplicity, throughout this proof, we denote

$$W \triangleq W_n, \ \hat{Q} \triangleq \hat{Q}_n, \ \hat{A} \triangleq \hat{A}_n.$$

We first prove

$$0 \preceq \Theta \triangleq W^T \hat{Q} W \prec \frac{4}{3} I. \tag{96}$$

Since $\text{eig}(\hat{Q}\hat{A}^T\hat{A}) \subseteq (0, \infty)$ and $\hat{A}$ is non-singular, thus $\hat{Q} \succ 0$. Then we have $\Theta = W^T\hat{Q}W \succeq 0$, which proves the first relation of (96). By the definition $W = \hat{A}^T A_n$ we have

$$\begin{aligned} \rho(\Theta) = \rho(A_n^T \hat{A}\hat{Q}\hat{A}^T A_n) &= \max_{v\in\mathbb{R}^{d_n\times 1}, \|v\|=1} v^T A_n^T \hat{A}\hat{Q}\hat{A}^T A_n v \\ \leq \rho(\hat{A}\hat{Q}\hat{A}^T) \max_{v\in\mathbb{R}^{d_n\times 1}, \|v\|=1} \|A_n v\|^2 &= \rho(\hat{A}\hat{Q}\hat{A}^T)\|A_n\|^2 = \rho(\hat{A}\hat{Q}\hat{A}^T) < \frac{4}{3}, \end{aligned} \tag{97}$$

where the last equality is due to the assumption $A_n^T A_n = I$, and the last inequality is due to the assumption (94). By (97) we have $\Theta \prec \frac{4}{3}I$, thus (96) is proved.

We apply a trick that we have previously used: factorize Q_n and change the order of multiplication. To be specific, Q_n defined in (95) can be factorized as

$$Q_n = \begin{bmatrix} I & 0 \\ -\frac{1}{2}W^T & I \end{bmatrix} \begin{bmatrix} \hat{Q} & 0 \\ 0 & I - \frac{1}{4}W^T\hat{Q}W \end{bmatrix} \begin{bmatrix} I & -\frac{1}{2}W \\ 0 & I \end{bmatrix} = J \begin{bmatrix} \hat{Q} & 0 \\ 0 & C \end{bmatrix} J^T, \tag{98}$$

where $J \triangleq \begin{bmatrix} I & 0 \\ -\frac{1}{2}W^T & I \end{bmatrix}$, I in the upper left block denotes the $(N-d_n)$-dimensional identity matrix, I in the lower right block denotes the d_n-dim identity matrix, and

$$C \triangleq I - \frac{1}{4}W^T\hat{Q}W \in \mathbb{R}^{d_n\times d_n}. \tag{99}$$

It is easy to prove

$$\text{eig}(AQ_nA^T) \subseteq (0, \infty). \tag{100}$$

In fact, we only need to prove $Q_n \succ 0$. According to (98), we only need to prove $\begin{bmatrix} \hat{Q} & 0 \\ 0 & C \end{bmatrix} \succ 0$. This follows from $\hat{Q} \succ 0$ and the fact $C = I - \frac{1}{4}W^T\hat{Q}W \overset{(96)}{\succ} I - \frac{1}{3}I \succ 0$. Thus (100) is proved.

It remains to prove

$$\rho(AQ_nA^T) < \frac{4}{3}. \tag{101}$$

Denote $\hat{B} \triangleq \hat{A}^T\hat{A} \in \mathbb{R}^{(N-d_n)\times(N-d_n)}$, then we can write A^TA as

$$A^TA = \begin{bmatrix} \hat{B} & W \\ W^T & I \end{bmatrix}. \tag{102}$$

We simplify the expression of $\rho(AQ_nA^T)$ as follows:

$$\rho(AQ_nA^T) = \rho\left(AJ\begin{bmatrix}\hat{Q} & 0\\ 0 & C\end{bmatrix}J^TA^T\right) = \rho\left(\begin{bmatrix}\hat{Q} & 0\\ 0 & C\end{bmatrix}J^TA^TAJ\right). \tag{103}$$

By algebraic computation, we have

$$\begin{aligned} J^TA^TAJ &= \begin{bmatrix}I & -\frac{1}{2}W\\ 0 & I\end{bmatrix}\begin{bmatrix}\hat{B} & W\\ W^T & I\end{bmatrix}\begin{bmatrix}I & 0\\ -\frac{1}{2}W^T & I\end{bmatrix}\\ &= \begin{bmatrix}I & -\frac{1}{2}W\\ 0 & I\end{bmatrix}\begin{bmatrix}\hat{B}-\frac{1}{2}WW^T & W\\ \frac{1}{2}W^T & I\end{bmatrix} = \begin{bmatrix}\hat{B}-\frac{3}{4}WW^T & \frac{1}{2}W\\ \frac{1}{2}W^T & I\end{bmatrix}, \end{aligned} \tag{104}$$

thus

$$Z \triangleq \begin{bmatrix}\hat{Q} & 0\\ 0 & C\end{bmatrix}J^TA^TAJ = \begin{bmatrix}\hat{Q} & 0\\ 0 & C\end{bmatrix}\begin{bmatrix}\hat{B}-\frac{3}{4}WW^T & \frac{1}{2}W\\ \frac{1}{2}W^T & I\end{bmatrix} = \begin{bmatrix}\hat{Q}\hat{B}-\frac{3}{4}\hat{Q}WW^T & \frac{1}{2}\hat{Q}W\\ \frac{1}{2}CW^T & C\end{bmatrix}. \tag{105}$$

According to (103), $\rho(AQ_nA^T) = \rho(Z)$, thus to prove (101) we only need to prove

$$\rho(Z) < \frac{4}{3}.$$

Suppose $\lambda > 0$ is an arbitrary eigenvalue of Z. In the rest, we will prove

$$\lambda < \frac{4}{3}. \tag{106}$$

Suppose $v \in \mathbb{R}^{N\times 1}\backslash\{0\}$ is the eigenvector corresponding to λ, i.e. $Zv = \lambda v$. Partition v into $v = \begin{bmatrix}v_1\\ v_0\end{bmatrix}$, where $v_1 \in \mathbb{R}^{N-d_n}$, $v_0 \in \mathbb{R}^{d_n}$. According to the expression of Z in (105), $Zv = \lambda v$ implies

$$(\hat{Q}\hat{B} - \frac{3}{4}\hat{Q}WW^T)v_1 + \frac{1}{2}\hat{Q}Wv_0 = \lambda v_1, \tag{107a}$$

$$\frac{1}{2}CW^Tv_1 + Cv_0 = \lambda v_0. \tag{107b}$$

If $\lambda I - C$ is singular, i.e. λ is an eigenvalue of C, then by (96) we have $\frac{2}{3}I \prec C = 1 - \frac{1}{4}\Theta \preceq I$, which implies $\lambda \leq 1$, thus (106) holds. In the following, we assume

$$\lambda I - C \text{ is non-singular.} \tag{108}$$

An immediate consequence is

$$v_1 \neq 0,$$

since otherwise (107b) implies $Cv_0 = \lambda v_0$, which combined with (108) leads to $v_0 = 0$ and thus $v = 0$, a contradiction.

By (107b) we get

$$v_0 = \frac{1}{2}(\lambda I - C)^{-1}CW^Tv_1.$$

Plugging into (107a), we obtain

$$\lambda v_1 = (\hat{Q}\hat{B} - \frac{3}{4}\hat{Q}WW^T)v_1 + \frac{1}{2}\hat{Q}W\frac{1}{2}(\lambda I - C)^{-1}CW^Tv_1 = (\hat{Q}\hat{B} + \hat{Q}W\Phi W^T)v_1, \tag{109}$$

where

$$\begin{aligned} \Phi &\triangleq -\frac{3}{4}I + \frac{1}{4}(\lambda I - C)^{-1}C = -I + \frac{1}{4}[I + (\lambda I - C)^{-1}C]\\ &= -I + \frac{\lambda}{4}(\lambda I - C)^{-1} = -I + \lambda[(4\lambda - 4)I + \Theta]^{-1}. \end{aligned} \tag{110}$$

Here we have used the definition $C = I - \frac{1}{4}W^T\hat{Q}W = I - \frac{1}{4}\Theta$. Since Θ is a symmetric matrix, Φ is also a symmetric matrix.

To prove (106), we consider two cases.

Case 1: $\lambda_{\max}(\Phi) > 0$.

According to (110), we have

$$\theta \in \text{eig}(\Theta) \Longleftrightarrow -1 + \frac{\lambda}{(4\lambda - 4) + \theta} \in \text{eig}(\Phi).$$

By the assumption $\lambda_{\max}(\Phi) > 0$ and the above relation, there exists $\theta \in \text{eig}(\Theta)$ such that

$$-1 + \frac{\lambda}{(4\lambda - 4) + \theta} > 0. \tag{111}$$

If $\lambda < 1$, then (106) already holds; so we can assume $\lambda > 1$. By $\Theta \succeq 0$ we have $\theta \geq 0$, thus (111) implies $1 < \frac{\lambda}{(4\lambda-4)+\theta} \leq \frac{\lambda}{4\lambda-4}$, which leads to $\lambda < \frac{4}{3}$. Thus in Case 1 we have proved (106).

Case 2: $\lambda_{\max}(\Phi) \leq 0$, i.e. $\Phi \preceq 0$.

By the assumption (94) we have

$$\hat{\lambda} \triangleq \rho(\hat{Q}\hat{B}) = \rho(\hat{Q}\hat{A}^T\hat{A}) \in (0, 4/3). \tag{112}$$

Since $\hat{Q} \in \mathbb{R}^{(N-d_n)\times(N-d_n)}$ is a (symmetric) positive definite matrix, there exists a non-singular matrix $U \in \mathbb{R}^{(N-d_n)\times(N-d_n)}$ such that

$$\hat{Q} = U^T U. \tag{113}$$

Pick a positive number g that is large enough (will specify how large later). By (109) we have $(g+\lambda)v_1 = (\hat{Q}\hat{B} + \hat{Q}W\Phi W^T + gI)v_1$. Consequently,

$$\begin{aligned} g + \lambda \in \text{eig}(\hat{Q}\hat{B} + \hat{Q}W\Phi W^T + gI) &\overset{(113)}{=} \text{eig}(U^T U\hat{B} + U^T U W\Phi W^T + gI) \\ &= \text{eig}(U\hat{B}U^T + UW\Phi W^T U^T + gI). \end{aligned} \tag{114}$$

Define $\Gamma \triangleq UW\Phi W^T U^T \in \mathbb{R}^{(N-d_n)\times(N-d_n)}$, then the above relation implies

$$\begin{aligned} g + \lambda &\leq \rho(U\hat{B}U^T + \Gamma + gI) \\ &\leq \rho(U\hat{B}U^T) + \rho(\Gamma + gI) \\ &= \hat{\lambda} + \rho(\Gamma + gI), \end{aligned} \tag{115}$$

where the last equality is due to $\rho(U\hat{B}U^T) = \rho(\hat{A}^T\hat{A}U^T U) = \rho(\hat{A}^T\hat{A}\hat{Q}) \overset{(112)}{=} \hat{\lambda}$.

For any vector $v \in \mathbb{R}^{(N-d_n)\times(N-d_n)}$ we have

$$v^T\Gamma v = v^T UW\Phi W^T U^T v = (W^T U^T v)^T \Phi (W^T U^T v) \leq 0,$$

where the last inequality follows from our assumption $\Phi \preceq 0$, thus $\Gamma \preceq 0$. Pick a large g so that $g > \rho(\Gamma)$, then $\rho(gI + \Gamma) \leq g$. Plugging into (115), we get

$$g + \lambda \leq \hat{\lambda} + g,$$

which implies $\lambda \leq \hat{\lambda} < \frac{4}{3}$. Thus in Case 2 we have also proved (106). $\square$

Remark: The following generalization of (78) is also true:

$$\lambda \leq \begin{cases} \hat{\lambda} + \lambda_{\max}(\Phi)\|\Theta\|, & \lambda_{\max}(\Phi) > 0, \\ \hat{\lambda}, & \Phi \preceq 0. \end{cases} \tag{116}$$

The proof of (116) is a bit longer than the proof for the scalar case, and we will omit it in this paper. Note that (116) is not necessary for the proof of Proposition 4. In particular, when $\lambda_{\max}(\Phi) > 0$, we do not need to use $\lambda \leq \hat{\lambda} + \lambda_{\max}(\Phi)\|\Theta\|$ to bound λ; instead, it is enough to just use $\lambda_{\max}(\Phi) > 0$ to bound λ, as shown in Case 1 of the above proof.

9 Proof of Proposition 1, the induction formula for the n-coordinate case

In the n-coordinate case, the ambient dimension $N = n$, and the i-th block of A is $a_i \in \mathbb{R}^{n\times 1}$. Denote (σ', k) and (k, σ') as permutations of $[n]$ that are formed by combining a permutation σ' and k. For example, if $\sigma' = (124)$, then $(\sigma', 3) = (1243)$ and $(3, \sigma') = (3124)$.

We divide the proof into two parts. First, we present a simple formula related to the permutation matrices. Second, we apply the three levels of symmetrization.

9.1 Step 1. Deal with Permutation Matrices S_k

We first prove

$$S_k^T L_{(\sigma',k)} S_k = \begin{bmatrix} L_{\sigma'} & w_k \\ 0 & 1 \end{bmatrix}. \tag{117}$$

We write $L_{\sigma'}$ as a 2×2 block matrix

$$L_{\sigma'} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}, \tag{118}$$

where $Z_{11} \in \mathbb{R}^{(k-1)\times(k-1)}$, $Z_{12} \in \mathbb{R}^{(k-1)\times(n-k)}$, $Z_{21} \in \mathbb{R}^{(n-k)\times(k-1)}$, $Z_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, and denote

$$U_k = (a_1, \dots, a_{k-1}) \in \mathbb{R}^{n\times(k-1)}, \;\; V_k = (a_{k+1}, \dots, a_n) \in \mathbb{R}^{n\times(n-k)},$$

which implies

$$\begin{aligned} w_k &\overset{(47)}{=} [a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_n]^T a_k \\ &= [U_k, V_k]^T a_k \\ &= \begin{bmatrix} U_k^T a_k \\ V_k^T a_k \end{bmatrix}. \end{aligned} \tag{119}$$

It is easy to verify that

$$L_{(\sigma',k)} = \begin{bmatrix} Z_{11} & U_k^T a_k & Z_{12} \\ 0 & 1 & 0 \\ Z_{21} & V_k^T a_k & Z_{22} \end{bmatrix}.$$

Note that in the above expression, $\begin{bmatrix} U_k^T a_k \\ 1 \\ V_k^T a_k \end{bmatrix}$ is the k'th column and $[0, \dots, 0, 1, 0, \dots 0] \in \mathbb{R}^{1\times n}$ with the entry 1 in the k'th position is the k'th row. By moving the k'th column to the end and then moving the k'th row to the end, we get

$$S_k^T \begin{bmatrix} Z_{11} & U_k^T a_k & Z_{12} \\ 0 & 1 & 0 \\ Z_{21} & V_k^T a_k & Z_{22} \end{bmatrix} S_k = S_k^T \begin{bmatrix} Z_{11} & Z_{12} & U_k^T a_k \\ 0 & 0 & 1 \\ Z_{21} & Z_{22} & V_k^T a_k \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} & U_k^T a_k \\ Z_{21} & Z_{22} & V_k^T a_k \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} L_{\sigma'} & w_k \\ 0 & 1 \end{bmatrix},$$

where the last equality follows from (118) and (119). Thus we have proved (117).

9.2 Step 2: Three Levels of Symmetrization

Taking the inverse of both sides of (117) and using $S_k^{-1} = S_k^T$, we obtain

$$S_k^T L_{(\sigma',k)}^{-1} S_k = \begin{bmatrix} L_{\sigma'}^{-1} & -L_{\sigma'}^{-1} w_k \\ 0 & 1 \end{bmatrix}. \tag{120}$$

As the first level symmetrization, summing up (120) for all $\sigma' \in \Gamma_k$ and dividing by $|\Gamma_k|$, we get

$$\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} S_k^T L_{(\sigma',k)}^{-1} S_k = \begin{bmatrix} \frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} L_{\sigma'}^{-1} & -\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} L_{\sigma'}^{-1} w_k \\ 0 & 1 \end{bmatrix} \overset{(46)}{=} \begin{bmatrix} \hat{Q}_k & -\hat{Q}_k w_k \\ 0 & 1 \end{bmatrix}. \tag{121}$$

As the second level symmetrization, we can prove

$$\frac{1}{2|\Gamma_k|} S_k^T \left(\sum_{\sigma' \in \Gamma_k} L_{(\sigma',k)}^{-1} + \sum_{\sigma' \in \Gamma_k} L_{(k,\sigma')}^{-1} \right) S_k = \begin{bmatrix} \hat{Q}_k & -\frac{1}{2}\hat{Q}_k w_k \\ \frac{1}{2} w_k^T \hat{Q}_k & 1 \end{bmatrix} = Q_k. \tag{122}$$

In fact, by the definition of L_σ in (18), it is easy to see that

$$L_\sigma^T = L_{\bar{\sigma}},$$

where $\bar{\sigma}$ is a "reverse permutation" of σ that satisfies $\bar{\sigma}(i) = \sigma(n+1-i), \forall i$. Thus we have $L_{(\sigma',k)}^T = L_{(k,\bar{\sigma'})}^T$, where $\bar{\sigma'}$ is a reverse permutation of σ'. Summing over all σ', we get $\sum_{\sigma' \in \Gamma_k} L_{(\sigma',k)}^{-1} = \sum_{\sigma' \in \Gamma_k} L_{(k,\bar{\sigma'})}^{-T} = \sum_{\sigma' \in \Gamma_k} L_{(k,\sigma')}^{-T}$, where the last equality is because summing over $\bar{\sigma'}$ is the same as summing over σ'. Thus, we have

$$\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} S_k^T L_{(k,\sigma')}^{-1} S_k = \left(\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} S_k^T L_{(\sigma',k)}^{-1} S_k \right)^T \overset{(121)}{=} \begin{bmatrix} \hat{Q}_k & 0 \\ -w_k^T \hat{Q}_k & 1 \end{bmatrix}.$$

Here we have used the fact that $\hat{Q}_k$ is symmetric. Combining the above relation and (121) and invoking the definition of Q_k in (51) yields (122).

According to (84) and the fact $|\Gamma_k| = (n-1)!$, we can rewrite (122) as

$$S_k Q_k S_k^T = \frac{1}{2(n-1)!}\left(\sum_{\sigma'\in\Gamma_k} L_{(\sigma',k)}^{-1} + \sum_{\sigma'\in\Gamma_k} L_{(k,\sigma')}^{-1}\right).$$

As the third level symmetrization, summing up the above relation for $k = 1,\dots,n$ and then dividing by n, we get

$$\frac{1}{n}\sum_{k=1}^{n} S_k Q_k S_k^T = \frac{1}{n}\frac{1}{2(n-1)!}\sum_{k=1}^{n}\left(\sum_{\sigma'\in\Gamma_k} L_{(\sigma',k)}^{-1} + \sum_{\sigma'\in\Gamma_k} L_{(k,\sigma')}^{-1}\right) = \frac{1}{2n!}2\sum_{\sigma\in\Gamma} L_\sigma^{-1} = Q,$$

which proves (50). **Q.E.D.**

10 Proof of Proposition 3, the induction formula for the general n-block case

This proof is a direct extension of the proof of Proposition 1, i.e. the induction formula for the n-coordinate case. The major difference is Step 1 (the proof of (123)), since the permutation matrix S_k here is a block-partitioned matrix. Step 2 is the same as the n-coordinate case.

10.1 Step 1. Deal with Permutation Matrices S_k

We will prove

$$S_k^T L_{(\sigma',k)} S_k = \begin{bmatrix} L_{\sigma'} & W_k \\ 0 & I \end{bmatrix}, \tag{123}$$

where I denotes $I_{d_k\times d_k}$.

Note that $L_{\sigma'} \in \mathbb{R}^{(N-d_k)\times(N-d_k)}$ can be viewed as a block partitioned matrix with $(n-1)\times(n-1)$ blocks, and both the row pattern and column pattern of $L_{\sigma'}$ are $(d_1,\dots,d_{k-1},d_{k+1},\dots,d_n)$. By grouping the first $(k-1)$ block-rows and the last $(n-k)$ block-rows respectively, and grouping the first $(k-1)$ block-columns and the last $(n-k)$ block-columns respectively, $L_{\sigma'}$ can be written as a 2×2 block matrix

$$L_{\sigma'} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}, \tag{124}$$

where $Z_{11} \in \mathbb{R}^{(d_1+\cdots+d_{k-1})\times(d_1+\cdots+d_{k-1})}$, $Z_{22} \in \mathbb{R}^{(d_{k+1}+\cdots+d_n)\times(d_{k+1}+\cdots+d_n)}$, and the size of Z_{12} and Z_{22} can be determined accordingly. We denote

$$U_k = (A_1,\dots,A_{k-1}) \in \mathbb{R}^{N\times(d_1+\cdots+d_{k-1})},\;\; V_k = (A_{k+1},\dots,A_n) \in \mathbb{R}^{N\times(d_{k+1}+\cdots+d_n)},$$

which implies

$$\begin{aligned} W_k &\overset{(87)}{=} [A_k^T A_1,\dots,A_k^T A_{k-1},A_k^T A_{k+1},\dots,A_k^T A_n]^T \\ &= [A_1,\dots,A_{k-1},A_{k+1},\dots,A_n]^T A_k \\ &= [U_k,V_k]^T A_k \\ &= \begin{bmatrix} U_k^T A_k \\ V_k^T A_k \end{bmatrix}. \end{aligned} \tag{125}$$

It is easy to verify that

$$L_{(\sigma',k)} = \begin{bmatrix} Z_{11} & U_k^T A_k & Z_{12} \\ 0 & I_{d_k\times d_k} & 0 \\ Z_{21} & V_k^T A_k & Z_{22} \end{bmatrix}.$$

Note that in the above expression, $\begin{bmatrix} U_k^T A_k \\ I_{d_k\times d_k} \\ V_k^T A_k \end{bmatrix}$ is the k'th block-column of $L_{(\sigma',k)}$ and

$$[0, I_{d_k\times d_k}, 0] = [0_{d_k\times d_1},\dots,0_{d_k\times d_{k-1}}, I_{d_k\times d_k}, 0_{d_k\times d_{k+1}},\dots 0_{d_k\times d_n}] \in \mathbb{R}^{d_k\times N}$$

with $I_{d_k\times d_k}$ being the k'th block is the k'th block-row of $L_{(\sigma',k)}$. By moving the k'th block-column to the end and then moving the k'th block-row to the end, we get

$$S_k^T \begin{bmatrix} Z_{11} & U_k^T a_k & Z_{12} \\ 0 & I_{d_k\times d_k} & 0 \\ Z_{21} & V_k^T a_k & Z_{22} \end{bmatrix} S_k = S_k^T \begin{bmatrix} Z_{11} & Z_{12} & U_k^T a_k \\ 0 & 0 & I_{d_k\times d_k} \\ Z_{21} & Z_{22} & V_k^T a_k \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} & U_k^T a_k \\ Z_{21} & Z_{22} & V_k^T a_k \\ 0 & 0 & I_{d_k\times d_k} \end{bmatrix} = \begin{bmatrix} L_{\sigma'} & W_k \\ 0 & I_{d_k\times d_k} \end{bmatrix},$$

where the last equality follows from (124) and (125). Thus we have proved (123).

10.2 Step 2: Three Levels of Symmetrization

The rest of the proof of Proposition 3 is the same as the proof of Proposition 1, except a minor difference that w_k is replaced by W_k. For completeness, we still present the proof of Step 2 in detail.

Taking the inverse of both sides of (123) and using $S_k^{-1} = S_k^T$, we obtain

$$S_k^T L_{(\sigma',k)}^{-1} S_k = \begin{bmatrix} L_{\sigma'}^{-1} & -L_{\sigma'}^{-1} W_k \\ 0 & I \end{bmatrix}. \tag{126}$$

As the first level symmetrization, summing up (126) for all $\sigma' \in \Gamma_k$ and dividing by $|\Gamma_k|$, we get

$$\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} S_k^T L_{(\sigma',k)}^{-1} S_k = \begin{bmatrix} \frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} L_{\sigma'}^{-1} & -\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} L_{\sigma'}^{-1} W_k \\ 0 & I \end{bmatrix} \overset{(86)}{=} \begin{bmatrix} \hat{Q}_k & -\hat{Q}_k W_k \\ 0 & I \end{bmatrix}. \tag{127}$$

As the second level symmetrization, we will prove

$$\frac{1}{2|\Gamma_k|} S_k^T \left(\sum_{\sigma' \in \Gamma_k} L_{(\sigma',k)}^{-1} + \sum_{\sigma' \in \Gamma_k} L_{(k,\sigma')}^{-1} \right) S_k = \begin{bmatrix} \hat{Q}_k & -\frac{1}{2}\hat{Q}_k W_k \\ -\frac{1}{2} W_k^T \hat{Q}_k & I_{d_k \times d_k} \end{bmatrix} = Q_k. \tag{128}$$

By the definition of L_σ in (18), it is easy to see that

$$L_\sigma^T = L_{\bar{\sigma}},$$

where $\bar{\sigma}$ is a "reverse permutation" of σ that satisfies $\bar{\sigma}(i) = \sigma(n+1-i), \forall i$. Thus we have $L_{(\sigma',k)} = L_{(k,\bar{\sigma'})}^T$, where $\bar{\sigma'}$ is a reverse permutation of σ'. Summing over all σ', we get $\sum_{\sigma' \in \Gamma_k} L_{(\sigma',k)}^{-1} = \sum_{\sigma' \in \Gamma_k} L_{(k,\bar{\sigma'})}^{-T} = \sum_{\sigma' \in \Gamma_k} L_{(k,\sigma')}^{-T}$, where the last equality is because summing over $\bar{\sigma'}$ is the same as summing over σ'. Thus, we have

$$\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} S_k^T L_{(k,\sigma')}^{-1} S_k = \left(\frac{1}{|\Gamma_k|} \sum_{\sigma' \in \Gamma_k} S_k^T L_{(\sigma',k)}^{-1} S_k \right)^T \overset{(127)}{=} \begin{bmatrix} \hat{Q}_k & 0 \\ -W_k^T \hat{Q}_k & 1 \end{bmatrix}.$$

Here we have used the fact that $\hat{Q}_k$ is symmetric. Combining the above relation and (127) and invoking the definition of Q_k in (89) yields (128).

According to (84) and the fact $|\Gamma_k| = (n-1)!$, we can rewrite (128) as

$$S_k Q_k S_k^T = \frac{1}{2(n-1)!} \left(\sum_{\sigma' \in \Gamma_k} L_{(\sigma',k)}^{-1} + \sum_{\sigma' \in \Gamma_k} L_{(k,\sigma')}^{-1} \right).$$

As the third level symmetrization, summing up the above relation for $k = 1, \dots, n$ and then dividing by n, we get

$$\frac{1}{n} \sum_{k=1}^n S_k Q_k S_k^T = \frac{1}{n} \frac{1}{2(n-1)!} \sum_{k=1}^n \left(\sum_{\sigma' \in \Gamma_k} L_{(\sigma',k)}^{-1} + \sum_{\sigma' \in \Gamma_k} L_{(k,\sigma')}^{-1} \right) = \frac{1}{2n!} 2 \sum_{\sigma \in \Gamma} L_\sigma^{-1} = Q,$$

which proves (88). **Q.E.D.**